

REPORT DOCUMENTATION PAGE					Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to the Department of Defense, Executive Service Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.						
1. REPORT DATE (DD-MM-YYYY) 16-02-2016		2. REPORT TYPE Final			3. DATES COVERED (From - To) May 2014 - Feb 2016; Feb 2016	
4. TITLE AND SUBTITLE PHASE 2: THE MARS GAME SUMMATIVE STUDY REPORT (TRUE GAME PROJECT)				5a. CONTRACT NUMBER W911QY-14-C-0027		
				5b. GRANT NUMBER		
				5c. PROGRAM ELEMENT NUMBER 0603769SE		
6. AUTHOR(S) Freeman, Barbara				5d. PROJECT NUMBER		
				5e. TASK NUMBER		
				5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) StandardsWork, Inc. - http://www.standardswork.org/ Lockheed Martin Corporation- http://www.lockheedmartin.com/					8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) ADL Initiative - http://www.adlnet.gov					10. SPONSOR/MONITOR'S ACRONYM(S)	
					11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.						
13. SUPPLEMENTARY NOTES						
14. ABSTRACT						
15. SUBJECT TERMS						
16. SECURITY CLASSIFICATION OF: a. REPORT b. ABSTRACT c. THIS PAGE			17. LIMITATION OF ABSTRACT		18. NUMBER OF PAGES 23	
19a. NAME OF RESPONSIBLE PERSON					19b. TELEPHONE NUMBER (Include area code)	

Reset



PHASE 2: THE MARS GAME SUMMATIVE STUDY REPORT (TRUE GAME PROJECT)

Dr. Barbara Freeman, Visiting Scholar UC Berkeley

STANDARDWORK, INC.

DELIVERED TO ADL - FEBRUARY 1, 2016



Table of Contents

Executive Summary	2
Hypothesis, Study Goals, and Background	2
Methods	3
Research Questions.....	3
Study Participants	4
Procedures and Data Collection	4
Research Design	5
Interventions.....	6
Measures and Instruments	8
Data Analysis	9
Results	11
Major Conclusions and Implications	16
Hypothesis.....	16
Study Goals	17
Other Conclusions and Implications.....	19
True Game Project Studies	20
Limitations and Opportunities for Further Study	20
References	21
 Tables and Figures	
Table 1: Summary of Study Participants by Characteristic	4
Table 2: Summary of Conditions.....	6
Table 3: Statistical Technique used to address each Research Question	10
Table 4: Summary of Pre-Post Scores for Research Questions 1 and 2	11
Table 5: Independent Samples <i>t</i> -Tests for Engagement Sub-Factors	12
Table 6: Results of Programming Posttest – Single versus Randomized Pairs.....	13
Table 7: Results of Engagement Posttest – Single versus Randomized Pairs.....	13
Table 8: Summary of the Main Effects of the Multivariate Analysis	14
Figure 1: Time on Task.....	16

1. EXECUTIVE SUMMARY

This paper reports the major results of The Mars Game study, a digital learning game designed to teach mathematics and programming knowledge and skills. The primary goal of this study was to evaluate whether games designed to impart STEM knowledge in general—and The Mars Game in particular—can be both engaging and effective in increasing learning outcomes.

We performed a randomized controlled trial with a pretest-posttest control group design (Gall, Gall, & Borg, 2007). The sample consisted of eighty students in high school who were randomly assigned into one of four conditions. There were two treatment and two control groups, with one condition in each group working individually on their respective intervention and one group working in randomized pairs. The interventions that were studied were supplemental and were not replacements for teacher or the core curriculum. Pre-posttests were used to measure learning in math and programming. The EGameFlow scale (Fu, Su, & Yu, 2009) was used to measure engagement.

The study demonstrated that The Mars Game treatment group's overall learning outcomes were significantly greater than the control group. The gains in achievement were driven largely through an increase in programming posttest scores compared to a control group baseline. It also showed that students in the treatment group were significantly more engaged than in the control group, with a large effect size for engagement as a whole and for the immersion, feedback, and perception of knowledge improvement sub-factors of the construct.

Confirming the findings of a previous experimental study of The Mars Game, this current study demonstrated that well-designed and well-integrated serious games, which strike a critical balance between the 'fun' and the learning, have the ability to engage and increase the knowledge and skills of students in STEM subjects. Furthermore, it showed that The Mars Game worked similarly well in an afterschool setting with a highly diverse student population in a high-poverty school.

Other findings that are reported in this paper are:

- Boys and girls learned equally by using the game and were equally engaged.
- The Mars Game intervention had the greatest success with students who are new to the content being taught (ninth graders).
- Digital learning games have the potential to improve the learning outcomes of students who may have previously been switched off from learning (think that math is “horrible”).
- Assignment to the treatment group condition mattered more than assignment to work with the respective intervention individually or collaboratively in randomized pairs.

2. HYPOTHESIS, STUDY GOALS, AND BACKGROUND

The main hypothesis evaluated in this study is that an engaging game design-led strategy will result in a STEM learning solution that is enjoyable to students, and will impart knowledge more effectively than traditional learning paradigms. This means that the educational content is seamlessly and invisibly embedded in the game, not that a few game-like elements or nice graphics were simply added on top of an instructional curriculum. We expect that the fact that the students

are engaged, that they are having fun, will result in their immersion in the story, the challenge, the game, and ultimately in the *learning*. We call this “True Game” vs. “Chocolate Covered Broccoli”. Inherent in our hypothesis is the belief that directly engaging students as players immersed in the gameplay will help students learn.

As part of this design-based research project, five rigorous studies of The Mars Game were conducted over two years: three formative and two summative. This is the fifth research study and the second summative study. The objective of the formative studies was to test the game to get students’ reactions and to see if the developers were on the right track with their development efforts, and to modify the game based on students’ feedback and in-depth observations (Freeman, 2015). The first summative study was an experimental study to investigate the impact of The Mars Game prototype on ninth and tenth grade students’ learning outcomes in mathematics and programming and on their engagement with the game. It demonstrated that The Mars Game prototype was significantly more effective in causing learning gains and significantly more engaging than the control intervention, with a high effect size for student engagement and immersion in the gameplay. Since that study, the game developers have substantially expanded the content provided in the game. The primary goal of this current study was to assess whether the expanded game was also *effective* (i.e., learning happens in the areas of mathematics and programming) and *engaging* (e.g., it is immersive). Based on the observations and results from our previous experiments, we wanted to evaluate the game under a different setting and conditions whilst once again conducting an experimental study and maintaining the central inquiry into the effectiveness and engagement of The Mars Game versus a control group. The new circumstances included how well the game worked in an afterschool setting with a highly-diverse student population in a high-poverty school, and whether the game would be more effective if students played collaboratively in randomized pairs or played it alone.

We provided the literature review in the first phase of this project. This study builds upon the same body of knowledge and is therefore not repeated in this report. This report includes the following sections: Methods, Results, Major Conclusions and Implications, Project Studies, and Limitations and Opportunities for Further Study. The appendices have been labelled and included separately as part of a package of deliverables.

3. METHODS

3.1 Research Questions

The study addressed four key research questions:

RQ1: What is the effect of The Mars Game on high school students’ knowledge and skills in mathematics and programming?

RQ2: What is the effect of The Mars Game on high school students’ engagement with STEM subjects?

RQ3a: What is the effect of individual gameplay versus randomized pairs on learning outcomes in STEM subjects?

RQ3b: What is the effect of individual gameplay versus randomized pairs on engagement with STEM subjects?

RQ4a: Is the relationship between learning and engagement affected by students' characteristics?

RQ4b: Is the relationship between learning and engagement affected by students' perceptions?

Research questions 3a and 3b were added because of observations in previous studies about student's successfully working with other students to solve problems in the game. Please see Research Design for further discussion as to why this question was included and how we mitigated the risk of its inclusion.

3.2 Study Participants

The study was conducted at Ocoee High School in Orange County Public School (OCPS), Florida. Ocoee is a high school of approximately 2,500 students. The school's population is highly diverse: 38% Black, 32% White, 23% Hispanic, 5% Asian, and 2% Other. It is a medium high-poverty school, with 61% of students eligible for free or reduced lunch. Fifty-nine percent of students are performing at basic or below basic proficiency in math.

The eighty high school students in the sample reflect the general population of the school. Table 1 provides a summary of the characteristics of the student participants, corresponding to treatment and control groups. Five teachers also participated in the study; one for each condition and one who organized the study for the school.

Table 1

Summary of Study Participants by Characteristic

Variable	Groups	Total (<i>n</i> = 80)	All Treatment* (<i>n</i> = 42)	All Control* (<i>n</i> = 38)
Gender	Male	33 (41.3%)	20 (47.6%)	13 (34.2%)
	Female	47 (58.8%)	22 (52.4%)	25 (65.8%)
Grade	9 th	54 (67%)	26 (61.9%)	28 (61.9%)
	10 th	17 (21.3%)	9 (21.4%)	8 (21.1%)
	11-12 th	9 (11.3%)	7 (16.7%)	2 (5.3%)

* All Treatment includes students in the treatment single and treatment pair conditions. All Control includes students in the control single and control pair conditions.

3.3 Procedures and Data Collection

The study was conducted for one month. First, teachers were recruited to participate. The Principal Researcher provided a training session to introduce the teachers to The Mars Game intervention, explained the purpose and procedures of the study, and selected an alternative intervention to be used by the control group (discussed below). Second, as this is a one-to-one computing school that provides laptops to each student, the schools' laptops were tested to ensure compatibility with the game. The Mars Game websites were 'whitelisted' by the school district in order that the game could be delivered from external servers and the data could be collected from The Mars Game logs as students interacted with the game (e.g., the mission reached, time-on-task). Third, teachers in all algebra and geometry classes recruited their students to participate in the study, including informing them of the end-of-study rewards (raffle for an iPad mini and Beats Headsets) and certificates (Appendix A). Students in both conditions were offered the same incentives and rewards to participate in the study. Participation was purely voluntary, and

students who signed up first were selected for participation. Although we did not hold formal interviews with students to ask them why they signed up for the study, anecdotal evidence and conversation with teachers suggests that it was curiosity about the video game, the incentives, and the desire to do something afterschool with their friends largely drove students to voluntary for the study. Fourth, the Principal Researcher worked with the Head Math Coach to randomize student participants into conditions, ensure data anonymity and other security issues (e.g., generating random login ids and passwords, obtaining consent letters), and address logistical issues (e.g., making sure students went to the classroom of their assigned intervention, ensuring students brought headsets and power cords).

Finally, the study was implemented afterschool during three sessions, over two weeks. One classroom for each condition was used. One teacher was assigned to observe each classroom, and the Head Math Coach worked with the Principal Researcher to check that the teachers observed the study protocol and procedures (Appendix B). Three facilitators from the StandardsWork/Lockheed Martin team were also present (including the Principal Researcher); two to oversee the fidelity of implementation and one to ensure that the technology ran smoothly. Teachers instructed students in both conditions to take all pre-posttests under “test conditions”. Teachers did not provide the same instructions to students regarding interaction with their respective interventions. That is, they were not informed that they could talk or collaborate with their peers, nor were they told that they could not do so.

Students interacted with their respective interventions for a total of approximately 220 minutes. Students in both conditions accessed their interventions using their personal school-issued laptops. During Session One, all treatment and control students took the Mathematics and Programming Learning Pretest (Appendix C, shows an example of Pretest A student version and version with answers and distractors and Posttest B student version and version with answers and distractors. There is an equivalent version A and B for all tests) and a Student Self-assessment about their math perceptions and programming and gaming experience (Appendix D). This was followed immediately by students working on their assigned interventions for approximately 60 minutes including setup and login. During Session Two, all treatment and control students continued working on their interventions for approximately 105 minutes. In Session Three, students worked on their interventions for about 55 minutes. Then, students in both conditions took the EGameFlow engagement posttest (Appendix E) and the Mathematics and Programming Learning Posttest (Appendix C). The study ended with a pizza party, where certificates and rewards were issued.

Only those students who attended all three sessions, took the pre-posttests seriously, and stayed on task were included in the final study participant count. There was an overall student attrition rate of 17 percent from Day 1 to Day 3 completion of all requirements, with greater attrition in the control conditions.

3.4 Research Design

This study used a randomized controlled trial with a pretest-posttest control group design (Gall, Gall, & Borg, 2007). The sample consisted of voluntary student participants who were randomly assigned into one of four conditions: (1) The Mars Game treatment working individually (Mars Game Single), (2) The Mars Game treatment working in randomly selected pairs (Mars Game

Pair), (3) the control intervention working individually (Control Single), and (4) the control intervention working in randomly selected pairs (Control Pair). Teachers were randomly assigned to each condition.

During the previous studies, we observed students in the treatment group working together, discussing the game, and mainly staying on task. We thought this was an interesting aspect of gameplay, and although The Mars Game was not designed as a multiplayer game, we wanted to observe and study this phenomena in a more formal way. We therefore structured pair collaboration into the study design, and created treatment and control conditions in which students were assigned randomly into pairs (i.e., we did not match pairs according to ability or along any other dimensions). The pairs played the game on either two computers (independent sessions) or one computer with one person driving. The pairs were given the choice which of these two ways they wished to interact. Sixty percent played on separate computers and 40 percent on one computer. Although students assigned to the ‘pairs’ conditions interacted with their respective interventions with a partner, they took the pre-posttests individually and were assessed at an individual level (see Measures and Instruments section).

We mitigated any potential confounds that could arise from this design by ensuring that the treatment and control group were randomly selected into conditions and all conditions were identical to one another in all aspects (e.g., same study protocol, etc.). Furthermore, participants were told that they would have access to The Mars Game intervention after the experiment; this together with the short duration of the study and randomized assignment across classrooms and teachers safeguarded and minimized threats to internal validity. Table 2 shows a summary of the conditions. The four conditions were evaluated in research questions 3a and 3b, though the combined treatment and control groups (see Table 1) were the central focus of comparison, as manifested in research questions 1, 2, and 4.

Table 2
Summary of Conditions

Condition	N	%
Mars Game Single	21	26.2
Mars Game Pair	21	26.3
Control Single	18	22.5
Control Pair	20	25.0

3.5 Interventions

The interventions used by the treatment group and by the control group were both supplemental interventions and were not meant to replace either the teacher or the core curriculum. Students in the treatment groups used The Mars Game and students in the control groups used MathXL (<http://www.pearsonmylabandmastering.com/northamerica/mathxl/>) for their intervention. MathXL is a well-established online mathematics supplemental instructional tool developed and sold by the world’s largest publisher. It is an interactive and adaptive supplemental instructional program that provides personalized recommendations and tutorial exercises pertaining to what an individual student should be practicing next, based on his or her current performance. MathXL was selected by the teachers in this study for a number of reasons. It is the supplemental

intervention used by the school, as well as by millions of other students nationally and globally, and it is considered to be highly representative of ‘conventional’ supplemental instruction, which is now largely web-based and interactive (as opposed to textbooks and pencil and paper methods of supplemental instruction).

The Mars Game and MathXL cover the same math content. The content in both interventions align to the Common Core State Standards for Mathematics (CCSSM), focusing on topics related to the coordinate plane, slope-intercept form, transformations, and triangles (see Appendix F for an alignments table and example of alignment development process). Two of the teachers involved in the study created assignments for the control group to ensure that students were working on the same content as the treatment group.

The Mars Game teaches both math and programming skills and knowledge, whereas MathXL only teaches math. The Mars Game teaches students how to use Blockly, (<https://developers.google.com/blockly/>), an open-source block-based programming language created by Google. It also teaches basic coding skills including sequencing commands, conditionals, and loops amongst others. As far as we are aware, there are no supplemental programs that teach both math and programming to high school students. There are programs designed to teach coding to students of any age (e.g., Hour of Code) and online courses designed to teach programming (e.g., MIT OpenCourseWare); however, these courses are not widely used as classroom supplements. It is important to note that no student participants in the treatment or control conditions were formally taking a computer programming course during the implementation of this study. This, in conjunction with the short duration of the study (as discussed above) eliminated the teacher effect.

There are numerous supplemental programs that teach CCSSM-aligned math content, which could serve as a true comparison to The Mars Game. Choosing the control intervention has several considerations, which are listed below.

- All students—in both the treatment and control conditions—took the overall learning pretest and posttest. Each test consists of 18 questions, 14 questions pertain to mathematics and 4 relate to programming. Treatment and control group students were asked to answer the programming questions as well as the mathematics questions. See Measures and Instruments for a further description of the assessments.
- With respect to the mathematics pre-posttest questions, which comprise 77 percent of the overall pre-posttest, we have a true comparison group, as the MathXL group received a comparable math intervention to the treatment group.
- Since students who used MathXL were required to answer the programming questions, for the 23 percent of pre-post questions that address only programming, we have a solid control group that can serve as a baseline against which we can assess treatment students’ performance. We do not, however, have a true comparison group because the control group students were not taught programming (i.e., we do not know if the game outperformed a different supplemental program designed to teach programming).
- Students in the control group had more time learning math than the treatment group for two main reasons. First, they were learning only one subject. Forty seven percent of all students

had no programming experience and 70 percent of students had never programmed or tried programming only once or twice prior to the study. Second, students use MathXL as their regular in-school supplemental intervention and did not need to spend time learning the features of the program.

An important reason for selecting a well-regarded and well-established conventional math supplemental program was to evaluate the efficacy of very different approaches to teaching and learning. With the control intervention, the teaching is explicit with an emphasis on repeated practice of specific math skills through the provision of a wealth of questions and answers. “In contrast, in The Mars Game, the instructional concepts are subtly embedded in the gameplay and learning is intended to take place within the context of a 3D simulated environment with a compelling story narrative. The Mars Game prototype provides an experience (as opposed to direct instruction) that is designed to induce students to solve complex but manageable math problems” (Dill, Freeman, Frazier, & Benito, 2015, p. 10).

3.6 Measures and Instruments

Measures and Instruments Learning Outcomes: Mathematics and Programming Outcomes Assessment. A pre-post assessment was used to measure the effect of the intervention on learning outcomes (Appendix C). Questions were aligned to CCSSM and were also aligned to Webb’s (2002) Depths of Knowledge, in order that questions were balanced and reflective of different depths of knowledge (e.g., declarative, procedural, problem-solving) and different depths of cognitive demand, and difficulty. The test therefore included multiple choice questions, and constructed response and short answer response questions that require students to write or draw their answers to explain their reasoning. The purpose of item difficulty modeling is to determine “whether the processes measured by the items and tests are those that were intended by the researcher” (Gorin, 2006). Though this process is typically used in large-scale test design, the factors that make it useful are applicable in smaller scale test designs, like this one. As part of this process, the pretest and posttest items were vetted, tested and retested by mathematics and programming experts to establish content validity (Appendix G, example of the item development review).

Scores on the overall Learning pre-posttest range from 0-21, with the Math Only questions ranging from 0-17 and the Programming Only questions ranging from 0-4. The pretest and posttest differed very slightly and were labelled Form A and Form B; Form A was delivered to half the treatment and half the control students as the pretest and Form B to half the treatment and half the control students as the posttest. The same procedure was followed for the posttest. (Appendix B shows one example of each from).

Engagement. Fu, Su, and Yu’s (2009) tests of the EGameFlow Scale were used to assess engagement with learning (Appendix E). EGame Flow contains 42 items, across eight factors: Concentration (6 items), Goal Clarity (4 items), Feedback (5 items), Challenge (6 items), Autonomy (3 items), Immersion (7 items), Social Interaction (6 items), and Knowledge Improvement (5 items). The maximum point score across the eight dimensions is 294. Due to the nature of the questions addressed by the scale, it is appropriate for the testing of both the treatment and control group interventions. This is largely because (a) MathXL is an interactive program with

adaptive game-like elements and (b) the EGameFlow questions address issues related to presentation of goals, feedback on progress, relationship between activities and objectives, hints to overcome challenges, and so forth – factors that are relevant to both interventions.

The EGameFlow Scale has demonstrated high internal consistency and reliability, with a Cronbach's alpha of 0.942 for the items as a whole and > 0.8 for each distinct subgroup (Fu, Su, Yu, 2009). Using the data from our first summative study, we conducted a confirmatory factor analysis (CFA) to determine how the eight sub-factors loaded on the latent construct engagement. Our results revalidated Fu et al.'s results. All eight sub-factors showed significant influence on the construct engagement, and model fit indices indicate reasonable fit (CFI = 0.95, RMSEA = .11; see Appendix H).

Math Perceptions. In order to assess students' perceptions in math, we used a Student Self-assessment, which has been validated and used in prior studies (Lawyer-Brook, 2009). The assessment has nine questions (Appendix D). We used items 1 and 2, which asked students to rate themselves on a 5-point Likert scale as to how they feel about math. Question 1 asked students to fill in the blank, "I think math is ____." The responses were, "*Horrible, Boring, OK, Kind of interesting, Lots of fun*". Question 2 asked students to fill in the blank, "I am ____ at math." The responses were, "*Awful, Not very good, OK, Pretty good, Very good*".

Data Capture. Instrumentation was built into The Mars Game and stored in a central database, which enabled the tracking and monitoring of each student's performance. Data captured by the system include: the scenario and level reached, the number of attempts and amount of time spent on each level, time-on-task, periods of idle time, and which hints were provided.

3.7 Data Analysis

Initially, data were cleaned and checked for normality, and descriptive statistics were then calculated. Independent samples *t*-tests were conducted to look at differences between the treatment and control conditions on the pretest.

Since the dependent variable is the posttest, we used the related pretest as a covariate in all analyses of variance. The pretest was used as a covariate for the following reasons. First, due to the similarity of the pre- and posttests, the pretest accounts for a portion of variance that is related to the dependent variable. Second, the use of a covariate reduced possible bias and provided added control against error, and a baseline from which the posttest could be compared. According to Stevens (2002), the purpose of using covariates is to reduce systematic bias and within-group error of variance. Third, it is common practice in game-related studies to use pretest covariates (for example, see Ke, 2008; Riconscente, 2013).

A range of statistical techniques were used to address the research questions (RQ). Table 3 shows the relationship between each of the four research questions to the statistical technique. For clarity, we describe RQ4 in two parts: (4a) characteristics and (4b) perceptions. The techniques are explained (beneath the table) by grouping together research questions that logically fit together and share underlying assumptions (e.g., use of covariate, reported effect size), rather than in order of the research question.

Table 3*Statistical Technique used to address each Research Question*

Research Question	Statistical technique	Dependent Variable	Covariate	Independent Variables
RQ 1: Learning Outcomes	ANCOVA	Overall posttest Math posttest Programming posttest	Overall posttest Math posttest Programming posttest	Treatment vs Control
RQ 2: Engagement	Independent Samples <i>t</i> -test	Engagement Sub-factors	n/a	Treatment vs Control
RQ3a: Single/Pair Learning	ANCOVA Bootstrap	Overall posttest Math posttest Programming posttest	Overall posttest Math posttest Programming posttest	4 Conditions (Singles and pairs)
RQ3b: Single/Pair Engagement	Kruskal-Wallis non-parametric test	Overall posttest Math posttest Programming posttest	n/a	4 Conditions (Singles and pairs)
RQ4a: Characteristics and Learning & Engagement	MANCOVA	Posttest Engagement	Overall pretest	Treatment vs Control Gender Grade
RQ4b: Perceptions and Learning & Engagement	Spearman's correlation coefficient	n/a	n/a	Survey Questions

Research questions 1, 3a, and 4a, all required the use of a covariate. For research question 1, an analysis of covariance (ANCOVA) was used to investigate the impact of the interventions on learning outcomes. For research questions 3a, an ANCOVA was also used, however when the treatment and control groups were divided into four groups the size of each group is too small to apply a standard parametric analysis, so we conducted an ANCOVA bootstrap. Use of the bootstrap method was applicable in this instance because the normality of the data permitted us to assume that the sampling distribution is also normal (Field, 2009). Furthermore, the distribution of the population in the four conditions is the same as the original population (Efron & Tibshirani, 1993). For research question 4a, a multivariate analysis of covariance (MANCOVA) was conducted. Multivariate analysis enabled us to investigate the overall effects of The Mars Game on both learning and engagement whilst simultaneously accounting for differences in subgroups. The independent variables included gender and grade.

Partial eta-squared is the reported effect size for all analyses of variance (RQ 1, 3a, and 4a). Miles and Shevlin's (2001) rules of thumb were used to estimate the magnitude of the effect size (which differ for ANCOVA and MANCOVA analyses). Since we chose univariate analysis and multivariate analysis over pairwise *t*-tests for example, it was necessary to conduct post hoc tests, as such tests consist of pairwise comparisons. Post hoc analyses enabled us to examine mean differences. We chose to use the Bonferroni correction post hoc for two main reasons. First, the Bonferroni test statistic is used to compare non-orthogonal means, and is commonly used in post hoc analyses comparing multiple means (Pedhazur & Schmelkin, 1991). Second, since we had a

good sample size relative to the number of variables in the study for the main treatment-control comparison, we were not concerned about the possible loss of power and opted for a more conservative post hoc analysis (Field, 2009).

To answer Research Question 2, independent samples *t*-tests were conducted to determine the differences in engagement by condition and level of math class. The sub-factors of engagement were also analyzed (see Measurements and Instruments section). Cohen's *d* is the reported effect size for all independent samples *t*-tests (Cohen, 1988).

A non-parametric test was conducted for Research Question 3b, because when we investigated the four conditions the size of the sample in each group fell below 30, the standard rule of thumb for parametric tests (Tanis & Hogg, 2008). We used the Kruskal-Wallis test, which is the parallel test to the one-way independent ANOVA (Field, 2009).

For Research Question 4b, a Spearman correlation was used to look at the relationship between perceptions and learning outcomes and perceptions and engagement. Our hypothesis about these relationships were both directional, therefore we used a one-tailed test (Field, 2009).

4. RESULTS

Pre-posttest Summary. Independent samples *t*-tests were conducted to look at differences between the treatment and control conditions on the pretest. The Levene's test was not significant for the overall learning pretest ($F_{(1,78)} = 1.57, p > .05$), nor was it significant for the mathematics or programming tests. Therefore, equality of variances were assumed and descriptive statistics were calculated. Table 4 provides a summary of the pre-post scores for Research Questions 1 and 2.

Table 4

Summary of Pre-Post Scores for Research Questions 1 and 2

Condition		Pretotal	Posttotal	Pre-post MD	Engagement
Mars Single	<i>M</i>	4.81	6.67	1.86	193.71
	<i>SD</i>	2.25	2.71		45.10
Mars Pair	<i>M</i>	3.78	5.17	2.19	186.95
	<i>SD</i>	2.71	2.60		53.03
Control Single	<i>M</i>	4.52	6.71	1.39	154.06
	<i>SD</i>	3.14	3.10		57.03
Control Pair	<i>M</i>	6.25	6.80	.55	153.65
	<i>SD</i>	3.58	4.57		36.79

Effect of The Mars Game on Learning Outcomes. Research Question 1 was: *What is the effect of The Mars Game on high school students' knowledge and skills in mathematics and programming?* A significant effect favoring the treatment over the control group's learning outcomes was found on the overall posttest, ($F_{(1,79)} = 4.17, p < .05, \eta_p^2 = .05$). Post hoc analyses showed the treatment group significantly outperformed the control group ($MD = 1.01$), denoting that the Mars game prototype had a small-to-medium effect on students' overall learning outcomes.

Although The Mars Game intervention outperformed the control group ($MD = 0.22$) on the mathematics posttest, this effect was not significant ($F_{(1,79)} = .24$, NS). The Mars Game did, however, have a significant effect for the treatment group over the control group on the programming posttest, ($F_{(1,79)} = 13.17$, $p < .001$, $\eta_p^2 = .15$). The post hoc analysis established that the treatment group significantly outperformed the control group on the programming posttest ($MD = 0.83$), showing a large effect on learning outcomes.

Effect of The Mars Game on Engagement. Research Question 2 was: *What is the effect of The Mars Game on high school students' engagement with STEM subjects?* On average, participants' scores on engagement were significantly greater for the treatment group ($M = 190.33$, $SD = 49.00$) than for the control group ($M = 153.84$, $SD = 46.73$), $t_{(78)} = 3.40$, $p = .001$. The results signify a robust relationship between *engagement* and The Mars Game, with a high effect size ($d = .76$) favoring treatment over control. The mean score was higher for the treatment than the control group on each of the sub-factors that comprise the engagement construct. The effect size was large for the feedback ($d = .91$), immersion ($d = .69$), and knowledge improvement ($d = .70$) sub-factors. Table 5 displays the findings of the independent samples t -tests.

Table 5
Independent Samples t -tests on Engagement Sub-factors

Measure	t	p	Cohen's d	MD
Engagement Overall	3.40	.00***	.76	T>C, 36.49
Concentration	1.70	.09	.38	T>C, 3.04
Goal Clarity	2.54	.01**	.57	T>C, 3.32
Feedback	4.03	.00***	.91	T>C, 6.43
Challenge	1.72	.09	.39	T>C, 3.11
Autonomy	2.49	.02*	.56	T>C 2.79
Immersion	3.08	.00***	.69	T>C, 6.90
Social Interaction	2.62	.01**	.59	T>C, 5.22
Knowledge Improvement	3.13	.00***	.70	T>C, 5.74

Notes: Score Maximums: Engagement: 294, Concentration: 42, Goal Clarity: 28, Feedback: 35, Challenge: 42, Autonomy: 21, Immersion: 48, Social Interaction: 42, Knowledge Improvement: 35.

*** $p < .001$; ** $p < .01$; * $p < .05$

Results for Single versus Randomized Pairs on Learning Outcomes. Research Question 3a was: *What is the effect of individual gameplay versus randomized pairs on learning outcomes in STEM subjects?* Employing an ANCOVA bootstrap, a significant effect for post-programming outcomes ($F_{(3,79)} = 4.48$, $p < .01$, $\eta_p^2 = .15$), with the greatest difference stemming between The Mars Game Single condition and the Control Pair condition, as shown in Table 6. Interestingly, no significant difference was found between singles and pairs when the treatment and control conditions were ignored; that is, when Treatment Single and Control Single were added and compared to Treatment Pair and Control Pair, ($F_{(1,79)} = .02$, NS). No significant difference was found on the overall posttest ($F_{(3,79)} = 1.55$, NS) or the math posttest, ($F_{(3,79)} = .26$, NS).

Table 6*Results of Programming Posttest – Single versus Randomized Pairs*

Condition	N	M	SE	MD**
Mars Single	21	1.898	.223	----
Mars Pair	21	1.738	.223	.160
Control Single	18	1.047	.244	.852
Control Pair	20	.940	.32	.958

Notes: * Covariate at pretest = .89. ** Bootstrap pairwise comparisons based on 1,000 samples

*** Mean difference compared to the Mars Single condition; Bonferroni adjustment

Results for Single versus Randomized Pairs on Engagement. Research Question 3b was: *What is the effect of individual gameplay versus randomized pairs on engagement with STEM subjects?* Engagement was found to be significantly affected by condition, ($H(3) = 10.76, p < .05$). As with learning outcomes, no significant difference was found between singles and pairs when the treatment and control conditions were ignored; that is, when Treatment Single and Control Single were added and compared to Treatment Pair and Control Pair, ($T(1,78) = .41, NS$). When viewed by the four conditions, significance was found for four of the eight sub-factors of engagement, including: Feedback, ($H(3) = 13.97, p < .01$), Immersion, ($H(3) = 11.95, p < .01$), Knowledge Improvement, ($H(3) = 10.76, p < .05$), and Autonomy, ($H(3) = 18.26, p < .05$). Table 7 shows the mean rank of the engagement posttest for all conditions based on the Kruskal-Wallis non-parametric test. The Mars Single condition ranked the highest on engagement and the Control Pair condition ranked the lowest. The difference observed stems from assignment to condition (treatment versus control), rather than from assignment to individual versus collaborative play.

Table 7*Results of Engagement Posttest – Single versus Randomized Pairs*

Condition	N	Mean Rank
Mars Single	21	49.90
Mars Pair	21	47.10
Control Single	18	33.03
Control Pair	20	20.43

Summary of Individual Play versus Pairs. Assignment to the treatment group condition mattered more than assignment to individual work or randomized pairs for both learning outcomes and engagement. The generally accepted idea that the synergy found in pair work and cooperative learning that leads to greater achievement was not manifested in the context of this study (Johnson, Johnson, Stanne, 2000). There are several possible explanations for these findings. It could be because the game was not designed as a multi-player game, and therefore collaboration was not structured into the experience. Another possibility is that students perform better when pairs are carefully chosen to complement one another (there are a variety of ways to do this). Alternatively, on the opposite end of the scale, it could be that with games, it is better when students naturally find their own partners. This would require further investigation to better understand this result.

Characteristics and the Relationship between Learning and Engagement. Research Question 4a was: *In what ways do students' characteristics and perceptions influence the relationship*

between learning and engagement? In this section, we looked at just the influence of students' characteristics.

To reiterate, from Research Question 1, a significant effect was found for the treatment condition on overall learning. As to be expected, when we conducted a MANCOVA, with the learning posttest and engagement posttest as the two dependent variables, a significant main effect on learning was again identified, ($F_{(1, 79)} = 7.10, p < .01, \eta_p^2 = .10$), as was a significant main effect on engagement, ($F_{(1, 79)} = 4.34, p < .05, \eta_p^2 = .06$). The post hoc showed the treatment group exceeding the control group, with a mean difference of 1.98 on learning outcomes and 36.19 on engagement. The Mars Game had a moderate effect on students' learning outcomes and a small-to-medium impact on their level of engagement.

When looking at subgroups, a significant main effect was found for grade on the overall learning outcomes posttest ($F_{(1, 79)} = 6.36, p < .001, \eta_p^2 = .16$). The post hoc analysis with respect to learning outcomes favored the 9th grade over all other grades, with the greatest difference in pretest to posttest gains occurring between the 9th and 11th/12th grade students ($MD = 2.96$). There may be a slight distortion because of the uneven number of participants in each group (see Table 1); however, the result points strongly in this direction. It should be noted that there was no significant difference between grades at pretest ($F_{(1, 79)} = .04, NS$).

A significant main effect for gender on the learning outcomes posttest was not found ($F_{(1, 79)} = .436, NS$), indicating that both boys and girls gained math and programming knowledge equally from The Mars Game. There was no significant effect for engagement either ($F_{(1, 79)} = .16, NS$), implying that boys and girls were equally engaged.

No interaction effect was found between any of the independent variable combinations, including treatment/control and gender, treatment/control/ and grade, or gender and grade.

A summary of the main effects of the multivariate analysis is provided in Table 8.

Table 8

Summary of the Main Effects of the Multivariate Analysis

Grouping Variable	Dependent Variable	<i>F</i>	<i>p</i>	η_p^2	Post hoc <i>MD</i>
Treatment/Control	Posttest	7.10	.01**	.10	T>C = 1.98
Treatment/Control	Engagement	4.34	.04*	.06	T>C = 36.19
Gender	Posttest	.436	.511	.05	M>F = -0.50
Gender	Engagement	0.16	.690	.00	M>F = -7.10
Grade	Posttest	6.36	.00***	.16	9>10 = 1.42
					9>11/12 = 2.96
Grade	Engagement	.420	.658	.01	9>10 = -1.06
					9>11/12 = 20.41

Note. *** $p < .001$; ** $p < .01$; * $p < .05$

Perceptions and the Relationship between Learning and Engagement. Research Question 4b was: *In what ways do students' characteristics and perceptions influence the relationship between*

learning and engagement? In this section, we looked at only the influence of students' perceptions.

Students' response in both groups to the question about their interest in math "I think math is ____" (the responses were, "*Horrible, Boring, OK, Kind of interesting, Lots of fun*") was significantly related to how engaged they were with their respective intervention, ($r_s = .22$, (one-tailed) $< .05$). The more "horrible" they found math, the greater their engagement with the intervention.

Data Capture. Figure 1 shows a snapshot of The Mars Game student participant's time-on-task, as captured by The Mars Game central database. Most importantly, Figure 1 shows that most students did not make it through all of the game's missions. Most students made it through Mission 1, which primarily teaches how to play the game, how to manipulate Blockly, and other introductory programming skills. About half the students were able to get deep into Mission 2, which teaches slope intercept form and coordinate plane. Very few players reached Mission 3 (teaches loops and conditioning and coordinate pairs) or Mission 4 (teaches triangles and deepens students' understanding of the coordinate plan). Other observations include:

- The vertical consistency (such as the reddish column down the last task in M1) indicates a level that students consistently took longer to solve. This denotes that students are experiencing a similar level of challenge to one another. It also indicates that the challenge itself falls within the right parameters of rigor (i.e., neither too easy nor too difficult) because it is not leaving a subset of students behind.
- From a development perspective, having this type of visualization of students' progress available throughout the development process would enable the development team to prioritize their efforts; for example, focus on tasks that were not expected to be hard, or on tasks that were expected to be harder than they turned out to be. In future development efforts, we will use this type of data to help guide our development efforts.
- Looking at the rightmost color in each row, and excluding a couple outliers, there tends to be more yellows than reds. This suggests that the players were progressing (especially if you look at the preceding colors per row) and would have continued to make it further into the game IF they had more time. In other words, most of the students did not get stuck to the point that they made no more progress (i.e., giving up on the task). Instead, the pattern suggests that given more time, they would have made more progress.
- In future development efforts, we will add a measure of frustration (to be defined) in addition to time-on-task, which we think will be a useful tool to help focus the development team's efforts at game improvement.
- It would be interesting to understand a student's "ideal" pattern from a development perspective. We speculate that an ideal pattern would be if the mission started out green, graduated to yellow (or even red), and then graduated back towards green. This would indicate that the challenges grew harder, but that the student grasped the concepts as they were exposed and practiced. Of course, some tasks would naturally take longer to work through than others, regardless of difficulty, but maybe that could be taken into account in the way we capture and visualize the data.

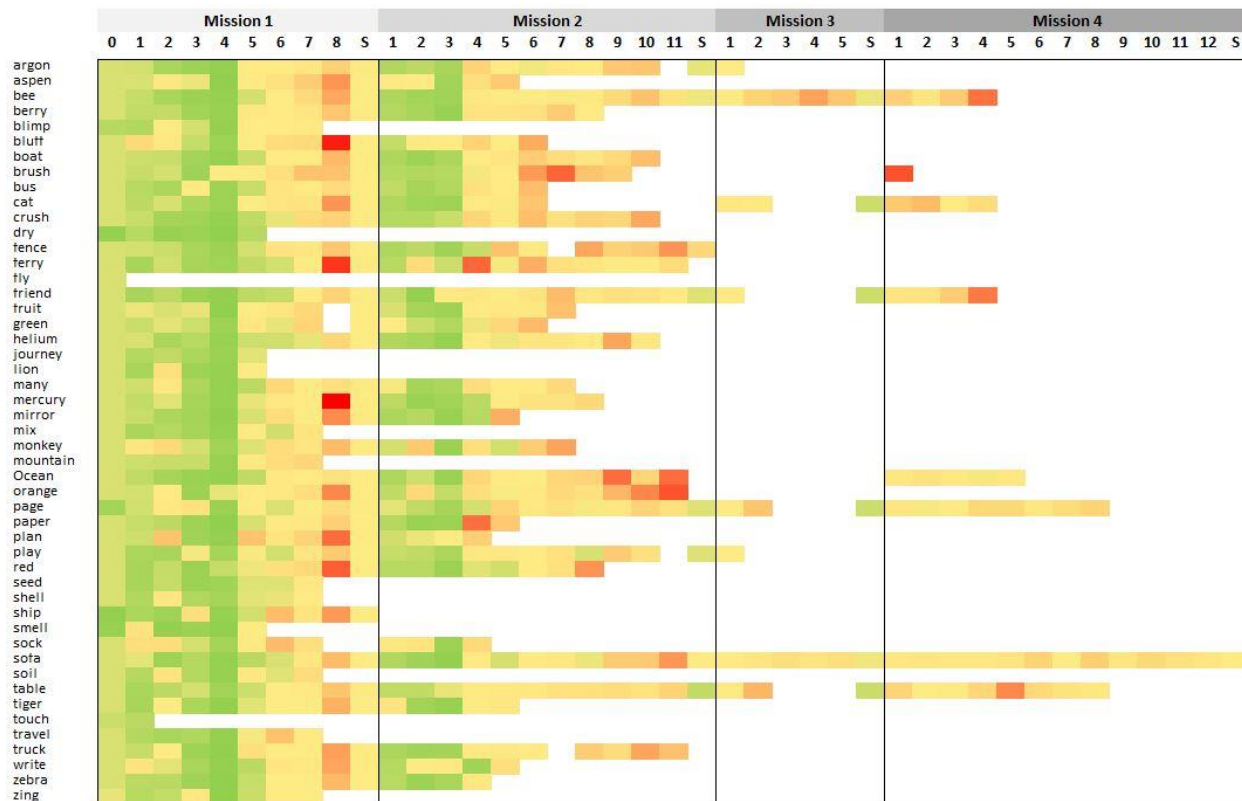


Figure 1: Time on Task

Notes. Green represents the lowest value, red represents the highest value, and yellow represents the 50th percentile, with “value” meaning the total time-on-task to complete each level, in seconds. The median is by percentile (not average), which smooths out the color gradation across the range. There is a column per task including the intro and success videos, and there is a row per student or pair.

5. MAJOR CONCLUSIONS AND IMPLICATIONS

In this section, we look at the study results with respect to our initial hypothesis and study goals. We then discuss the implications of our findings.

5.1 Hypothesis

These results support our hypothesis that an engaging design-led strategy will result in a STEM learning solution that is enjoyable to students, and will impart knowledge more effectively than traditional learning paradigms. Scholars have maintained that it is through students’ active engagement with intentionally designed immersive experiences and rich content that people learn (Gee, 2014; Virou, Katsionis, & Manos, 2005). A digital learning game therefore should engender a sense of enjoyment that captures students’ imagination and holds their interest (Ke, 2008; Papert, 1993; Virou, Katsionis, & Manos). Yet, the entertainment must not diminish the pedagogical rigor of the content (Dede, 2009), as the intensity of engagement is related to students’ interest and intimately intertwined with learning. Previous studies have shown that students can learn from games. This study is rare in that it used a randomized controlled trial to compare the effectiveness of the learning using a serious game to learning by a more traditional method, and thus contributed

to this body of literature.

This study also demonstrated that attaining engagement and rigor in a game-based solution is both possible and critical for learning (i.e., a “True Game” vs. “Chocolate Covered Broccoli”). The trade-off between engagement and learning is evidenced in the comparison of the results of the EGameFlow engagement posttest of the first summative study (Dill, Freeman, Frazier, & Benito, 2015) to the results from this current study. Although we found a significant difference in engagement favoring the treatment over the control group in both the first and second experimental studies, in the year 1 study, the sub-factor that was weakest was feedback ($d = .38$). During the second phase of this project, the game developers addressed this feature; accordingly, in the year 2 study, feedback was shown to have a large effect ($d = .91$). As the impact of feedback rose, however, the effect size of immersion—although still high—decreased from ($d = .97$) in year one to ($d = .69$) in year 2.

Our results were similar during our formative studies. We observed that whilst support and feedback were required to help several students reach the next level, the ability to reach the next level seemed also to be driven by students’ interest and motivation to actually want to undertake the next mission (Freeman, 2015). This motivation, in turn, is partially rooted in the “cognitive conflict”—the struggle to achieve a goal—inherent in the game’s challenge (Sanchez-Lozano, 2010, p. 93). Therefore, we wanted to make sure that students received the extra help they required whilst simultaneously ensuring that we did not diminish their immersion in the game’s challenges. Without the added feedback and scaffolding, we were concerned that students’ positive engagement could slip into frustration; nonetheless, we were concerned that too much support could convert the students’ engagement to boredom (D’Mello & Graesser, 2011; Freeman, 2015). A critical balance needs to be struck between providing sufficient support and meaningful feedback and yet not undermining the immersion and engagement in the game.

Finding this sweet spot between fun and pedagogical rigor has important implications for the design and development of digital learning games. In the papers, *Students crash-land on Mars* (Freeman, 2015) and *Interview with Barbara Freeman: The Mars Game* (Dudley, 2015), the role of design-based research toward helping to find this balance is discussed. In our manuscript, *Mars Game: Creating and Evaluating an Engaging Educational Game* (Dill, Freeman, Frazier, Benito, 2015), we describe key elements of The Mars Game design, and in so doing illuminate this tension between fun and rigor in order to steer future development of digital learning games.

5.2 Study Goals

The main goal of the study was to evaluate if the expanded version of The Mars Game was engaging and effective when evaluated under a different setting (high poverty school, afterschool) and conditions (played in pairs). To reiterate, The Mars Game had a significant effect for the treatment group over control group on engagement. The effect size was large for engagement and for the feedback, immersion, and knowledge improvement sub-factors of the construct. The study demonstrated that The Mars Game treatment group showed statistically significant improvements in learning outcomes when compared to the control group. The gains in achievement were driven largely through the increase in programming posttest scores compared to a control group baseline.

Students playing The Mars Game showed greater gains in math than their control group

counterparts, but these were not significant. Nonetheless, this finding is impressive for several reasons, which may also explain the lack of significance in math:

- The game is new and the control group used an established curriculum of the world's largest educational publisher.
- Students already knew how to use the control intervention, and students in the treatment group spent more time learning how to play the game.
- Many of the treatment students did not make it all the way through the game and therefore were not exposed to all of the math content (see Figure 1).
- Students in the control group were only practicing math whereas students in the treatment group were also learning to code; 70 percent of students had never programmed or tried it only once or twice prior to the study (as reported in the Student Self-Assessment).
- Any gains in programming came from interaction with The Mars Game because none of the students in the study were taking a programming class during the time of the study. The fact that there was a comparison group enables us to assert that the gains were not purely from guessing because one group's guess would not be better (statistically significantly) than another group's guess.

Furthermore, this study showed that The Mars Game worked similarly well in an afterschool setting as during the school day. In the first summative study, we found that of all sub-groups involved in the study, boys from Spanish-speaking households who played the game showed the greatest increase in learning and highest levels of engagement (Dill, Freeman, Frazier, & Benito, 2015). This second study substantiated this finding by showing that The Mars Game also successfully engaged students in a high-poverty, minority-majority school (70-30%), and it helped them learn more.

By demonstrating gains in learning and engagement, this evaluation reveals great promise in the pedagogical approach of game-based learning (i.e., stimulating students to learn through problem solving and experimentation, as opposed to direct instruction). In the context of serious games, problem solving is understood as a process of finding patterns and relationships between variables (Gee, 2007). This stands in stark contrast to more conventional ways of approaching problem solving, in which "knowledge is separated from skills and presented as revealed truth, not as an understanding that is discovered and constructed; this separation results in students learning data about a topic rather than learning how to extend their understanding beyond information available for assimilation" (Dede, 2009, p. 2).

Problems in games are solved by students staying involved and working their way through the tasks. Students who are immersed in a game often persist in order to complete a challenge. By doing so, they repeatedly practice the tasks, fail and try again. In the game development community, this is referred to as "failing better." During The Mars Game, students figure out problems for themselves, with scaffolds, hints and perhaps collaboration, but without anybody explicitly showing them how to achieve the mission (Freeman, 2015). In a well-designed learning experience, students apply their own adaptive reasoning to make sense of increasingly complex problems (Freeman, Higgins, & Horney, in review). This is a particularly critical point for students

who may be disengaged or feel disempowered with respect to achievement in STEM subjects.

Nonetheless, experimentation may take more time to achieve positive results than explicit instruction. This is evidenced by the difficulty students playing The Mars Game had in completing all levels of the game during the same time period that students in the control group completed their whole assignment; both The Mars Game and the control group's assignments covered the full content targeted in this study. If we wish to confirm this point about time and experimentation, however, we would need to conduct a different study in which students already know how to play the game in order to remove the learning curve for students playing the game, as it would be a confounding variable in this context.

5.3 Other Conclusions and Implications

Confirming the findings of the previous Mars Game experimental study, this current study demonstrated that well-designed and well-integrated serious games, which are not simply dressed up homework, have the ability to engage and increase the knowledge and skills of students in STEM subjects.

Ninth grade students, who were the main focus of this study, learned significantly more than students in the higher grades, although it should be noted that the numbers of students were not evenly distributed across grades (see Table 1). Nonetheless, The Mars Game intervention had the greatest success with students who are new to the content being taught. Again this finding was consistent with previous studies in which we found that the game had the greatest impact on this same group. These findings provide some evidence that games can be used as a primer to introduce new content – perhaps because of the confidence that some students gain as a result of the introduction of concepts and skills in a less threatening and more engaging way. Taken together, the learning, engagement, and perception outcomes indicate that The Mars Game has the potential to improve students' self-efficacy, which is important for STEM success. This substantiates Lieberman's (2012) findings, from multiple studies, which showed that students' actions and decision-making within a game environment can impact their self-efficacy. Immersive educational games provide players with choice and possibilities as well as access to the tools and information they need to manipulate objects and characters to make decisions within an immersive digital world (Lieberman). She found that this allows players to control their learning within the safety of the game.

It is also possible that digital learning games could be used as a primer to introduce content to students – because of the experience that students gain as they move from more tangible experience to more abstract understanding. In the context of The Mars Game, teaching mathematics and programming occurs concurrently. By doing so, the game is able to introduce abstract algebraic concepts unobtrusively by enabling students to experiment and discover using concrete manipulation (e.g., maneuvering Blockly). Dale (1969) maintained that learners are better prepared to process abstract information once they have accumulated concrete experiences, thus, acquiring the requisite background knowledge to be able to assign meaning to more “abstract representations of reality” (Heinich, Molenda, Russell & Smaldino, 2002, p.11).

Another notable finding is that boys and girls learned equally by using the game and were equally engaged. This too echoed our findings from the previous study, and thus provokes the question:

Are gender differences a thing of the past when it comes to gaming? Although this result may be partially explained by The Mars Game’s storyline, which is concerned with helping ‘cute’ rovers to survive on Mars, we are not alone in the general finding about lack of difference between the genders. According to an Entertainment Software Association 2015 report, of the 155 million US game players, 56% are male and 44% are female (ESA, 2015).

Lastly, the analysis demonstrated that digital learning games have the ability to improve the engagement of students who may have previously been switched off from learning STEM subjects. We observed that the lower students’ interest in the subject (e.g., the more “horrible” they found math), the higher their engagement. This suggests a potential to use The Mars Game to spark the interest and improve the learning experience of younger students in STEM content. And, importantly, to do so before failure sets in when some students become uninterested in STEM subjects or disengaged with learning. Thus, digital learning games provide an opportunity to re-imagine learning, and could be mindset-changing for many students.

6. TRUE GAME PROJECT STUDIES

We conducted five studies over two years: two formative, one online, and two summative. The various styles of study are useful and provided different information to inform development. As a result of this design-based research process, we believe that receiving a lot of feedback from different users, often, and throughout the entire process leads to a better experience for learners. The more feedback, the better. Early feedback tends to be shorter-lived but readily absorbed into game improvement, whilst later feedback tends to be harder to react to because there is less time left to modify the game. This project makes clear that data captured through the game’s instrumentation should also play a key role in guiding iterative development efforts.

This study shows the importance of repeated studies as a necessary component of any serious game development plan. The use of the same instruments (pre-posttests, surveys) from a representative sample (versus the same sample in a longitudinal study) provides an opportunity to evaluate learning in a variety of contexts, better understand the interrelationships between engagement, learning, and perceptions, and better grasp the dynamics of learning overtime (Laurie, 2013).

7. LIMITATIONS AND OPPORTUNITIES FOR FURTHER STUDY

One limitation of this study is its short length. It is possible that the high scores on engagement were in part due to the novelty of the game. Familiarity with the control group intervention may have dampened the novelty factor that was provided by the treatment group’s intervention. However, an offsetting factor is that the control group had more time learning math than the treatment group because they were being taught only one subject. Furthermore, because of familiarity with the control group’s supplemental program, students did not need to spend time learning how to use the intervention. They were able to get straight to the math practice.

A second possible limitation of the study is that we do not know if students’ learning gains are short-term or if the knowledge acquisition demonstrated during the study will endure for either the content or the learning process (e.g., practice, concentration). Interestingly, a comprehensive literature review of studies related to mathematics, technology, and affective characteristics (motivation, self-efficacy, and attitudes) shows that students show the highest levels of

achievement and learning with 6-10 weeks of intervention; however, their affective characteristics are more positive in the beginning (0-5 weeks of intervention) and after 11 weeks of intervention (Huscroft-D'Angelo, Higgins, & Crawford, 2015). Thus, conducting a longer study could allow for further assessment of long-term gains in both knowledge acquisition and engagement with the program. A longitudinal study would also enable us to assess whether digital learning games can increase self-efficacy, help at-risk students perform better in school, generate more students for STEM study in college and beyond.

During this study we identified a significant relationship between learning outcomes and engagement, however, we did not explore whether one mediates the other. Similarly, we identified a significant relationship between perceptions and engagement and learning outcomes, but did not explore the directionality of this relationship. It will be interesting to delve further into these relationships through path analysis and investigation of the sub-factors of engagement and the direct and indirect effects of engagement on learning outcomes.

Finally, we have an opportunity to continue with the panel study that we have already begun and evaluate the game again (perhaps with new features, new content, etc.), with a new representative sample of students in different settings (conceivably at scale).

REFERENCES

- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). New York: Academic Press.
- Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, 323, 66nce, doi: 10.1126/science.1167311
- Dale, E. (1969). *Audio-visual methods in teaching*. New York: Holt, Rinehart & Winston.
- Dill, K., Freeman, B., Frazier, S., & Benito, J. (2015). Mars Game: Creating and Evaluating an Engaging Educational Game (paper #15105). *Proceedings of the Interservice/Industry Training, Simulation and Education Conference 2015*.
- D'Mello, S., & Graesser, A. (2011). Emotions during learning with AutoTutor. In P. J. Durlach & A. M. Lesgold (Eds.), *Adaptive technologies for training and education* (Locations 64-65). Cambridge University Press. Kindle Edition.
- Dudley, A. (2015). Interview with Barbara Freeman: The Mars Game. *StemJobs*. Retrieved from: <http://stemjobs.com/interview-with-barbara-freeman/>
- ESA (2015, April). The 2015 Essential facts about the computer and video game industry. Retrieved from: <http://www.theesa.com/wp-content/uploads/2015/04/ESA-Essential-Facts-2015.pdf>
- Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: SAGE Publications.
- Freeman, B. (2015, Feb. 24). Students crash-land on Mars. *EdTechDigest*. Retrieved from: <https://edtechdigest.wordpress.com/2015/02/24/students-land-on-mars/>
- Freeman, B., Higgins, K., & Horney, M. (in review). Mathematical Communication and Writing

using Digital Technologies.

- Fu, F-L., Su, R-C., & Yu, S-C. (2009). EGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education*, 52(2009), pp. 101-112. doi: 10.1016/j.compedu.2008.07.004
- Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational research: An introduction*, (8th ed.). Boston, MA: Pearson Education.
- Gee, J. P. (2007). *What video games teach us about learning and literacy*. New York, NY: Palgrave, Macmillan.
- Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, Winter 2006, 21-35.
- Heinich, R., Molends, M., Russell, J. D., & Smaldino, S. (2002). *Instructional Media and Technologies for Learning*. Upper Saddle River, N.J.: Pearson Education, Inc.
- Huscroft-D'Angelo, J., Higgins, K., & Crawford, L. (2015). *Technology, affective characteristics and mathematics: A review of the literature from 1985-2013*. Manuscript submitted for publication.
- Johnson, D. W.; Johnson, R. T., Stanne, M. E. (2000). *Cooperative Learning Methods: A meta-analysis*. University of Minnesota, Minneapolis: Cooperative Learning Center.
- Ke, F. (2008). Computer games application within alternative classroom goal structures: cognitive, metacognitive, and affective evaluation. *Education Tech Research Dev*, 56, pp. 539-556. Doi: 10.1007/s11423-008-9086-5
- Laurie, H. (2013, August). Panel studies. *Oxford Bibliographic Online*. doi: 10.1093/obo/9780199756384-0108. Retrieved from: <http://www.oxfordbibliographies.com/view/document/obo-9780199756384/obo-9780199756384-0108.xml>
- Lawyer-Brook, D. (2009). Annual performance report: Help with English language proficiency. Ready to teach grant. Unpublished report submitted to U.S. Department of Education.
- Lieberman, D. A. (2012). Designing digital games, social media, and mobile technologies to motivate and support health behavior change. In R. E. Rice & C. K. Atkins (Eds.), *Public communication campaigns, 4th Edition*. Thousand Oaks, CA: Sage Publications, pp. 273-287.
- Miles, J. N. V., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. London: Sage.
- Papert, S. (1998). Does easy do it? Children, games and learning. *Game Developer*, 87-88.
- Pedhazur, E., & Schmelkin, L. (1991). *Measurement, design and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.
- Riconscente, M. M. (2013). Results from a controlled study of the iPad fractions game Motion Math. *SAGE*, 8(4), pp. 186-214. doi: 10.1117/1555412013496894
- Sanchez-Lozano, J. C. (2010). Exploratory digital games for advanced skills: Theory and application. In P. Zemliansky, & D. M. Wilcox (Eds.), *Design and implementation of educational games: theoretical and practical perspectives* (pp. 92-107). Hershey, PA:

Information Science Reference. doi: 10.4018/978-1-61520-781-7.ch007

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences (4th ed.)*. Hillsdale, NJ: Erlbaum.

Tanis, E. A. & Hogg, R. V. (2008). *A brief course in mathematical statistics*. Upper Saddle River, NJ: Pearson Prentice.

Virou, M., Katsionis, G., & Manos, K. (2005). Combining software games with education: evaluation of its educational effectiveness. *Educational Technology and Society*, 8(2), 54–65.

Webb, N. L. (2002). Depth-of-Knowledge levels for four content areas. Wisconsin Center for Education Research, Madison, WI. Retrieved from:
<http://facstaff.wcer.wisc.edu/normw/All%20content%20areas%20%20DOK%20levels%2032802.doc>