THE IMPACT OF VIRTUAL MATHEMATICS GAMES ON STUDENT ACHIEVEMENT AND ENGAGEMENT IN THE MATHEMATICS CLASSROOM

Amanda A. Peterson

Western Connecticut State University

BS Secondary Education Mathematics, Western Connecticut State University, 2009

BS History, Western Connecticut State University, 2009

MS Mathematics, Fairfield University, 2012

A Dissertation Proposal

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Education in Instructional Leadership

in the

Department of Education and Educational Psychology

at

Western Connecticut State University

2024

THE IMPACT OF VIRTUAL MATHEMATICS GAMES ON STUDENT ACHIEVEMENT AND ENGAGEMENT IN THE MATHEMATICS CLASSROOM

Abstract

The purpose of this study was to investigate the impact of playing a virtual mathematics game on students' mathematics achievement scores and student engagement regarding gender (male, female) and program type (game-based learning (GBL) enhanced mathematics program, traditional mathematics program). The current study took place at two urban middle schools within the same district. The researcher used a sample of convenience comprised of 431 seventh and eighth grade mathematics students taught by seven mathematics teachers. This quantitative study was quasi-experimental, with a treatment and a comparison group using intact classrooms of students. Student data were collected using five tools: a standardized mathematics assessment, student mathematics grades, student engagement in mathematics, demographic information, and data from the virtual game program. There was no statistically significant impact on student achievement scores. There was a significant interaction between engagement and gender with females who played the virtual mathematics game having significantly higher scores for the emotional subscale of engagement than males or females who did not play the game.

DEDICATION

To my sister Karie, who was always my biggest cheerleader and pushed me to be the best I could even if it was because I always thought we were in a competition.

APPROVAL PAGE

School of Professional Studies
Department of Education and Educational Psychology
Doctor of Education in Instructional Leadership

Doctor of Education Dissertation

THE IMPACT OF VIRTUAL MATHEMATICS GAMES ON STUDENT ACHIEVEMENT AND ENGAGEMENT IN THE MATHEMATICS CLASSROOM

Presented by

Amanda A. Peterson

Kristy L. Zaleta, EdD	Kristy L. Zaleta	June 13, 2024
Dissertation Committee Co-Chair	Signature	Date
Marcia A. B. Delcourt, PhD	Marcia A. B. Delcourt	June 13, 2024
Dissertation Committee Co-Chair	Signature	Date
Harry E. Rosvally, Jr., EdD	Harry E. Rosvally	June 13, 2024
Dissertation Committee Member	Signature	Date

2024

TABLE OF CONTENTS

Page

Chapter 1: Introduction to the Study 1

Rationale for Selecting the Topic 2

Statement of the Problem 4

Potential Benefits of Research and Significance of the Study 5

Definition of Key Terms 6

Summary of Chapter One 6

Chapter 2: Related Literature 7

Theory 8

Engagement 8

Activity Theory 12

Related Research 15

Game-Based Learning Studies in Mathematics 15

Game-Based Learning Across Subject Area 26

Examining Gamification and Game-Based Research 29

Conclusion 32

Chapter 3: Methodology 34

Description of the Setting, Sampling Procedures, and Participants 34

Selection and Description of the Setting 34

TABLE OF CONTENTS

Sampling Procedures for Particpants 36

Adult Participation 36

Student participants 39

Research Question and Hypotheses 41

Description of the Research Design 43

Research Question One 44

Research Question Two 44

Research Question Three 44

Description of the Instruments and Data Collection Tools 45

Student Demographic Survey 45

Teacher Demographic Survey 45

STAR Math Assessment 46

Mathematics Grades 47

Student Engagement in Mathematics Scale 48

Description and Justification of the Analyses 48

Research Question One 49

Research Question Two 49

Research Question Three 50

Description of the Activities 50

TABLE OF CONTENTS

Treatment Group Activities 50

Comparison Group Activities 51

Posttest Administration 51

Description of Data Collection Procedures and Timeline 51

Ethics Statement 52

Chapter 4: Data Analysis 54

Research Questions and Hypotheses 54

Description of the Data 56

Research Question One 56

Research Question Two 56

Research Question Three 56

Data Coding and Entry 58

Data Screening and Cleaning 61

Demographic Results 63

Analysis of Outliers 65

Descriptive Statistics 70

Research Question One 72

Testing the Assumptions for Pre-Data 72

Normality 73

TABLE OF CONTENTS

Linearity 73

Homoscedasticity 74

Revised Normality 75

Revise Linearity 81

Revised Homoscedasticity 84

Results for Pre-Data MANOVA 85

Testing the Assumptions for Post-Data 88

Normality 89

Linearity 94

Homoscedasticity 97

Results for Research Question One 98

Research Question Two 102

Descriptive Statistics of Data 102

Testing the Assumptions 106

Normality 106

Multivariate Outliers 107

Linearity 109

Results for Research Question Two 110

Research Question Three 112

TABLE OF CONTENTS

Data Cleaning 112

Descriptive Statistics 114

Testing the Assumptions 115

Normality 115

Linearity 121

Homoscedasticity 125

Results for Research Question Three 125

Summary of Chapter Four 129

Chapter 5: Summary and Conclusions 132

Overview and Discussion of the Study 132

Research Questions 132

Research Question One 132

Results 133

Limitations 134

Relationship to theoretical framework 134

Related Research 134

Implications for Education 134

Potential Future Research 135

Research Question Two 135

Results 135

TABLE OF CONTENTS

Limitations 136

Relationship to the theoretical foundation 136

Related Research 136

Implications for Education 137

Potential Future Research 137

Research Question Three 137

Results 138

Limitations 139

Relationship to theoretical foundation 139

Related Research 139

Implications for Education 139

Potential Future Research 140

Study Validity 140

Internal Validity 140

External Validity 143

Conclusion 144

References 146

APPENDICES

Appendix A:	IRB Approval Letter	153
Appendix B:	Superintendent Consent Form	154
Appendix C:	Principal Consent Form	156
Appendix D:	Teacher Consent Form	158
Appendix E:	Parent Consent Form	160
Appendix F:	Student Consent Form	162
Appendix G:	Teacher Demographic Survey	163
Appendix H:	Student Demographic Survey	165

TABLES

Table 1	Studies Related to Game-Based Learning in Mathematics Classrooms	17
Table 2	Studies Related to Game-Based Learning Across Subject Areas	26
Table 3	Examining Gamification and Game-Based Learning Research	29
Table 4	Summary of Game-Based Learning-Related Research	33
Table 5	Racial/Ethnic Status for Student Population for District and Participating Middle Schools	36
Table 6	Demographic Characteristics of Participating Teachers	38
Table 7	Student Participants per Teacher based on Consent Received	40
Table 8	Student Participants per Teacher by Group	41
Table 9	Dependent and Independent Variables Based on Research Question	57
Table 10	SPSS Codebook of Student Demographic Variables	60
Table 11	SPSS Codebook of Independent and Dependent Variables	61

Table 12	Frequency of Participants Who Completed the Study	62
Table 13	Gender and Grade of Student Participants between Comparison and Treatment Groups	64
Table 14	Ethnicity of Treatment and Comparison Group Student Participants	65
Table 15	Investigation of Potential Outliers for Research Question One	66
Table 16	Gender of Student Participants in Treatment and Comparison Groups after Data Cleaning	68
Table 17	Skewness and Kurtosis Values by Gender	68
Table 18	Skewness and Kurtosis Values by Program Type	69
Table 19	Tests of Normality Achievement by Gender	69
Table 20	Tests of Normality Achievement by Program Type	70
Table 21	Pretest Data for Program Type (n = 205)	71
Table 22	Pretest Data for Gender (n = 205)	71
Table 23	Posttest Data for Program Type (n = 205)	71
Table 24	Posttest Data for Gender (n = 205)	72
Table 25	Correlations by Gender Pretest Data (n = 205)	73
Table 26	Correlations by Program Type Pretest Data (n = 205)	74
Table 27	Summary of Correlations for all Pretest Data	74
Table 28	Summary of Comparison Means for Pretest STAR Scores with equal gender (n = 200)	75
Table 29	Skewness and Kurtosis Values by Gender Pretest Data (n = 200)	76
Table 30	Skewness and Kurtosis Values by Group Pretest Data (n = 200)	76
Table 31	Tests of Normality Pretest by Program Type (n = 200)	77

Table 32	Tests of Normality Pre-Test by Gender (*n* = 200)	77
Table 33	Revised Correlations by Pretest Data (*n* = 200)	82
Table 34	Revised Correlations by Program Type Pretest Data (*n* = 200)	82
Table 35	Summary of Revised Correlations for all Pretest Data *(n = 200)*	82
Table 36	Descriptive Statistics for the Pretest Scores	86
Table 37	Multivariate Tests for the Pretest Scores	87
Table 38	Two-Way MANOVA Results for Mean Pretest Scores	88
Table 39	Skewness and Kurtosis Values by Program Type for Posttest Scores	89
Table 40	Skewness and Kurtosis Values by Gender for Posttest Scores	89
Table 41	Tests of Normality Posttest Scores by Program Type (*n* = 200)	90
Table 42	Tests of Normality Posttest Scores by Gender (*n* = 200)	90
Table 43	Correlations by Gender for Posttest (*n* = 200)	95
Table 44	Correlations by Program Type for Posttest (*n* = 200)	95
Table 45	Summary of Correlations for all Posttest Data (n = 200)	95
Table 46	Descriptive Statistics for the Posttest Scores	99
Table 47	Multivariate Tests for the Posttest Scores	100
Table 48	Follow-up Results for Mean Posttest Scores	101
Table 49	Estimated Marginal Means for Covariate for Posttest Scores	101
Table 50	Investigation of Potential Outliers for Research Question Two	103
Table 51	Descriptive Statistics for Posttest of Independent Variables with Respect to Each Dependent Variable for Multiple Regression	105
Table 52	Skewness and Kurtosis for Achievement Regression Dependent Variables	106
Table 53	Tests of Normality for Achievement Regression Dependent Variables	107

Table 54	Mahalanobis Distance for Regression Assumptions	108
Table 55	Correlations for Regression Assumptions for Achievement Regression	109
Table 56	Model Summary for Achievement Linear Regression	111
Table 57	ANOVA Results for Achievement Linear Regression	111
Table 58	Analyzing Outliers for Engagement Analysis	113
Table 59	Gender of Student Participants in Treatment and Comparison Groups after Data Cleaning for SEMS	113
Table 60	Descriptive Statistics for Posttest of Independent Variables with Respect to Each Dependent Variable for SEMS	114
Table 61	Skewness and Kurtosis Values by Program Type for Posttest Scores for Engagement	115
Table 62	Skewness and Kurtosis Values by Gender for Posttest Scores for Engagement	115
Table 63	Tests of Normality Posttest Engagement Scores by Program Type (*n* = 117)	116
Table 64	Tests of Normality Posttest Engagement Scores by Gender (*n* = 117)	116
Table 65	Descriptive Statistics for Posttest of Independent variables with Respect to Each Dependent Variable (social and emotional) for SEMS	117
Table 66	Summary of Correlations for Engagement by Groups (Social and Emotional)	124
Table 67	Summary of Correlations for Engagement by Gender (Social and Emotional)	124

Table 68 Summary of Correlations for all variables (social, emotional, gender and program type) 124

Table 69 Levene's Test of Equality or Error Variances for Engagement 125

Table 70 Multivariate Test for Engagement 127

Table 71 Tests of Between-Subject Effects for Engagement 128

FIGURES

Figure 1	Comparison-Group Design Research Question 1	44
Figure 2	Design Research Question 2	44
Figure 3	Static-Group Comparison Design Research Question 3	45
Figure 4	Histogram of the mean pretest Winter STAR scores for treatment group	78
Figure 5	Histogram of the mean pretest Winter STAR scores for comparison group	78
Figure 6	Histogram of the mean pretest Winter STAR scores for males	79
Figure 7	Histogram of the mean pretest Winter STAR scores for females	79
Figure 8	Histogram of the mean pretest T2 Grade for treatment group	80
Figure 9	Histogram of the mean pretest T2 Grades for comparison group	80
Figure 10	Histogram of the mean pretest T2 Grades for males	81
Figure 11	Histogram of the mean pretest T2 Grade for females	81
Figure 12	Scatterplot for pretest scores for treatment group	83
Figure 13	Scatterplot for pretest scores for comparison group	83
Figure 14	Scatterplot for pretest scores for males	84
Figure 15	Scatterplot for pretest scores for females	84
Figure 16	Histogram of the mean posttest Spring STAR scores for treatment group	91
Figure 17	Histogram of the mean posttest Spring STAR scores for comparison group	91
Figure 18	Histogram of the mean posttest Spring STAR scores for males	92
Figure 19	Histogram of the mean posttest Spring STAR scores for females	92
Figure 20	Histogram of the mean posttest T3 Grades for treatment group	93
Figure 21	Histogram of the mean posttest T3 Grades for comparison group	93

Figure 22	Histogram of the mean posttest T3 Grades for males	94
Figure 23	Histogram of the mean posttest T3 Grades for females	94
Figure 24	Scatterplot for posttest scores for treatment group	96
Figure 25	Scatterplot for posttest scores for comparison group	96
Figure 26	Scatterplot for posttest scores for males	97
Figure 27	Scatterplot for posttest scores for females	97
Figure 28	Box and Whisker Plot After Three Rounds of Cleaning Data	104
Figure 29	Scatterplot for multivariate outliers	109
Figure 30	Scatterplot for variables in multiple regression for gender	110
Figure 31	Histogram of the mean posttest scores of the Emotional for treatment group	118
Figure 32	Histogram of the mean posttest scores of the Emotional for comparison group	118
Figure 33	Histogram of the mean posttest scores of the Emotional for males	119
Figure 34	Histogram of the mean posttest scores of the Emotional for females	119
Figure 35	Histogram of the mean posttest scores of the Social for treatment group	120
Figure 36	Histogram of the mean posttest scores of the Social for comparison group	120
Figure 37	Histogram of the mean posttest scores of the Social for males	121
Figure 38	Histogram of the mean posttest scores of the Social for females	121
Figure 39	Scatterplot for posttest score for treatment group for engagement (SEMS)	122
Figure 40	Scatterplot for posttest score for comparison group for engagement (SEMS)	122
Figure 41	Scatterplot for posttest score for males for engagement	123

Figure 42 Scatterplot for posttest score for females for engagement 123

Figure 43 Graph of the posttest means for Emotional scores for the interaction of gender and program type 129

CHAPTER 1: INTRODUCTION TO THE STUDY

Gamification is the process of using game mechanics, in a classroom, school, workplace, etc., to engage and motivate people to achieve their goals (Burke, 2014). In turn, game-based learning involves using video games to support teaching and learning (Anastasiadis, Lampropoulos, & Siakas, 2018; Perrotta et al., 2013). There appears to be "a positive relationship between gaming in the classroom, learning outcomes and motivation and engagement" (Perrotta, Featherstone, Aston, & Houghton, 2013, p. 19).

Experts have harnessed gamification and game-based learning (GBL) in multiple subject areas to increase student engagement in classes. As Dicheva, Dichev, Agre, and Angelova (2015) stated, engagement is missing from many mathematics classrooms across various grade levels. According to Burke (2014), engagement happens at an emotional level because students are more invested in reaching the next level when they are striving to win at a game. This motivation occurs because gamifying a classroom activity encourages participants to achieve the goals for which there are rewards (Burke, 2014). Both "motivation" and "engagement" are related constructs; as such, some scholars use the words interchangeably (Reschly & Christenson, 2012, p. 3). In this study, the researcher primarily uses the term engagement. Liu and Chu (2010) stated that game-based learning combines learning and game playing, hypothesizing that GBL will improve the ability of the player to retain educational knowledge and apply it to the real world. Games can introduce and teach specific skills in a user-friendly environment "where students are able to play, try, make mistakes and learn" (Hwa, 2018, p. 265). By using a virtual game program, students will be able to get personalized intervention based on their level of participation in the game (Buckley et al., 2016).

As Homer, Raffaele, and Henderson (2019) state, "By applying the concept of playful learning, we can realize educational experiences in digital games that may go beyond even what was envisioned by Vygotsky and Piaget in supporting the cognitive development and learning of children" (p. 44). The current study examined whether students using GBL demonstrated the ability to retain and apply educational knowledge, specifically in mathematics.

The Rationale for Selecting the Topic

The purpose of this study was to investigate the impact of playing a virtual mathematics game, Prodigy, on students' mathematics achievement scores and student engagement with consideration of program type (GBL enhanced mathematics program, traditional mathematics program) and gender (male, female).

There are mixed results regarding whether gamification or GBL increases student achievement. Gamification has the potential to improve learning (Dicheva et al., 2015) and games are effective in meeting particular learning needs (Kapp, Blair, & Mesch, 2014) if they are well-designed and implemented correctly. Four out of the five studies analyzed by this researcher showed that student performance increased when game-based learning was used in the classroom (Ke & Grabowski, 2007; Kebritchi, Hirumi, & Bai, 2010; Ku, Chen, Wu, Lao & Chan, 2014; Miller & Robertson, 2011). Kim and Chang (2010) showed that there was an improvement in performance for the students who played mathematics games on occasion, but a negative outcome for students who played daily. The current study continues Kim and Chang's (2011) work by examining the virtual game, Prodigy. As Perrotta et al. (2013) stated, "There is still a lot we don't know about the impact of video games on learning" (p. ii).

Using an education taxonomy, such as Bloom's Taxonomy, can help to determine when to use games (Kapp et al., 2014). The cognitive domain focuses on understanding,

comprehending, and synthesizing knowledge, while the affective domain represents emotional concepts (Kapp et al., 2014). In turn, these two domains were used to access students' engagement in the current research study, where Bloom's thinking processes related to the cognitive engagement subscale and the affective domain related to the emotional and social engagement subscales.

To address cognitive attributes, "games provide complex systems of rules for players to explore through active experimentation and discovery" (Lee & Hammer, 2011, p. 149). As such, students stay engaged because games reward them for leveling up, moving onward in the game, or gaining new game abilities. Students' own abilities and achievement increase when these rewards are linked directly to gaining content knowledge.

To address effective attributes, when a student's level of interest is sufficiently high and they are therefore engaged, this student will persist in the face of difficulty (Schlechty, 2011). Indeed, games have been shown to help positively change behavior (Kapp et al., 2014). Also, "people believe that educational games are effective because they motivate children to actively engage in learning as they play" (Hwa, 2018, p. 264).

Prodigy is the virtual mathematics game chosen for this study. Two previous studies showed that Prodigy can improve student achievement (Morrison, Risman, Reilly & Eisinger, 2020; Mahimker, 2014) and engagement (Mahimker, 2014). The game aligns with the Common Core Mathematics Standards. More importantly, the program has an initial diagnostic assessment, known as the placement test, which places students in their appropriate grade level based on their knowledge. One of the key factors needed to make GBL a successful strategy is the careful design of its implementation and use. According to Lee and Hammer (2011), a "critical game design technique is to deliver concrete challenges that are perfectly tailored to the

player's skill level, increasing the difficulty as the player's skill expands" (p. 149). Prodigy's diagnostic assessment creates a pathway for students that aligns with the critical game design of meeting the particular learning needs of each individual (Kapp et al., 2014). In addition, Prodigy engages students in friendly competitions with others called battles and adventures. In these competitions, students can become more engaged in learning and practicing mathematics concepts. Morrison et al. (2020) found that "student engagement was consistently described as a key strength of the program and was consistently married with students' preoccupation and independent learning" (p. 25) and that "school-based adults described the program as keeping students focused and engaged in mathematics content that is both challenging and relevant" (p. 25). For these reasons, Prodigy was selected as it best fit the study's purpose.

Statement of the Problem

The National Assessment of Educational Progress (NAEP) assesses achievement according to what students know and can do. Unfortunately, mathematics scores for American students have been generally flat during the past three administrations of the test "with 66 percent of eighth-grade students scoring below the proficient level in the 2019 administration of the test" (NCTM, 2020, p. 18). *Catalyzing Change in Middle School Mathematics* (NCTM, 2020) identified four challenges:

> to explicitly broaden the stated purposes for teaching and learning mathematics in middle school … to eliminate structures that impede development of student mathematical identity, agency, and mathematical growth … to support, enhance, and adopt instructional practices that are equitable and based on providing high-quality learning opportunities to motivate and engage students in learning mathematics … [and] to

> support students as they continue their study of mathematics and navigate their lives. (NCTM, 2020, p. 19)

The current research addressed the instructional practices that the NCTM (2020) identified as key to supporting students by providing equitable learning opportunities to help increase students' engagement in mathematics.

When students are not engaged in their mathematics classes, they cannot access personalized instructions and practices (Dicheva et al., 2015; Lee & Hammer, 2011). Research supports the use of games as a method to help motivate students (Klein & Freitag, 1991; Lee & Hammer, 2011) and increase engagement (Klien & Freitage, 1995; Liu & Chu, 2010; Sun-Lin & Chiou, 2019). However, such results do not indicate that GBL increases student achievement (Kim & Chang, 2010; Klein & Freitag, 1991).

Understanding whether a virtual game can increase student achievement and engagement is significant based on the current trends in mathematics education. National data shows society needs students' mathematics scores to start trending up as the overall progress made in the last three decades has not been sufficient (NCTM, 2020). Along with the influence of gender, it is important to consider and understand the impact of these variables. GBL is a classroom strategy that required further investigation. This is especially true in an era of ever-improving and ever-present technology, and where teachers are more frequently implementing virtual games in and out of the classroom.

Potential Benefits of the Research and Significance of the Study

The results of this study may assist researchers in determining whether using the virtual mathematics game Prodigy can provide students with a mechanism to engagingly receive the interventions, extensions, and practices they need in their mathematics classes.

Definition of Key Terms

This research study uses the following terms:

1. **Cluster** applies to a group of students who are taught by the same group of core teachers across subject areas (Mathematics, English, Social Studies, and Science).
2. **Gamification** is a strategy where the use of "game mechanics, dynamics, and frameworks" are applied to education (Lee & Hammer, 2011).
3. **Game-Based Learning** is the use of games to support teaching and learning (Anastasiadis et al., 2018; Perrotta et al., 2013).
4. **Student Growth Percentiles** are the norm-referenced quantifications of individual student growth derived using quantile regression techniques (Renaissance Learning Inc., 2018).

Summary of Chapter One

The purpose of the current study was to investigate the impact that playing a virtual mathematics game, Prodigy, has on students' mathematics achievement scores and student engagement regarding gender (male or female) and Program type (a GBL-enhanced mathematics program compared with a traditional mathematics program). This program allows students to engagingly practice their mathematics skills. Hopefully, this research will help teachers implement support experiences for students through GBL on the Prodigy game platform.

CHAPTER 2: RELATED LITERATURE

Many researchers have investigated the use of games in mathematics (Ke & Grabowski, 2007; Kerbritchi et al., 2010; Kim & Chang, 2010; Ku et al., 2014; Miller & Robertson, 2011; Sun-Lin & Chiou, 2019). Typically, these studies are based on gamification (Dicheva et al., 2015; Lee & Hammer, 2011) or game-based learning (Ke & Grabowski, 2007; Kerbritchi et al., 2010; Kim & Chang, 2010; Ku et al., 2014; Liu & Chu, 2010; Miller & Robertson, 2011; Perrotta et al., 2013; Saez-Lopez, Miller, Vázquez-Cano, & Domínguez-Garrido, 2015; Sun-Lin & Chiou, 2019).

Gamification is not a prescribed curriculum; rather, it is the use of gaming elements to enhance activities within the curriculum. According to Burke (2014), the most valuable parts of gamification are the student's engagement level and motivation to participate.

Burke is in favor of adding gamification to the curriculum and Kapp et al. (2020) warn that there need to be specific design elements into the game or gamification to elicit the desired behavior change. At times, people confuse gamification with game-based learning, as often the terms "gamification" and "game-based learning" are used interchangeably (p. 149). Therefore, a distinction must be made between the two. Perrotta et al. (2013) stated that game-based learning is the use of games to support teaching and learning; in contrast, gamification is a newer concept that uses game elements in a lesson or class activity. Anastasiadis et al. (2018) define GBL as the pedagogical approach of utilizing physical or virtual games in education. Plass et al. (2019) state that, "in contrast to gamification, *game-based learning* means a learning task is redesigned to make it more interesting, meaningful, and ultimately, more effective for learning than either a nongame or gamified task" (p. 4). All the noted definitions refer to using game components in the context of education. It is important to discuss both concepts as they are so intertwined;

however, Erickson et al. (2020) state that, "...simply bringing a game into the classroom is not gamification" (p. 149) and using one or more elements of gaming such as awarding points for responses is not GBL. Moreover, Liu and Chu (2010) stated that game-based learning is designed to combine learning and gameplay. Gameplay unaccompanied is the "treatment of topics and ideas as rules, actions decisions, and consequences, rather than as content to be communicated or assimilated" (Perrotta et al., 2013, p. ii). GBL will improve the ability of the players to retain education subjects and apply them beyond the game itself. Wohlwend and Peppler (2015) suggest that play benefits learning by deepening understanding.

Theory

Engagement

Engagement is a theory related to gamification and game-based learning. As the "power of games for education is often described in terms of their potential to enhance learning by engaging users" (Schwartz & Plass, 2019, p. 53). Reschly and Christenson (2012) define engagement as, "an observable, action-oriented subtype (behavioral) and two internal ones (cognitive and affective engagement) but then [it] is differentiated from motivation as engagement being action (observable behavior) motivation as intent (internal)" (p. 14). Engagement and motivation are related constructs, and they are often used together or interchangeably (Harrison et al., 2017). Notably, "motivation is necessary but not sufficient for engagement" (Reschly & Christenson, 2012, p.14). Through their extensive review of the literature, Reschly and Christenson (2012) defined three main types of engagement: behavioral, emotional, and cognitive. Behavioral engagement is defined by participation. Emotional engagement is comprised of affect, positive and negative, and interactions. Finally, cognitive engagement is rooted in personal investment. Reschly and Christenson (2012) state that "student

engagement is the glue, or mediator, that links important context – home, school, peers, and community – to students and in turn to outcomes of interest" (p. 3).

Engagement is a necessary element for improving the outcomes of students (Reschly & Christenson, 2012). In academic tasks, student engagement is the most proximal factor to achievement outcomes; greater engagement leads to greater success (Boykin & Noguera, 2011).

One of the top researchers of engagement, Schlechty (2002), suggests eight qualities of engagement: product focus, clear product standards, protection from adverse consequences for initial failures, affirmation, affiliation, choice, novelty and variety, and authenticity. Game-based learning exemplifies all these qualities by providing students with a task that needs to be completed. A game focuses on a product, and it has clear product standards because it has created activities based on specific content. It is acceptable to fail in a game because one can try again, often in perpetuity. Affiliation is important, and one sees it often when one plays games in the presence and cooperation of others. Schlechty (2011) states, "One of the major design features of many computer games is the affirmation built into them. Players know where they stand and that their efforts count, and this knowledge keeps them trying even when they initially experience failure" (p. 55). In addition, people can choose whether to play the game, the amount of time they will play, and the strategy they will use. They can also choose to alter their selected strategy based on feedback they receive while playing the game. Many games offer various levels or challenges. Most virtual games simulate an authentic experience by involving tasks and offering rewards for completion.

Authenticity indicates linking learning tasks to things that are of interest to students. This is especially important when students are disinterested in learning what teachers and administrators determine is necessary to learn. Plass et al. (2019) state, "games have features

that can enhance learner engagement, make task meaningful and relevance, and adaptively respond to learners' specific needs and conditions" (p. 15). When a student's level of interest is sufficiently high, they will persist – even through difficulty (Schlechty, 2011). Being able to persist connects with the Common Core State Standards and the Standards for Mathematical Practice 1, which expect students to make sense of problems and persevere in solving them (National Governors Association Center for Best Practices, 2010).

Regarding the eight qualities, one should remember that Schlechty (2002) stated, "It is not assumed that all these attributes need to be present in each lesson or even in every unit. Rather it is a list of possibilities-things a teacher might want to consider when attempting to design work that is engaging to students …" (p. 43). In his later work, Schlechty (2011) added two new elements to the original eight qualities to better engage students: content and substance and the organization of the knowledge. Content and substance are embedded in games. Indeed, games provide extrinsic rewards, and they make the subject more interesting. For content and substance, Schlechty (2011) states, "designing work that students want to do and that results in their learning what they need to learn is at the heart of great teaching" (p. 52). The organization of knowledge focuses on the learning style of the students, stating that this should determine how the instructions are delivered.

According to Schlechty (2002, 2011), engagement comes down to the tasks assigned to students, which need to be the central concern for the teacher. As stated by Schlechty (2011), "the primary goal of teachers is that of designing engaging work for students and guiding them to the sources of instruction they need to do this work successfully" (p. 58). Authentic engagement requires that the task, activity, or work must be associated with a result or outcome that has a clear meaning and relatively immediate value to the student being engaged (Schlechty, 2011).

Engagement paves the way for learning so "assessing engagement is a means of ensuring the possibility of preventing deficiencies in learning" (Schlechty, 2002, p. 41).

Kapp et al. (2014) suggest that one can classify the use of games into three learning domains using Bloom's cognitive, affective, and psychomotor taxonomy. For the cognitive aspect, "games provide complex systems of rules for players to explore through active experimentation and discovery" (Lee & Hammer, 2011, p. 149). Games in Bloom's cognitive taxonomy have students create, evaluate, analyze, apply, understand, and remember (Kapp et al., 2014). As such, students will stay engaged because games reward students by allowing them to level up, move onward in the game, or gain new game abilities as the students' ability increases. Students exhibit many emotions when playing games, but especially disappointment in failure. Students are constantly experimenting before succeeding or leveling up. This leads to many failures, allowing them to build resilience and teach them to keep trying. Finally, role-playing games (RPGs) include a social aspect that allows students to "try on new identities and roles" (Lee & Hammer, 2011, p. 150) by playing different characters in a game. One of the key factors to make gamification a successful intervention tool is the careful design of the implementation and use of a game. According to Lee and Hammer (2011), a "critical game design technique is to deliver concrete challenges that are perfectly tailored to the player's skill level, increasing the difficulty as the player's skill expands" (p. 149).

To help engage students, one should meet them at their level; games can achieve that with relative ease (Lee & Hammer, 2011). An approach to meeting each student's needs is to provide tasks within each student's zone of proximal development (Vygotsky, 1978). A zone of proximal development is a term used to describe the "distance between the actual development 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" (Vygotsky, 1978, p. 86). Vygotsky saw play as a way to create the zone of proximal development (ZPD), as play serves as a scaffolding function to support children, acting as a critical tool for learning new things (Homer et al., 2019).

To teach according to a student's zone of proximal development, the teacher must first determine what that zone is. To do this, the teacher must gather diagnostic information to assess the student's mathematics development level. When developing or choosing a game to use, there must be "… importance [given to] designing learner-adaptive engagement and a balanced integration of content and play in educational games" (Ke & Abras, 2013, p. 241). The adaptive nature of many educational games appears in the diagnostic or placement tests, which place the students at the correct level (i.e., in their zone of proximal development). As Hung (2001) states, "recognizing individual current levels of cognition and tailoring a *personalized* 'curriculum' for the learner within the stipulated ZPD would be the ideal learning situation" (p. 34).

Boykin and Noguera (2011) stated that engagement is three-dimensional and acts as a precursor to achievement and even growth in students' achievement levels over the years. They also found that engagement has a greater effect on math gains, especially for the lowest-achieving groups compared with instructional time (Boykin & Noguera, 2011). Engagement is a multidimensional construct with key building blocks, and when students are engaged it increases student achievement (Boykin & Noguera, 2011; Reschly & Christenson, 2012).

Activity Theory

Activity theory started within the psychology of play, learning, cognition, and child development; it evolved to include the development of work activities, computer technologies, and therapy (Engestrom, Miettinen, Raija-Leena, & Punamaki, 1999). Activity theory is based

on the classical German philosophies of Kant and Hegel as well as Marx and Engels and the Soviet-Russian cultural-historical psychology of Vygotsky, Leont'ev, and Luria (Engestrom et al., 1999; Jonassen & Rohrer-Murphy, 1999). "Activity theory posits that conscious learning emerges from activity (performance), not as a precursor to it. So, activity theory provides us with an alternative way of viewing human thinking and activity" (Jonassen & Rohrer-Murphy, 1999, p. 62). The basis of the theory is to consider everything happening in the activity, the purposeful interactions between the subject, and the object necessary to complete a goal.

To fully understand activity theory, one must first consider the activity system which is sometimes called the *international activity-theoretical collaboration* (Engestrom et al., 1999). The focus of the activity system is the production of the goal or outcome, depicted as a triangle. The system includes the subject (an individual or group), the object of the activity (the physical or mental product that is sought), and the community as the main components of the system. The items that help those components and affect the outcome include the tools used for the activity, the actions and operations, such as the rules, and the division of labor (Carvalho, Bellotti, Berta, De Gloria, Sedona, Hauge, Hu, & Rauterberg, 2015; Jonassen & Rohrer-Murphy, 1999). There is also a hierarchical structure of activity theory where an activity happens simultaneously at three levels: activity, action, and operation. This is the basis of activity theory. From there, it has been applied to multiple different disciplines and research.

Carvalho and colleagues (2015) developed the Activity Theory-based Model of Serious Games (ATMSG). The ATMSG uses the concepts of activity theory and includes a serious games-component taxonomy. In addition, one can use this model as a tool for the conceptual design of serious games. Serious games, games with a purpose, have been shown to have instructional potential (Carvalho et al., 2015). To analyze games and to better understand how

learning takes place, ATMSG is concerned with three components: actions, tools, and goals. In turn, it considers how each of these three is related to gaming, learning, and instruction (intrinsic and extrinsic).

ATMSG can support serious game design during two phases. Phase 1 focuses on the high-level characteristics of the serious games – actions, tools, and goals. Phase 2, "should be applied in conjunction with prototyping techniques, preferably, but not limited to low-fidelity ones (sketches, storyboard, game diagrams, etc.) (Carvalho et al., 2015, p. 175). These phases support the game design process and both are needed to create a serious game. In support of these phases, Plass et al. (2019) stated that "the design process therefore has to carefully calibrate how much each of these two design objectives - learning outcomes versus playfulness - should influence specific design decisions" (p. 4).

This model has created a way to have a detailed analysis of educational games. "The ATMSG model provides a comprehensive way to investigate, in detail, how a serious game is structured, using activity theory as the theoretical background" (Carvalho et al., 2015, p. 180). This is a conceptual framework that combines analyzing the game's elements and its pedagogical goals.

One can examine the game Prodigy using the ATMSG model for serious games provided by Carvalho et al. (2015), which breaks serious games down into three parts – the gaming activity, learning activity, and instructional activity. The motive of the gaming activity is to have fun whereas the learning activity is to have students work on mathematical concepts. The instructional activity is for the instructor and helps evaluate the extent to which the instruction components of the game support the stated learning outcome. Prodigy shows this with reports of students' progress, comprehension, and usage. The ATMSG model also makes a distinction

between intrinsic and extrinsic instruction. In Prodigy, the activities are all intrinsic as the instructional activity takes place solely inside the game, wherein the game is the one supporting the learning.

Related Research

Game-Based Learning Studies in Mathematics

Most game-based learning aims to motivate participants to increase their levels of attainment. For this current research, studies related to game-based learning were found by searching the following key terms: game-based learning, education, mathematics, achievement, and engagement. For this research, game-based learning, education, and mathematics were put in as "and" to ensure they were included, while achievement and engagement were put in as "or." That search provided 714,451 results. Subsequently, the researcher narrowed down these results to 1,650, by including achievement and engagement as "and" instead of "or." In turn, the researcher further narrowed down the search to articles published in English during the last 20 years, which resulted in 1,307 results. At this point, she started reading the abstracts.

The researcher kept the studies that pertained to mathematics, those that had a similar population to the current study (middle school), and studies that explained performance or achievement as well as motivation or engagement. The terms motivation or engagement were both used because they are related constructs, and some scholars use the words interchangeably (Reschly & Christenson, 2012). In addition, motivation is seen as a subset of cognitive engagement. Ainley (2013) stated, "Describing students' engagement with learning mathematics implicates motivation for the activity, just as describing students' disengagement or disaffection with learning mathematics implicates lack of motivation for the activity" (p. 286).

The studies related to game-based learning and cognitive or affective constructs, then were divided into two categories. These categories were studies related to game-based learning in mathematics classrooms, see Table 1, and studies related to game-based learning across subject areas, see Table 2. Both sets of studies pertained to middle school students.

Table 1

Studies Related to Game-Based Learning in Mathematics Classrooms

Reference	Type	Sample	Data Collection	Results
Ke and Grabowski (2007)	Experimental design using a pretest-posttest design to measure the effects of game playing on math performance and attitudes. Students were assigned to one of three groups: cooperative gameplaying, competitive game playing, or no game playing.	There were 125 5th grade students from 6 public schools. All students volunteered to be in the study. Members of the intervention group played a series of web-based computer games that focused on academic standards for mathematics.	The study included data from a 30-item math skills assessment, for mathematical performance, and Tapia's Attitudes Towards Math Inventory, for attitudes toward mathematics.	There was a significant effect on the gameplaying students' math performance and attitudes. The cooperative gameplaying group had a significantly higher positive math attitude compared with the other groups.
Kebritchi, Hirumi, and Bai (2010)	An experimental design was used to research achievement and motivation using an instructional math game called DimensionM. Teachers were randomly assigned to have their students in the control or treatment groups.	There was a total of 193 algebra and pre-algebra students and 10 teachers, with the teachers evenly and randomly assigned to the control group (n = 76) or a treatment group (n = 117)	To measure mathematics achievement, district-wide benchmark exams were used. Motivation was assessed using a 20-item survey.	There was a significant difference in achievement. The treatment group had a greater gain, with a mean increase of 3.74. There was an insignificant difference in motivation between the two groups.
Kim and Chang (2010)	Empirical research using the 2005	The participants included 3,732,411 U.S.	Data were collected from the NAEP	A multiple regression model was used for analysis. The results

Reference	Type	Sample	Data Collection	Results
	National Assessment of Educational Progress (NAEP) and a multi-stage probability sample design to collect the data by performing a regression analysis on whether playing computer games had an effect on math achievement.	4th grade students and 32,898 teachers	performance indicator and related surveys; information was collected on teacher instruction practices, educational context, and student behaviors provided by the National Center for Education Statistics. Three separate hierarchical regression analyses were conducted: the ELL Model, the Gender Model, and the Interaction Model.	showed that computer games had greater positive effects on males than on females. ELL students who played math games daily had a higher math performance compared to non-ELL students who did not play games. However, non-ELL students who played the same games daily also performed lower in math.
Ku, Chen, Wu, Lao and Chan (2014)	An experimental design that examined the effects of game-based learning with a focus on student confidence towards mathematics and performance using a game-based environment compared to a	A sample of convenience comprised of two 4th grade classes (n = 51) were randomly selected to either be in the control group or the experimental group.	Personal confidence in mathematics and mathematics performance using a computational skills test were used.	The students who played mini-games, which focused on the math content, showed significantly greater gains in confidence and achievement than members of the comparison group.

Reference	Type	Sample	Data Collection	Results
	paper-based one.			
Miller and Robertson (2011)	Experimental design using a pre-post test measuring mathematics performance, in mental computation (sums correct and time), and self-perception of mathematics and school. The experimental group played a commercial off-the-shelf computer game for 20 minutes a day.	A randomized controlled trial with 634 primary students, ages 10 to 11, separated into control and experimental groups.	To measure performance, speed, and accuracy, a 100-item Number Challenge assessment tool was used. A self-perception instrument focusing on 2 sub-scales, mathematics self-concepts, and general academic self-concept.	All students, regardless of group, had an increase in accuracy of computation and speed of processing. The experimental group improved at a faster rate for accuracy and the gain for them was over 50% higher than it was for the control group. There was no significant difference in math self-perception between the student groups.
Sun-Lin & Chiou (2019)	Quasi-experimental design using a gamified comparison to measure the effect of gameplay on student achievement in algebra word problems and attitudes.	72 sixth-grade students (39 females and 33 males) were invited to take part in a 4-week experiment where they were randomly assigned to one of 3 groups: gamified comparison, comparison, or control.	Student achievement of algebra word problems was based on test scores for similar and transfer problems; Algebra learning attitude scale (4 dimensions: enjoyment, motivation, confidence, and perceived value).	There was a significant effect on solving similar problems and transfer problems. The gamified comparison group performed better than the other 2 groups. There was a significant effect on learning attitudes, with the gamified comparison group offering more positive responses; however, there was no significant difference between the groups in terms of confidence.

In these six studies, all researchers examined the effect that game-based learning had on a mathematics class. Five of the six studies showed an increase in student performance (Ke & Grabowski, 2007; Kebritchi et al., 2010; Ku et al., 2014; Miller & Robertson, 2011; Sun-Lin & Chiou, 2019). Kim and Chang (2010) reported an improvement in performance for the students who played mathematics games on occasion. They also showed a negative outcome for non-ELL male students who played the games daily, but there was no negative effect for ELL male students that played the game every day. They attributed the negative impact to, "daily games for English-speaking male students can be [a] distracting factor for their school engagement, but daily games for ELL male students can be an educational stimulator" (Kim & Chang, 2010, p. 231). Two of the studies showed positive outcomes in students' attitudes and confidence (Ku et al., 2014; Ke & Grabowski, 2007). One study showed an increase in students' motivation, enjoyment, and perceived value (Sun-Lin & Chiou, 2019), however Kebritchi et al. (2010) found no change in motivation.

Ku et al. (2014) created an experimental design using convenience sampling that examined the effects of game-based learning on fourth-grade elementary students, with a focus on the students' performance and confidence regarding mathematics. The experiment spanned five weeks, and each session lasted 20 minutes two times a week. The students in the control group (CG) learned their academic content in a paper-based condition, whereas the students in the experimental group (EG) learned the content in a digital, game-based environment. All students worked on mental computation. The experimental group played mini-games such as Battleship and Math Kicker, a soccer game. To do anything in the game, the students needed to answer a question correctly. The students who played the mini-games showed more gains in confidence than did those in the control group. In achievement, the gain in the experimental

group was significantly greater than that of the comparison group (U = 193.000, p = .013), as shown in the Mann-Whitney test that was performed. The experimental group members also scored 14.18 more points on the mean posttest than did the students in the control group.

In another game-based study, Miller and Robertson (2011) conducted a randomized controlled trial with 634 primary school children, ages 10 and 11. The students in the experimental group played Dr. Kawashima's Brain Training on a Nintendo DS Lite for 20 minutes a day during the nine-week trial, where the control group continued with their normal routine. The study aimed to investigate changes in the children's mental computation performance and self-perception related to school and mathematics. The mental computation performance of students was assessed using a 100-item "Number Challenge," and self-perception was measured as a shortened form of Marsh's Self-Description Questionnaire that measured only two of the eight subscales – mathematics and general school.

All students, regardless of group, showed an increase in the accuracy of computation and speed of processing. The experimental group's improvement was statically significant and at a faster rate for accuracy than the control group; the gain for the experimental group was over 50% higher. The experimental group also had a better mean post-processing speed time (13:29) compared to the control group (15:34). There was no significant difference between the two groups in terms of the students' mathematics self-concept. There was a significant interaction between time and condition, $F(1, 632) = 2.74$, $p = 0.049$, but no significant effect on children's attitudes related to time. No significant gender differences were noted.

While Miller and Robertson's (2011) study showed an increase in computation accuracy and speed as a result of playing math games, Kim and Chang's study (2010) did not. Many researchers have suggested that games increase motivation and that they create a positive

environment while ensuring that the content is not boring (Klein & Freitag, 1991). Playing games also helps to foster a collaborative and competitive environment (Dichev & Dicheva, 2017). In addition, a "digital game-based learning approach and serious games in general can be utilized as an educational tool which can boost students' wellbeing and self-esteem, help them improve their soft skills, develop their critical thinking, decision-making and problem-solving skills, as well as maintain a healthy mental and psychological balance" (Anastasiadis et al., 2018, p. 142).

Kim and Chang (2010) investigated the effects of computer games on the math achievement of 4th-grade students. They examined the differences between gender and linguistic groups by using data from the National Assessment of Education Progress. The participants were comprised of 3,732,411 4th-grade students. Their methods included, "a multi-stage probability sample design: in the first stage, units are the US counties, the second-stage units are schools within the counties, and the last-stage units are students" (p. 226). In addition, Kim and Chang (2010) also surveyed teachers to collect data on their instructional practices and the frequency of computer games they used in math class. Teachers (n = 32,898) responded to the survey, scored by more than one person. The variables they used were Math Item Response Theory Score, Free and Reduced Lunch Eligibility, English Language Learner (ELL), and designation of frequency of gameplay (Game Sometimes and Game Daily). The researchers then segregated the participants into groups based on gender and ELL status. Kim and Chang (2010) used a covariate of the socioeconomic status of the students to help to control the contextual variable for student academic performance. Their results showed that the effects of computer games had a greater positive effect on males than they did on females in all three of the categories, with an average mean difference of 2.3. Regression analysis showed that the ELL

students who played math games daily had a higher math performance than did those who did not play daily. Interestingly, non-ELL students who played the same games daily performed lower in math than did those who did not play daily. In each comparison group, males benefited from playing the games, while females did not. Kim and Chang (2010) concluded that when males played the educational games too frequently, they showed lower math performance than those who played them only occasionally during math class. The researchers suggested that future studies should examine the optimal frequency of play to ensure that computer games increase student achievement. In addition, they suggested that researchers should investigate the types of games students are playing and whether they align with the content of the classroom.

Kebritchi et al. (2010) performed an experimental study that included 193 Algebra and Pre-Algebra students with a control group comprised of 76 students and a treatment group of 117 participants. Ten teachers were randomly assigned to either the treatment or the control group. The researchers studied achievement and motivation. Achievement was measured using a district-wide benchmark exam where the reliability and validity were determined by the school district. Reliability for the Grade 9 pretest was 0.73, and for the posttest, it was 0.84; for Grade 10 students, the pretest was 0.82 and the posttest was 0.86. Data on motivation were acquired through a 20-item survey "based on Keller's ARCS Model. This model measures the motivation along four major attributes of Attention, Relevance, Confidence and Satisfaction (ARCS)" (Kebritchi et al., 2010, p. 432). This test met the requirement of reliability by being higher than 0.7. In addition, interviews were conducted to obtain information on the perceptions of the experiences of the participants.

The treatment group played a set of instructional math games called DimensionsM^{TM}. These games, "have a learner centered nature in which the players learn the mathematical

concepts by completing virtual missions" (Kebritchi et al., 2010, p. 431). The students in the treatment group played the games in DimensionsM^{TM} for 30 minutes each week for 18 weeks during regular class time. In turn, students in the control group spent that same time receiving regular mathematics instruction.

For the analysis, a Multivariate analysis of covariance (MANCOVA) was used. The results showed a significant difference in terms of achievement between the groups on the posttest, $F(1, 188) = 6.93$, $p < .01$. A paired-samples *t*-test, indicated that each group achieved significant gains; however, the students who played the game showed greater gains from the pretest to posttest, with a mean increase of 8.07. In contrast, students in the control group had a mean increase of 3.74. The analysis showed an insignificant difference in motivation across the treatment and control groups.

Ke and Grabowski (2007) used a pretest-posttest experimental design to study the effects of both cooperative and interpersonal game-playing on students' math performances and math attitudes. Their population was comprised of 125 fifth-grade students. The students in the experimental group played a web-based game developed by the Center for Advanced Technologies called ASTRA EAGLE. The program was designed to reinforce academic standards for mathematics; it was comprised of four games with multiple levels for the students to work through.

A 30-item math skills assessment named Game Skill Arithmetic Test (GSAT) was developed with a reliability of 0.80. In addition, the Tapia's Attitudes Towards Math Inventory (ATMI) was modified for fifth grade, and it consisted of 40 items with a 5-point Likert scale. A MANCOVA was done on the post-GSAT and ATMI scores. It showed a positive significant effect of game playing on the participants' math performances and math attitudes, $F = 4.395$, $p =$

0.002, (Ke & Gradbowski,2007, p. 254). The study showed no significant interactions between the treatment and gender, $F = 0.518$, $p = 0.723$, on either math performance or attitudes, and there was no indication of the main effects on gender (Ke & Grabowski, 2007, p. 255). Based on their results, Ke and Grabowski (2007) recommended, "that educational practitioners use gaming within meaningful learning environments or tasks to promote learning" (p. 257).

Sun-Lin and Chiou (2019) conducted a quasi-experimental study to measure the effect of gameplay on students' learning achievement. Specifically, they studied algebra word problems regarding students' ability to solve similar and transfer problems. They also assessed student learning attitudes. The students were randomly assigned to one of the three groups, each of which was comprised of 24 sixth graders. The measures used for analysis were the students' test scores on similar and transfer problems as well as the algebra learning attitude scale.

The three groups used in the study were gamified comparison, comparison, and control group. During the study, both the gamified comparison group and the comparison group learned to solve algebra word problems by completing four comparison tasks. In addition, the gamified comparison group played a game that involved space and aircrafts; their aim was to prepare their aircraft to be able to protect the Earth from a space attack. By solving the questions correctly, the students would level up to create better aircrafts. In contrast, the control group learned to solve the word problems, but they were not given comparison problems.

The gamified comparison group produced significantly higher scores than the other two groups on the comparison tasks, transfer problems, and solving similar problems. The researchers concluded that the method used by the gamified comparison group could effectively help students learn to solve algebra word problems. In addition, the gamified comparison group made significantly more positive responses than did members of the other groups in terms of

enjoyment, motivation, and perceived value. Additionally, Sun-Lin and Chiou (2019) concluded there was no significant difference between the groups in terms of the students' confidence. The researchers also found that combining game rules with meta-cognitive strategies was as effective tool for facilitating students' achievement in solving algebra word problems; moreover, it increased their enjoyment, motivation, and perception of learning.

Game-Based Learning Across Subject Areas

The studies in Table 2, relate to game-based learning across subject areas in a middle school setting.

Table 2

Studies Related to Game-Based Learning Across Subject Areas

Reference	Type	Sample	Data Collection	Results
Liu and Chu (2010)	Quasi-experimental design where the students were divided into control and experimental groups. The researchers investigated how ubiquitous games, games that blend everyday experiences, influence English learning achievement and motivation.	The participants were 64 seventh-grade students, ages 13 or 14, and 3 teachers. The students were assigned to either the control group, using non-game learning, or the experimental group, using ubiquitous game-based learning.	The mixed methods study included data from students' motivation survey and interviews about students' opinions and they explanations of experimental results.	Student achievement scores for those in the experimental group significantly exceeded those of the control group for all measures. Through the qualitative data, the students stated the increase in achievement was due to using interesting materials in real-life situations. There was a positive relationship between learning outcomes and motivation in the experimental group.
Saez-Lopez, Miller, Vasquez-Cano, and	A quasi-experimental study assessing the use of MinecraftEdu in	There were 131 students in the treatment group and 50 students acting in the	Data were collected for the study by using an achievement	There was no significant difference in scores between the control group and the experimental group on

Reference	Type	Sample	Data Collection	Results
Dominquez-Garrido (2015)	a classroom in regard to pedagogical benefit, achievement, attitudes, and interactions.	control group. The participants were from the US and Spain, ages 11 – 14.	test (P1mp), an attitudes questionnaire (APMA) and messages on the Edmodo platform to determine interactions.	the P1mp. The attitudes of teachers, students, and parents were all positive. The APMA indicated that using MinecraftEdu was fun (98.5%), enabled the discovery of new things (96.6%), encouraged learning about historical content, (97.1%), and encouraged learning (83.4%).

Saez-Lopez, Miller, Vasquez-Cano, and Dominquez (2015) and Liu and Chu (2010) examined middle-school populations playing learning games. Both studies showed a significant positive increase in students' attitudes and motivation. Saez-Lopez and colleagues (2015) found no change in achievement whereas Liu and Chu (2010) determined an increase in academic performance.

Saez-Lopez and colleagues (2015) performed a quasi-experimental study that included students aged 11 to 14 years old from several history classes in the US and Spain. There were 131 students in the experimental group and 50 students in the control group. The students used MinecraftEdu, an open virtual world in which there is no plot or story; the game design provides the students with the freedom to explore the virtual world. As an avatar, a student can explore, create, discover, and experiment in this immersive environment in collaboration with classmates. In addition, they can receive tutoring from the teacher, who also has an avatar in this world. The teacher designs the unit and creates a map for the students to provide them with numerous interactions, explorations, and discoveries.

Information was collected using three sources: an academic unit test (P1mp), an attitudinal questionnaire (APMA), and messages on the Edmodo platform. The questionnaire was designed to gather data about the attitudes of parents, teachers, and students regarding working with the MinecraftEdu program. The posttest results from the P1mp showed an increased median score of 0.2 for the experimental group, but the difference was not significant. In contrast, there was no significant increase for the control group. The results from the APMA revealed that the subjects in the sample group scored 96% or above in fun, discovery, encouraging learning, and rich interactions using virtual environments. In total, 71.7% of the participants thought that applying game-based learning in class was appropriate for the learning process.

Liu and Chu (2010) also implemented a quasi-experimental study to investigate how ubiquitous games, defined as games that blend with everyday experiences, influence English learning achievement and motivation. Their sample was comprised of 64 seventh grade students who were placed in either an experimental or control group. Each group had four teams with eight members each. The experimental group used a ubiquitous game-based learning environment called HELLO. The students played different computer games in which they moved their characters around activities inside the HELLO learning environment. The control group used printed materials and audio CDs to learn. The results showed that students in the experimental group significantly exceeded those in the control group for all measures. The research results also showed better results for the experimental group in attention, relevance, confidence, and satisfaction. The indicators that showed a statistically significant difference in terms of attention were the attractiveness of the presentation of learning materials, the attractiveness of the active nature of the learning activities, the students' curiosity toward the

active nature of the learning activities, and the students' interest in the activities. The indicator that showed a statistically significant difference in terms of relevance was that the students thought the three activities were helpful to them. The indicators that showed a statistically significant difference in terms of confidence were that the students believe that the progressing method of the learning activities met their expectations and that the students had the confidence to apply what they learned from this course to their everyday lives. The indicators that showed a statistically significant difference in terms of satisfaction were that the students enjoyed the three activities and that they were satisfied with their learning achievement. In addition, "according to the interviews, students in the experimental group thought the ubiquitous games interesting and useful for assisting them to learn" (p. 641).

Examining Gamification and Game-Based Research

The following two studies were chosen because they both used a systematic review of literature; one was used to examine gamification research and the other examined game-based learning research. The results are shown in Table 3.

Table 3

Examining Gamification and Game-Based Learning Research

Reference	Type	Sample	Data Collection	Results
Dicheva, Dichhev, Agre, and Angelova, (2015)	Use a systematic mapping study.	Reviewed 34 case studies	Published empirical research.	Gamification has the potential to improve learning if it is designed and used correctly. The results showed that there were many publications on the use of gamification in education, but the majority described only some game mechanisms and dynamics and simply

Reference	Type	Sample	Data Collection	Results
				reiterated their possible use in an educational context. In contrast, true empirical research on the effectiveness of incorporating game elements in learning environments was still quite sparse.
Perrotta, Feathersto ne, Aston, and Houghton (2013)	Used empirical research and evaluation.	485 items were coded and shortlisted to 31 items.	Published empirical research with two levels. High level: large-scale quantitative studies, meta-analyses, or in-depth qualitative and systematic reviews. Medium level: quantitative or qualitative with smaller sample sizes.	The results showed that the literature is split on the extent to which games can impact overall academic performance. The studies consistently found that video games can positively impact problem-solving skills, broader knowledge of acquisition motivation, and engagement.

A systematic mapping study by Dicheva, et al. (2015) was used to examine the empirical research on the application of gamification in education. The authors collected evidence by reviewing existing seminal and conceptual publications of literature reviews as well as 34 papers reporting empirical studies. The researchers found that the design principles of gamification that were predominantly investigated concerned students knowing their visible status, social engagement, freedom of choice, freedom to fail, rapid feedback, and goals/challenges. In gamification, students are shown that taking risks is okay and not succeeding is not perceived as a failure. The results of the game mechanics show that badges were the most awarded type of

prize that the student received, appearing 27% of the time; in turn, leader board and points followed, appearing 26% and 24%, respectively. The least used mechanics were virtual goods and avatars, each of which appeared less than 2% of the time. The most prominent type of learning was blended learning, which was seen 53% of the time. The most prominent classes that used gamification were computer science and information technology, followed by math, science, and engineering. The reported results were synthesized and showed that 62% of the participants provided positive feedback from gamification and only 3% gave negative feedback. The results showed that there were many publications on the use of gamification in education, but the majority described only some game mechanisms and dynamics and reiterated the possible use of gamification in an educational context. In contrast, true empirical research on the effectiveness of incorporating game elements in learning environments was still quite sparse (Dicheva et al., 2015). The researchers concluded that the "reviewed papers share the opinion that gamification has the potential to improve learning if it is well designed and correctly used" (p. 83).

Perrotta, et al. (2013) conducted empirical research using 31 items that met their criteria after doing a review of the literature. Each item was classified as either high or medium based on the nature of the study or the sample size. The high category included large-scale quantitative studies, meta-analyses, and in-depth qualitative case studies that covered a range of settings and stakeholders; triangulated and systematic reviews were also included in this category. The report focused on definitions, the impact and potential impact of game-based learning, implications on future research, and implications for teachers and schools. The first part of the paper focused on defining game-based learning and comparing it to gamification and gameplay. Game-based learning, "broadly refers to the use of video games to support teaching and learning" (p. 5). In

turn, gamification refers to those who are, “interested in ‘gaming’ as a set of principles” (p. 7). Overall, the analysis of Perrotta et al. (2013) demonstrated a positive relationship between gaming in the classroom, learning outcomes, and motivation and engagement. However, the overall strength of the evidence is not high due to inconsistencies in the research designs or lack of information. The researchers also issued a warning, “Don’t rely on games to improve attainments - there is still a lot we don’t know about the impact of video games on learning” (Perrotta et al., 2013 p. 27).

Conclusion

Ke and Grabowski (2007) commented that the diversity in the results of GBL can be attributed to “researchers having used different configurations of games, such as networking or being competitive, or examined different outcomes such as exam performance or learning attitudes” (p. 250). Seven of the nine studies about GBL showed an increase in performance or achievement using games. Two of those nine studies, Kim & Chang (2010) and Klein and Freitag (1991), reported a decrease in performance using games, while Saez-Lopez et al. (2015) revealed no change in performance. All the studies reported either an increase in motivation or attitudes or no change at all. Miller and Robertson (2011) stated, “the evidence that does exist points to benefits of gameplaying in terms of faster processing of information, enhanced selection of relevant material and high levels of engagement and interest” (p. 852). In short, it depends on how well the game is designed and whether it is used correctly (Dicheva et al., 2015), but games should be incorporated for both educational and motivational reasons (Leonard & Tracy, 1993). A summary of the studies and the results based on achievement or performance and engagement or motivation can be found in Table 4.

Table 4

Summary of Game-Based Learning-Related Research

Reference	Population	Achievement or Performance for the Group that Played Games	Engagement or Motivation for the Group that Played Games
Ke & Grabowski (2007)	5th-grade math students	Increase in performance	Increase in positive math attributes
Kebritchi, Hirumi & Bai (2010)	Pre-algebra and algebra students	Increase in achievement	No difference in motivation
Kim & Chang (2010)	4th-grade math students	Decrease in performance if used every day. Increase in performance if used sometimes	Not studied
Klein & Freitag (1991)	Education undergraduates	Performance decreased	Increase in motivation
Ku, Chen, Wu, Lao & Chan (2014)	4th-grade math students	Increase in performance	Increase in confidence
Liu & Chu (2010)	7th grade English (ELL) students	Increase in performance	Increase in motivation
Miller & Robatson (2011)	Students aged 10 to 11 years old in mathematics classes	Increase in performance	No change in self-concept
Saez-Lopez, Miller, Vázquez-Cano, Domínguez-Garrido (2015)	Middle school history students	Not studied	Increase in attitudes
Sun-Lin & Chiou (2019)	6th-grade math students	Increase in achievement	Increase in motivation, enjoyment, and perceived value

CHAPTER 3: METHODOLOGY

The purpose of this study was to investigate the impact that playing a virtual mathematics game, Prodigy, had on students' mathematics achievement scores and student engagement, with consideration of gender (male, female) and program type (GBL enhanced mathematics program, traditional mathematics program). This chapter provides details of the methodology used to examine this topic, and it is divided into the following sections: (a) setting, sampling procedures, and participants; (b) the research questions and hypotheses; (c) the research design; (d) research design; (e) the instrumentation and data collection tools; (f) the justification of the analyses; (g) student activities; (h) the data collection procedures and timeline; and (j) an ethics statement.

Description of the Setting, Sampling Procedures and Participants

Selection and Description of the Setting

This research took place at two middle schools in an urban district in the northeastern United States. The school district was conveniently located near the work environment of the researcher. The researcher was not an employee of either school. The district contains three middle schools, one of which is a magnet school. The middle schools are in an urban community with a population of approximately 85,000 people (United States Census, 2017). The district enrolled 12,000 students in pre-kindergarten through grade 12 (CSDE, 2018). At the time of the study 6.5% of students in the district identified as Asian, 6.9% Black or African American, 50.6% Hispanic, and 33.0% White (CSDE, 2018). Approximately 26.2% of students in the district came from homes where English was not the primary language; 66.2% were eligible for a free or reduced-price lunch based on their socioeconomic status (CSDE, 2018). This community consisted of 19 schools: 2 early childhood education centers served students in preschool; 12 elementary schools enrolled students from kindergarten through grade 5; 3 middle

schools including students from grades 6 through 8; and 2 high schools, 1 alternative and 1 comprehensive, served students in grades 9 through 12 (CSDE, 2018). In 2017, the estimated median household income for the district was $67,430, the national median household income was $60, 336 (United States Census, 2017).

The two selected middle schools served approximately 2,000 students. In Middle School A, 979 students attended while 977 students attended Middle School B (CSDE, 2018). This study used a sample of convenience of seventh and eighth-grade students as well as teacher participants. Seventh and eighth-grade students were selected because their current math programs did not include a game-based approach. In contrast, the sixth grade used a game program called MobyMax at both schools, therefore, they were omitted from the participant pool. The absence of a program in the seventh and eighth grades allowed the researcher to examine the effects of this game-based platform without influence of another GBL program.

Table 5 presents the breakdown of ethnicity, English language learner (ELL) status, and free or reduced-price lunch status for students attending the two middle schools. Middle School A employed 140 full-time equivalent staff members, including 80 teachers (CSDE, 2018). In addition, 1.0% of staff identified as Asian, 5.0% Black or African American, and 82.0% White (CSDE, 2018). Middle School B employed 117 full-time equivalent staff members, including 78 teachers (CSDE, 2018). 2.9% of staff identified as Asian, 2.9% Black or African American, 2.0% Hispanic or Latino of any race, 2.0% two or more races, and 90.2% White (CSDE, 2018).

Table 5

Racial/Ethnic Status for Student Population for District and Participating Middle Schools

Ethnicity	District	Middle School A	Middle School B
Asian American	6.5	5.4	5.7
Black or African American	6.9	6.9	7.4
Hispanic or Latino of any race	50.6	59.1	49.0
White	33.0	25.8	33.9
Two or More	2.8	0.0	0.0
Total	99.8	97.2	96.0
ELL Status	26.2	27.3	13.3
Free/Reduced Lunch Status	66.2	80.6	68.3

Note. CSDE, 2018

Sampling Procedures for Participants

The study included a sample of convenience comprised of students in grades seven and eight as well as teacher participants. Before starting the study, the researcher obtained consent from the Western Connecticut State University Institutional Review Board (IRB; Appendix A).

Adult Participation. After IRB approval was obtained, the researcher met with the district Superintendent and Assistant Superintendent to describe and propose her study. After they granted permission (Appendix B), the researcher met with the two middle school principals to describe the study and seek permission for the study to take place in their schools (Appendix C). Once permission was obtained from both principals, the researcher met with each school's grade-level PLCs (professional learning communities) to explain the study, answer questions, and invite the teachers to participate in the study either as a part of the comparison group or the

treatment group. Teachers were able to choose if they wanted to participate in either group. Each grade-level professional learning community (PLC) consisted of three core mathematics teachers; therefore, there were four teams, with a total of, 12 teachers across the two schools. Consent forms were then provided to the potential teachers, who were invited to participate in the study and return signed forms to the researcher (Appendix D). At School A, all three seventh-grade math teachers and two of the three eighth-grade math teachers consented to participate. At School B, two of the three seventh-grade math teachers and all three of the eighth-grade math teachers consented to participate. Each participating teacher was provided with an identification code (Teacher 1 through Teacher 10) to ensure confidentiality. The demographic characteristics of the teachers are included in Table 6.

Table 6

Demographic Characteristics of Participating Teachers

Characteristics	Number of Teachers (n = 10)	Percentage of the Sample
School		
School A	5	50
School B	5	50
Grade Level Taught		
Grade 7	5	50
Grade 8	5	50
Teaching Experience		
0 – 10 years	4	40
11 – 20 years	1	10
21 or more years	5	50
Identity		
African American	1	10
White – non-Hispanic	8	80
Two or more races	1	10

At each middle school students were placed into clusters, with each grade level having three clusters. The term cluster refers to a group of students taught by the same core teachers, with one teacher for each subject (Mathematics, English, Social Studies, and Science). All the teachers in this study taught as part of a four-teacher cluster, therefore both the teacher and cluster were assigned an average of 107.4 students. All grade-level teachers taught the same

curriculum in the same sequence with the same standards. The study took place in the seventh- and eighth-grade mathematics classrooms where the cluster system was maintained throughout the study.

Mathematics was chosen for this study because the teachers had an enrichment/intervention block within their schedule during which they did not use any specific program. The teachers and administrators had expressed interest in using a digital platform to meet the individual needs of their students during this enrichment/intervention block. This study used the virtual mathematics game program Prodigy to provide intervention and enrichment for the students in the treatment group of the study. The comparison group students continued to work on teacher-provided material during the enrichment block.

Student participants. After the teachers had agreed to participate in the study, the researcher provided them with an outline and timeline of the study. Subsequently, the teachers invited the students in the classes of their choosing to participate. The teachers were given the freedom to either invite their whole roster of students or specific classes to participate.

A letter (Appendix E) was sent home to student parents/guardians, explaining the study and asking for their consent to allow their student to participate. If guardians did not want their child to participate in the study, the guardians sent back a signed form. Once a guardian consented, the student was given an assent form (Appendix F) to complete if they were interested in participating. The participating teachers collected the completed forms and gave them to the researcher. Three teachers who provided their consent did not provide the researcher with the consent forms for their students. Therefore, those students were not included in the data set. Seven teachers and their students remained in the study.

Of the 743 targeted seventh- and eighth-grade students who were initially invited to participate in the study, consent and assent were received from 413 students. This is a response rate of 55.6%. As seen in Table 7, the targeted population was each teacher's entire roster, and the total participants were the students from whom consent and assent were received.

Table 7

Student Participants per Teacher based on Consent Received

Teacher Identification	Target Participation *n*	Total Participation *n*	Participation Rate by Teacher
1	116	74	63.8
3	99	66	66.7
4	113	27	23.9
5	113	50	44.2
6	112	75	67.0
9	116	86	74.1
10	74	35	47.3
Total	743	413	55.6

This study used a sample of convenience. The teachers decided if they wanted their students to be in the treatment or comparison group. Moreover, the teachers chose to either invite their whole roster of students or simply students from specific classes. A breakdown of teacher and student participation in either the treatment or comparison group is presented in Table 8.

Table 8

Student Participants per Teacher by Group

	Treatment Group		Comparison group		Total	
Teacher Identification	*n*	%	*n*	%	*n*	%
1	0	0	74	54	74	18
3	57	21	9	7	66	16
4	0	0	27	20	27	7
5	50	18	0	0	50	12
6	65	24	10	7	75	18
9	81	29	5	4	86	21
10	23	8	12	9	35	8
Total	276	100	137	101	413	100

Research Questions and Hypotheses

This study was used to examine if different program types (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female) affect student achievement and mathematics class engagement. Using a systematic approach, this research addressed the following research questions:

1. For grade 7 and 8 students, is there a statistically significant difference in posttest student achievement (STAR Assessment, mathematics grade) between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female)?

a. Is there a statistically significant difference in student achievement (STAR Assessment, mathematics grade) between program types (a GBL enhanced mathematics program, a traditional mathematics program)?
b. Is there a statistically significant difference in student achievement (STAR Assessment, mathematics grade) between males and females?
c. Is there a statistically significant interaction between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female) for student achievement (STAR Assessment, mathematics grade)?

Non-Directional Hypothesis: There will be a significant difference in posttest student achievement between gender and program type.

2. For grade 7 and 8 students, to what degree and in what manner do gender (male, female), mathematics grades, and amount of time spent using a virtual mathematics game predict student achievement on the STAR Mathematics Assessment?

Non-Directional Hypothesis: Gender, mathematics grades, and amount of time spent using the virtual grade will predict student achievement on the STAR Mathematics assessment.

3. For grade 7 and 8 students, is there a statistically significant difference in mathematics class engagement (Social, Emotional, Cognitive) between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female)?
    a. Is there a statistically significant difference in mathematics class engagement between males and females?

b. Is there a statistically significant difference in student engagement between the program types being used (a GBL enhanced mathematics program, a traditional mathematics program)?

c. Is there a statistically significant interaction between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female), a traditional mathematics program) for student mathematics class engagement?

Non-Directional Hypothesis: There will be a statistically significant difference in mathematics engagement between gender and program type.

Description of the Research Design

The researcher utilized a quasi-experimental research design. This design was chosen because neither individuals or classrooms could be randomly assigned to the treatment or the comparison group. Gall, Gall, and Borg (2003) call this a non-equivalent control-group design, which is the most commonly used quasi-experimental design. It means that participants are not randomly assigned to groups and that both groups complete both a pretest and a posttest. Teachers selected which classes and students would participate in the study.

Research Question One

For research question one, the data collected for the pretests and posttests were a district measures of student mathematics achievement and mathematics grades. The research design, for question one, is depicted in Figure 1.

Figure 1

Comparison-Group Design Research Question 1

Group	Pretest	Treatment	Posttest
Treatment (Played Prodigy)	O_1	X	O_2
Control Group (Did not play Prodigy)	O_1		O_2

Note. Adapted from Gall et al., 2003.

Research Question Two

Research question two used a correlation design which enabled the researcher to analyze the relationships among multiple variables (Gall et al., 2003). The design can be seen in Figure 2.

Figure 2

Design Research Question 2

Group	Treatment	Posttest
Treatment (Played Prodigy)	X	O_1

Note. Adapted from Gall et al., 2003.

Research Question Three

In research question three, the researcher used a posttest only design for student engagement that was administered after the treatment group played the game and after a typical mathematics class for the comparison group. This is a static-group comparison design depicted in Figure 3.

Figure 3

Static-Group Comparison Design Research Question 3

Group	Treatment	Posttest
Treatment (Played Prodigy)	X	O_1
Control Group (Did not play Prodigy)		O_1

Note. Adapted from Gall et al., 2003.

Description of the Instruments and Data Collection Tools

Data were collected using the following instruments: (a) student demographic survey, (b) teacher demographic survey, (c) STAR Math Assessment, (d) students' mathematics grades, and (e) Student Engagement in Mathematics Scale (SEMS-PCM).

Student Demographic Survey

At the start of the study, the students were asked to complete a brief, researcher-designed demographic survey (Appendix G). It took them approximately 5-10 minutes to complete. The purpose of this survey was to collect background information about the student participants to provide a detailed description of the sample. The survey consisted of 14 items that included information about their demographics, how the student felt about mathematics class, the types and frequency of academic games they played in mathematics class, how often the student played any type of digital game, and the types of computer/game systems they used. All students were asked what gender they identified with, whether male, female, or other. For this study, no student chose other, so only male and female were used in the analysis and report.

Teacher Demographic Survey

At the start of the study, the teachers were asked to complete a brief, researcher-designed demographic survey (Appendix H). It took them approximately 5-10 minutes to complete. The purpose of the survey was to collect background information about the teacher participants to

provide a detailed description of the sample. In addition, the researcher aimed to gain some insight into the teachers' experiences with digital academic games. The survey consisted of 12 items and included questions about their demographics, years of teaching experience, the types of academic games they have used in class, and the frequency with which they have used academic games.

STAR Math Assessment

The STAR Math Assessment (Renaissance Learning Inc., 2014) is a computer-adaptive screening assessment used by students from kindergarten through grade 12 in this school district. The STAR Math Assessment is used to monitor student growth in mathematical concepts, and it helps teachers obtain information about their students' mathematical skills. There are many different models regarding the proper frequency of test administration. For example, it can be used in conjunction with an intervention framework, to monitor student progress, or just to scrutinize yearly growth.

The STAR Math Assessment is administered three times throughout the school year, once in the fall (September), winter (January), and spring (April). For this study, the winter and spring STAR scores were used.

The STAR Math Assessment has a test bank of more than 5,000 items. Each item is a multiple-choice question, and there are four possible answers for each. When a test is administered, the students receive 34 items; on average, students are given 20 minutes to complete the assessment. The STAR Math Assessment considers four domains that are aligned with the Common Core State Standards: numbers and operations, algebra, data analysis, and statistics and probability. Shapiro, Dennis, and Fu (2015) explained, "Within each of the domains are 54 skill sets, for a total of more than 550 core skills" (p. 474). To score the

assessment, the software produces a scaled score between 0 and 1,400, which relates to the relative grade expectation. These scores include a percentile rank of all students who took the test nationally, a scaled score, and a grade level equivalent. The output also provides a student growth percentile that projects student improvement across the year with consideration of their baseline data.

The dependability of the instrument is very high as it has internal consistency reliability coefficients ranging from 0.90 to 0.95. The instrument's retest reliability is also strong, with coefficients ranging from 0.76 to 0.84. For grade levels one through grade level 12, the average predictive validity is 0.72 and the average concurrent validity is 0.63 (Renaissance, 2014).

At the time of the study, the STAR program costs roughly $1,200 per district to use the program plus an additional $8 per student. Specific technology is needed to implement the assessment, including computer and Internet access. Many districts use this platform to obtain standardized data and to monitor student growth. Several districts also use STAR data to determine student placement into mathematics courses and possible math advancement in middle schools. In addition, STAR data are used to help school intervention experts target and monitor the students who need help.

Mathematics Grade

A mathematics grade is based on a combination of data points for each student at the end of each trimester. All teachers follow the district grading policy. For middle schools, there are three categories: homework, worth 10%; assignments, worth 20%; and assessments, worth 70%. Each category has a minimum number of required assignments. Each student is given a numerical grade, though district policy is to change any grade lower than 50 to a grade of 50. Grades for the second and third trimester were used in this study.

Student Engagement in Mathematics Scale

The Student Engagement in Mathematics Scale (SEMS) comprises 18 items rated on a 4-point scale and is divided into three sub-categories: social engagement, emotional engagement, and cognitive engagement (Leis, Schmidt, & Rimm-Kaufmann, 2015). Leis et al. (2015) define these categories as follows: *social engagement* refers to "students' day-to-day social exchange with peers that are tethered to the instructional content"; *emotional engagement* refers to "feelings of connection to content, interest in learning, and enjoyment of solving problems and thinking about content"; and *cognitive engagement* refers to "extent to which children show a willingness to exert effort to understand content, work through difficult problems, and manage and direct their attention toward the task at hand" (p. 253). Based on their analysis of the scale, Leis et al. (2015) suggested eliminating items 2, 3, 9, and 12 in future administration of the questionnaire. Therefore, these items were removed from the final analysis of the current research. The emotional engagement subscale consists of five items with a person-separation reliability of 0.74 for non-extreme persons and 0.66 for all respondents. The social engagement subscale consisted of four items with a person-separation reliability of 0.62. The cognitive subscale consisted of five items with an item separation reliability value of 0.96. The pilot study indicated the instrument had a high-reliability coefficient, Cronbach's alpha = 0.90, and a confirmatory factor analysis resulted in the three factors.

Description and Justification of the Analyses

The data for the research questions were collected and first entered into Microsoft Excel Version 16.38. Next, the data were transferred into the statistical package SPSS for further analysis.

Research Question One

Research question one was used to investigate whether a significant difference existed in pre and post student achievement (STAR Assessment, mathematics grade) for grade 7 and 8 students between gender (male, female) and program type (GBL enhanced mathematics program, traditional mathematics program). A two-way multivariate analysis of variance (MANOVA) was used to analyze the dependent variables, student achievement, which was measured by postscores on the Spring STAR Assessment and the Trimester 3 mathematics grades. The independent variables, program type, consisted of two levels: GBL enhanced mathematics program (Treatment) and traditional mathematics program (Comparison). The second independent variable, gender, consisted of two levels: male and female. As Meyers, Gamst, and Guarino (2006) stated, MANOVAs are used when one is, "interested in assessing the effects of [at least] one dichotomous (two-group) independent variable on two or more quantitative dependent variables" (p. 365). Therefore, a MANOVA was the appropriate test to use since the effects of student achievement, as measured by two outcome variables, were examined in relation to both gender and program type.

Research Question Two

To address research question two a stepwise multiple regression analysis of collected data was conducted to answer this question. The dependent variable was the students' spring STAR scores, and the independent variables were gender (male, female), trimester 3 (T3) mathematics grade, and time spent playing the virtual mathematics game in hours. All variables, except gender, were on an interval-level measurement scale.

Multiple linear regression is appropriate to use as this research question involved a single metric dependent variable, student achievement on STAR, that was presumed to be related to the independent variables listed above (Hair, Black, Babin, & Anderson 2010).

Research Question Three

Research question three was developed to explore whether there was a statistically significant difference in mathematics class engagement in relation to gender (male, female) and Program type (a GBL enhanced mathematics program, a traditional mathematics program). A MANOVA was used to answer this question. The dependent variables variable consisted of the student posttest scores on the Student Engagement in Mathematics Scale across three subscales: Cognitive, Social, and Emotional engagement. The three subscales represented interval-level data. The independent variables were (a) Program type, which consisted of two levels: a GBL enhanced mathematics program (Treatment) and a traditional mathematics program (Comparison); and (b) Gender, which consisted of two levels: male and female. A MANOVA was the appropriate test to use as per Hair et al. (2010): "a MANOVA is useful when the researcher designs an experimental situation to test hypotheses concerning the variance in group responses on two or more metric dependent variables" (p. 18).

Description of the Activities

Treatment Group Activities

At the start of the study, the researcher trained each teacher to use Prodigy (Peters and Mahimker, 2011), the digital game platform. The teachers were also provided with additional resources, such as directions and webinars, which allowed them to further review the information included in the training. The researcher met weekly with the teachers to monitor their progress. In addition, the researcher made herself available via email if a teacher had a question or needed

additional information. The teachers were asked to assign 30 minutes a week for students to work on the program during their enrichment classes.

Prodigy is an interactive, virtual, role-playing game in which students engage in friendly mathematics battles. To battle, students must answer questions correctly. Teachers can select questions or topics; if preferred, Prodigy's adaptive learning system will customize content for the students based on their pre-assessments. Prodigy questions are either multiple choice or free response, and they are aligned to the curriculum standards. Prodigy has 1,300 math skills for grades one through eight along with 50,000 mathematics questions.

Comparison Group Activities

The comparison group followed the traditional mathematics curriculum approved by the district. After the study was conducted, the comparison group was offered the treatment as well.

Posttest Administration

At the end of the study, the teachers administer the STAR assessment, based on the school assessment calendar that lined up with the study, and were asked to administered the Student Engagement in Mathematics Scale (SEMS; Leis et al., 2015).

Description of Data Collection Procedures and Timeline

The following procedures were followed according to the proposed timeline. The researcher recorded the actual dates in a dissertation journal.

1. Submitted proposal for IRB approval (January 31, 2019); subsequently, approval was granted (February 19, 2019).
2. Requested and received permission from the district superintendent and assistant superintendent (March 6, 2019).

3. Requested and received permission from the school principals to conduct the study at the chosen middle schools, School A (March 13, 2019) and School B (March 19, 2010)
4. Distributed, collected, and received consent from the teachers at School A (March 14, 2019) and School B (March 26, 2019) to be a part of the study.
5. Received consent from parents and assent from students (March 2019)
6. Provided training for teacher participants; clearly outlined the specific steps and expectations for the study; administered the Teacher Demographic Survey at School A (March 14, 2019, and April 2, 2019) from School B (April 2, 2019).
7. Collected students' trimester two grades and Winter STAR scores (April 2019).
8. Administered the Student Demographic Survey (March and April 2019).
9. Provided weekly check-ins and support to participating teachers (April 2019 – June 2019).
10. Teachers administered the SEMS scale and the Spring STAR assessment (June 2019).
11. Researcher collected students' trimester three grades (June 2019).
12. Conducted data analysis (June 2019 – December 2021).

Ethics Statement

Approval to conduct the study was obtained from the Institutional Review Board (IRB) of Western Connecticut State University. Permission to participate in this research was obtained from the school district's superintendent, assistant superintendent, the two school principals, and the 10 participating teachers. Letters were sent home with all students in the study to obtain consent from their parents or guardians. All students were informed about the study and invited to assent to participate. All participants were informed that their participation was voluntary and

that they could withdraw at any time. To ensure confidentiality, each participant was assigned a coded identification number. These coded data were made available to the researchers related to the study for the purposes of data verification and analysis. When not being analyzed, the data were locked in a data file by the researcher to protect the school, teacher, and student privacy. The aggregated data were only made available to the researcher's co-advisory and interested parties upon request.

CHAPTER 4: DATA ANALYSIS AND RESULTS

The purpose of this study was to investigate the impact of playing a virtual mathematics game, Prodigy, on students' mathematics achievement scores and grades in mathematics between gender (male, female) compared to students who did not play the game. In addition, the impact on student engagement between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female) was investigated. Chapter four presents the findings and statistical procedures related to the research questions for this study. The results are presented in eight sections: (a) research questions and hypothesis, (b) description of the data, (c) demographic results, (d) analysis of outliers, (e) descriptive statistics, (f) quantitative data analysis for research question one, (g) quantitative data analysis for research question two, and (h) quantitative data analysis for research question three.

Research Questions and Hypotheses

Using a systematic approach, this research addressed the following research questions:

1. For the grade 7 and 8 students, is there a statistically significant difference in post student achievement (STAR Assessment, mathematics grade) between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female)
    a. Is there a statistically significant difference in student achievement (STAR Assessment, mathematics grade) between program types (GBL enhanced mathematics program, traditional mathematics program)?
    b. Is there a statistically significant difference in student achievement (STAR Assessment, mathematics grade) between males and females?

c. Is there a statistically significant interaction between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female) for student achievement (STAR Assessment, mathematics grade)?

Non-Directional Hypothesis: There will be a significant difference in pre and post student achievements between gender and program type.

2. For grade 7 and 8 students, to what degree and in what manner do gender (male, female), mathematics grade, amount of time spent using the virtual mathematics game predict student achievement on the STAR Assessment?

Non-Directional Hypothesis: Gender, mathematics grades and amount of time spent using the virtual grade will predict student achievement on STAR assessment.

3. For grade 7 and 8 students, is there a statistically significant difference in mathematics class engagement (social, emotional, and cognitive) between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female)?
    a. Is there a statistically significant difference in mathematics class engagement (social, emotional, and cognitive) between males and females?
    b. Is there a significant difference in student mathematics engagement (social, emotional, and cognitive) between program types (GBL enhanced mathematics program, traditional mathematics program)?
    c. Is there a significant interaction between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female) for student mathematics class engagement (social, emotional, and cognitive)?

Non-Directional Hypothesis: There will be a significant difference in mathematics engagement (social, emotional, and cognitive) between gender and program type.

Description of the Data

Research Questions One

To complete the MANOVA, data from two dependent variables were collected. Interval level data from the STAR assessment, Winter STAR for the pre-scores and Spring STAR for the post-scores, were collected. In addition, data from students' mathematics grades, T2 (Trimester 2) for the pre-scores and T3 (Trimester 3) for the post-scores, were collected. The independent variables were as follows: (a) Gender, two levels: male and female, and (b) Program type, two levels: traditional mathematics program and GBL enhanced mathematics program. The dependent variable for research question one was Spring STAR, which consisted of a total score, and T3 mathematics grade.

Research Question Two

For this regression analysis quantitative data from the STAR assessment, Winter STAR and Spring STAR, were collected, as well as mathematics grades, T2 and T3. In addition, data from time spent using a virtual math program, were collected, which consisted of a total time in hours students had played the virtual mathematics game. The independent variables were as follows: (a) Gender, two levels: male and female, (b) Mathematics grade, one level: T3, and (c) Amount of time continuous data. The dependent variable was Spring STAR, which consisted of a total score.

Research Question Three

The Student Engagement in Mathematics Scale (SEMS) data were represented the dependent variables for research question three. This instrument was scored using a 4-point

scale and was divided into three sub-categories: emotional engagement, social engagement, and cognitive engagement (Leis, Schmidt, & Rimm-Kaufmann, 2015). The independent variables were Program type, two levels: traditional mathematics program and GBL enhanced mathematics program and gender, two levels: male and female. There were three dependent variables for Student Engagement in Mathematics Scale, students' mean scores on each of the three subscales: (a) emotional engagement, (b) social engagement, and (c) cognitive engagement.

This information for each research question is summarized in Table 9. The total number of participants showed in the table reflect the final count after data cleaning and checking assumptions for analysis.

Table 9

Dependent and Independent Variables Based on Research Question

Research Question	Dependent Variable	Independent Variable
One	Mathematics Grades (T3) Spring STAR Scores (n = 200)	Program Type (treatment: n = 84; comparison: n = 116), Gender (males: n = 100; females: n = 100)
Two	Spring STAR Scores (n = 79)	Amount of Time in Hours Gender (males: n = 47; females: n = 32), Mathematics Grade
Three	Emotional engagement, Social engagement, and Cognitive engagement (n = 117)	Program Type (treatment: n = 65; comparison: n = 52), Gender (males: n = 62; females: n = 55)

Data Coding and Entry

Prior to analyzing the data, the researcher combined multiple data sources into one spreadsheet, coded, verified, and cleaned the data. This process, according to Meyers et al. (2006) is referred to as, "code and value cleaning where the researcher follows the checks for the appropriateness of the numerical codes for the values of each variable under study" (p. 44).

The data set included student names, short names which are district assigned unique email name, student ID number, mathematics teacher name, grade level, school, and ethnicity. The researcher then assigned a Unique ID to each student and teacher. Once the demographic data were assembled the researcher added to that data file the students' T2 and T3 mathematics grades along with their Winter and Spring STAR scores. This was done using an Excel formula of VLookUp to match the data accurately in Microsoft Excel for Mac v. 16.49 (Microsoft, 2021). This formula was used multiple times in the data cleaning process. Next, the researcher matched students to their Prodigy usernames by looking at the teachers' rosters of students' names and matching them accordingly, the students who signed in using Google were able to be matched using the Excel formula, but those who created their own username had to be matched by hand. Data were collected weekly from Prodigy, which was then combined into one data set. The researcher then matched the individual Prodigy student data, how much time the student spent playing each week and the number of questions answered, from each week of the study to the main data set. In order to determine the total number of weeks each student played Prodigy over the course of the study; the researcher utilized the data collected to create an additional column to calculate the total number of weeks played. Additionally, the amount of time spent playing each week was combined and captured in a separate column and converted from minutes to hours to calculate the total number of hours spent playing. Finally, the researcher calculated the

total number of questions answered throughout all Prodigy sessions in a separate column by combining the weekly data.

The next task was to assemble the Student Engagement in Mathematics Scale data. First, question 7 in the engagement scale was reverse coded. Then the mean for, questions 1, 9, 10, and 13 was calculated for cognitive engagement; the mean for questions 2, 3, 4, and 5 was calculated for social engagement; and the mean for questions, 6, 7, 8, 11 and 12 was calculated for emotional engagement. The mean scores were then matched to the correct students in the Microsoft Excel for Mac v. 16.49 (Microsoft, 2021) data set.

Once all the data were combined, the researcher eliminated the students' names, short names and student ID to create a codebook. These data were then uploaded to SPSS v. 27 (SPSS Inc., 2020) where codes were assigned to the data to create the codebook. From the codebook in SPSS, the researcher kept the Unique ID for each student, ethnicity, grade, teacher code, and school code and combined those with the additional data from the independent and dependent variables; entries for the codebook are presented in Table 10. The researcher changed the Prodigy Username to the code Prodigy, two-levels, yes or no. This code book was used to ensure that consistent codes and information were entered into SPSS v. 27 (SPSS Inc., 2020). The additional codes were assigned to the data representing the independent and dependent variables as presented in Table 11.

Table 10

SPSS Codebook of Student Demographic Variables

Code Name	Type of SPSS Field	Measurement Scale	Assigned Values
Unique ID	Numeric,	Nominal	1 – 663
Gender Identity	Numeric	Nominal	0 = Male 1 = Female
School	String	Nominal	A = School A B = School B
Grade Level	Numeric	Nominal	7 = 7th Grade 8 = 8th Grade
Teacher	String	Nominal	1T = Teacher 1 through 10T = Teacher 10
Program Type	Numeric	Nominal	0 = No (0 weeks playing Prodigy) 1 = Yes (Played Prodigy for 7+ weeks)
Identity	String	Nominal	A = Asian America B = Black or African American H = Hispanic or Latino of any race M = Two or More W = White

Table 11

SPSS Codebook of Independent and Dependent Variables

Label	Code Name	Type of SPSS Field	Measurement Scale	Assigned Values
Winter STAR Score	Winter STAR	Numeric	Continuous	492 – 1023
Spring STAR Score	Spring STAR	Numeric	Continuous	431 – 1056
T2 Grades	T2Grades	Numeric	Continuous	50 – 100
T3 Grades	T3Grades	Numeric	Continuous	50 – 100
SEMS Cognitive Score	SEMSPostCognitve	Numeric	Continuous	1.75 – 4
SEMS Social Score	SEMSPostSocial	Numeric	Continuous	1 – 4
SEMS Emotional Score	SEMSPostEmotional	Numeric	Continuous	1 – 4
Total Time in Hours	TotalTimeinHoursroundedtohundreths	Numeric	Continuous	0 – 20.11
Total Number of Questions Answered	TotalNumberofQuestionsAnswered	Numeric	Continuous	0 – 2814
Number of Weeks Prodigy was Played	NumberofWeeksPlayed	Numeric	Continuous	0-10

Data Screening and Cleaning

Prior to any analysis, the data were screened and cleaned which, as Meyers et al. (2006) states, is a verification procedure that checks for appropriateness, such as legitimate numerical codes or values, for each variable under study. First, a visual inspection was conducted of the data to determine if there were any missing data, if any extreme values were noticed. Through this process, the researcher noticed that there was an inconsistency in the number of weeks that students played Prodigy. The researcher first checked the number of weeks students played or participated in the treatment using a frequency table. The researcher found 28 students who were selected to the treatment but did not participate. These students were eliminated from the study

because they had exposure to the treatment but did not participate in playing Prodigy. Of these 28 students who played zero weeks, 23 of them were female and five were male.

To better understand the data and population, the researcher then grouped the students into different brackets based on how many weeks they played Prodigy. The groups were as follows; Group 0 included students in the comparison group who did not use Prodigy, Group 1 represented students who played the game for 1-3 weeks, Group 2 included students who played the game for 4-6 weeks, and Group 3 included students who played for seven weeks or more. While the study ended after eight weeks, some students continued to play Prodigy during their spring break. The frequency for students who completed the study can be found in Table 12.

Table 12

Frequency of Participants Who Completed the Study

Group	Frequency (n = 385)	Percentage
Comparison Group		
0 (did not play prodigy)	137	35.6
Treatment Group		
1 (1-3 weeks)	72	18.7
2 (4-6 weeks)	83	21.6
3 (7+ weeks)	93	24.2
Total	385	100.0

There was a total of 93 students who consistently played Prodigy during the study, some even played during school vacation. The other 155 students in Groups 1 and 2 did not continuously play the game during the enrichment block due to many factors, such as the teachers not enforcing that the students play the game, especially if students had other work to

complete during this time. For this reason, it was decided that for the analysis of all research questions, Groups 0 and 3 were used, $n = 230$, which is 59.8% of the total sample. These groups were chosen to study for maximal variation.

The researcher then examined the data set for any missing data or data that were not in the correct numerical or value form. No data were found that used an incorrect format. There were no missing values recorded for gender, program type, T2 or T3 mathematics grades, Winter or Spring STAR. There were missing data for the engagement scale. This resulted in two data sets being used to address the three research questions. The students who were missing the engagement scale data were not included in research question three but were represented in research questions one and two.

Through the cleaning process, the researcher was able to find and eliminate the errors that were noted above. Using the two data sets the researcher then continued with the analysis process.

Demographic Results

Participating students were asked to complete a Student Demographic Form (Appendix G) to obtain additional information from the students, these results were then cross referenced with the grade report sheet provided to the researcher from the teachers. This information was used to better describe the population of the students involved in the study. Table 13 presents the breakdown for Gender and Grade level of student participants based on treatment and comparison groups.

Table 13

Gender and Grade of Student Participants between Comparison and Treatment Groups

Group	Male		Female			7th Grade		8th Grade		
	n	Percent	*n*	Percent	Total	*n*	Percent	*n*	Percent	Total
Comparison	65	55.1	72	64.3	137	110	60.1	27	57.4	137
Treatment	53	44.9	40	35.7	93	73	39.9	20	42.6	93
Total	118	100.0	112	100.0	230	183	100.0	47	100.0	230

In this study there were more students in the treatment group (*n* = 137 or 59.6%) than in the comparison group (*n* = 93 or 40.4%). There were more 7th grade students (79.6%) than 8th grade students (20.4%), therefore, no analysis was done to compare the two groups by grade level.

The middle schools, where this study took place, were ethnically diverse. The breakdown of ethnicity for the district includes: 6.5% Asian, 6.9% Black or African American, 50.6% Hispanic, and 33.0% White (CSDE, 2018). The sample size reflects this diversity as illustrated in Tables 14. The only ethnic status that was not aligned with the district breakdown was Black or African American which made up only 1.1% of the student treatment group sample compared to the district average of 6.9% of the student population.

Table 14

Ethnicity of Treatment and Comparison Group Student Participants

Group	Comparison Group		Treatment Group		Total	
	n	Percent	*n*	Percent	*n*	Percent
Asian American	12	8.8	7	7.5	19	8.3
Black or African American	15	10.9	1	1.1	16	7.0
Hispanic or Latino of any race	68	49.6	48	51.6	116	50.4
Two or More	3	2.2	3	3.2	6	2.6
White	39	28.5	34	36.6	73	31.7
Total	137	100.0	93	100.0	230	100.0

Analysis of Outliers

Outliers are distinctly different and stand out from the observation set with a value that is unusually high or low (Hair et al., 2010). In order to identify the outliers in the data set, the researcher first examined each variable's box and whisker plot. As Meyers et al. (2006) stated, "box plots convey a considerable amount of information about that distribution in one fairly condensed display" (p. 52). While analyzing the box plots, regions beyond the inner fences are extreme scores and a data point that is farther than ±1.5 IQRs (inter quartile range) but less than ±3.0 IQRs is labeled as an outlier by SPSS (Meyers et al., 2006). Once the outliers were visually detected in this way, the researcher also used the rule by Hair et al. (2010) that outliers are defined as cases with scores of ±2.5 standard deviations from the mean.

To analyze for outliers, the researcher used SPSS to calculate the descriptive statistics for all variables beginning with the pretest data of Winter STAR and T2 mathematics grades with a focus on gender followed by an analysis for program type using the same dependent variables.

The data analysis started with 118 males and 112 females. There were 137 students in the control group and 93 in the treatment group.

There was a total of six rounds to identify and eliminate outliers in the data set. Table 15 shows the students who were identified as outliers from the box and whisker plots. If the student values for any indicator were 2.5 standard deviations below or above the mean they were eliminated from the data set.

Table 15

Investigation of Potential Outliers for Pretest and Posttest Data for Research Question One

						Criteria		
Case	IV	DV	Case Mean	Group Mean	*SD*	2.5 *SD* Below Mean	2.5 *SD* Above Mean	Eliminated
				Round 1				
6	Female	T2 Grades	50	85.44	10.114	60	111	Yes
51	Female	T2 Grades	60	85.44	10.114	60	111	Yes
98	Female	T2 Grades	50	85.44	10.114	60	111	Yes
337	Female	T2 Grades	50	85.44	10.114	60	111	Yes
28	Male	T2 Grades	50	80.69	10.661	54	107	Yes
46	Male	T2 Grades	54	80.69	10.661	54	107	Yes
89	Male	T2 Grades	50	80.69	10.661	54	107	Yes
28	Comparison	T2 Grades	50	81.82	11.211	54	110	Yes
46	Comparison	T2 Grades	54	81.82	11.211	54	110	Yes
89	Comparison	T2 Grades	50	81.82	11.211	54	110	Yes
98	Comparison	T2 Grades	50	81.82	11.211	54	110	Yes
337	Comparison	T2 Grades	50	81.82	11.211	54	110	Yes
186	Treatment	T2 Grades	58	84.74	9.546	61	109	Yes
46	Male	Winter STAR	501	770.12	98.754	523	1017	Yes
89	Male	Winter STAR	502	770.12	98.754	523	1017	Yes
240	Male	Winter STAR	494	770.12	98.754	523	1017	Yes
297	Male	Winter STAR	494	770.12	98.754	523	1017	Yes
46	Comparison	Winter STAR	501	759.45	96.781	517	1001	Yes
89	Comparison	Winter STAR	502	759.45	96.781	517	1001	Yes
240	Comparison	Winter STAR	494	759.45	96.781	517	1001	Yes
297	Comparison	Winter STAR	494	759.45	96.781	517	1001	Yes

Case	IV	DV	Case Mean	Group Mean	*SD*	Criteria 2.5 *SD* Below Mean	Criteria 2.5 *SD* Above Mean	Eliminated
				Round 2				
43	Female	T2 Grades	67	86.66	7.96	67	107	Yes
135	Female	T2 Grades	66	86.66	7.96	67	107	Yes
192	Female	T2 Grades	65	86.66	7.96	67	107	Yes
323	No	T2 Grades	60	83.65	8.88	61	106	Yes
34	No	T2 Grades	58	83.65	8.88	61	106	Yes
13	Female	Winter STAR	996	762.52	86.093	547	978	Yes
214	Male	Winter STAR	557	781.99	84.695	570	994	Yes
323	Male	Winter STAR	546	781.99	84.695	570	994	Yes
323	No	Winter STAR	546	772.66	83.568	564	982	Yes
				Round 3				
137	Male	T2 Grades	62	82.45	8.824	60	105	No
160	Yes	T2 Grades	64	85.30	8.790	63	107	No
218	No	Winter STAR	566	777.00	80.679	575	979	Yes
281	No	Winter STAR	573	777.00	80.679	575	979	Yes
				Round 4				
137	Male	T2 Grades	62	82.45	8.824	60	105	No
160	Yes	T2 Grades	64	85.30	8.790	63	107	No
				Round 5				
233	Female	Spring STAR	546	779.19	80.627	578	981	Yes
154	Female	Spring STAR	508	779.19	80.627	578	981	Yes
110	Male	Spring STAR	578	808.65	83.358	600	1017	Yes
293	Male	Spring STAR	549	808.65	83.358	600	1017	Yes
293	No	Spring STAR	549	795.41	79.083	598	993	Yes
233	No	Spring STAR	546	795.41	79.083	598	993	Yes
154	Yes	Spring STAR	508	793.03	88.915	571	1015	Yes
223	Male	T3 Grades	50	80.40	10.044	55	106	Yes
223	No	T3 Grades	50	81.89	10.039	57	107	Yes
				Round 6				
137	No	T2 Grades	62	84.82	7.861	65	104	Yes
160	Yes	T2 Grades	64	85.31	8.890	63	108	No

There was a total of 25 outliers identified. After the final analysis for outliers, the sample size was determined to be 205 in total as shown in Table 16.

Table 16

Gender of Student Participants in Treatment and Comparison Groups after Data Cleaning

Group	Treatment Group		Comparison Group		Total	
	n	Percent	*n*	Percent	*n*	Percent
Male (*n* = 105)	50	57.5	55	46.6	105	51.2
Female (*n* = 100)	37	42.5	63	53.4	100	48.8
Total	87	100.0	118	100.0	205	100.0

Once the outliers were removed the researcher validated that the skewness and kurtosis were within the range for normality meaning there was a symmetrical distribution of the data. This was done first with a visual check of the histograms to make sure they had a normal distribution (Hair et al., 2010) and then by checking the skewness and kurtosis values for all dependent variables by the categories for each independent variable to make sure that values were not greater than ±1.00 (Meyers et al., 2006) as presented in Tables 17 and 18, for gender and program type, respectfully.

Table 17

Skewness and Kurtosis Values by Gender

Variable	Skewness		Kurtosis	
	Male (*n* = 105)	Female (*n* = 100)	Male (*n* = 105)	Female (*n* = 100)
Winter STAR	-0.359	-0.238	-0.675	-0.464
Spring STAR	-0.603	-0.456	-0.219	-0.407
T2 Grades	-0.351	-0.578	-0.325	-0.005
T3 Grades	-0.104	-0.477	-0.513	-0.368

Table 18

Skewness and Kurtosis Values by Program Type

Variable	Skewness		Kurtosis	
	Comparison (n = 118)	Treatment (n = 87)	Comparison (n = 118)	Treatment (n = 87)
Winter STAR	-0.364	-0.193	-0.457	-0.734
Spring STAR	-0.567	-0.381	-0.205	-0.550
T2 Grades	-0.457	-0.592	-0.113	-0.203
T3 Grades	-0.254	-0.350	-0.639	-0.364

As recommended by Meyers et al. (2006) the Kolmogorov-Smirnov and Shapiro-Wilk tests of normality were not significant at $p < .001$, resulting in the conclusion that the assumption of normality was not violated. The tests of normality can be seen in Table 19 and Table 20 for gender and program type, respectively.

Table 19

Tests of Normality Achievement by Gender

		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	*df*	Sig.	Statistic	*df*	Sig.
Winter STAR	Male	0.086	105	0.056	0.970	105	0.017
	Female	0.050	100	0.200	0.986	100	0.390
Spring STAR	Male	0.089	105	0.040	0.962	105	0.004
	Female	0.093	100	0.032	0.971	100	0.025
T2 Grade	Male	0.075	105	0.179	0.975	105	0.048
	Female	0.090	100	0.046	0.966	100	0.011
T3 Grades	Male	0.050	105	0.200	0.987	105	0.401
	Female	0.082	100	0.094	0.969	100	0.017

Table 20

Tests of Normality Achievement by Program Type

		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	*df*	Sig.	Statistic	*df*	Sig.
Winter STAR	Comparison	0.090	118	0.019	0.972	118	0.016
	Treatment	0.073	87	0.200	0.975	87	0.090
Spring STAR	Comparison	0.093	118	0.013	0.970	118	0.009
	Treatment	0.110	87	0.012	0.972	87	0.059
T2 Grade	Comparison	0.074	118	0.165	0.976	118	0.036
	Treatment	0.095	87	0.049	0.957	87	0.006
T3 Grades	Comparison	0.069	118	0.200	0.979	118	0.064
	Treatment	0.059	87	0.200	0.979	87	0.162

In addition, after examining the Normal Q-Q plots the distribution was normal because, "the line representing the actual data distribution closely follows the diagonal" (Hair et al., 2010, p. 72).

Descriptive Statistics

Descriptive statistics for the quantitative data, Winter STAR, Spring STAR, T2 mathematics grades, T3 mathematics grades, were calculated for the treatment and comparison groups as well as for gender. Winter STAR and T2 mathematics grades were designated as pretest data and Spring STAR and T3 mathematics grades were identified for use as posttest data, seen in Tables 21 through 24.

Table 21

Pretest Data for Program Type (n = 205)

Variable	Program Type	Minimum	Maximum	Mean	Standard Deviation
Winter STAR	Treatment	592	916	778.40	77.802
Winter STAR	Comparison	609	920	782.68	73.598
T2 Grades	Treatment	64	100	85.31	8.890
T2 Grades	Comparison	65	99	85.01	7.605

Table 22

Pretest Data for Gender (n = 205)

Variable	Gender	Minimum	Maximum	Mean	Standard Deviation
Winter STAR	Male	625	920	790.12	75.255
Winter STAR	Female	592	910	771.14	74.385
T2 Grades	Male	64	100	81.27	8.500
T2 Grades	Female	68	99	87.47	7.100

Table 23

Posttest Data for Program Type (n = 205)

Variable	Program Type	Minimum	Maximum	Mean	Standard Deviation
Spring STAR	Treatment	613	944	798.78	81.080
Spring STAR	Comparison	598	945	799.80	72.842
T3 Grades	Treatment	58	100	83.61	9.557
T3 Grades	Comparison	60	100	82.45	9.548

Table 24

Posttest Data for Gender (n = 205)

Variable	Gender	Minimum	Maximum	Mean	Standard Deviation
Spring STAR	Male	613	945	813.78	76.923
Spring STAR	Female	598	910	784.23	72.898
T3 Grades	Male	58	100	80.81	9.639
T3 Grades	Female	62	100	85.18	8.959

Research Question One

Research question one was used to investigate whether a significant difference existed in posttest student achievement (STAR Assessment, mathematics grade) between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female) for middle school students. A two-way multivariate analysis of variance (MANOVA) was used to analyze the dependent variables, achievement, measured by postscores on the Spring STAR and T3 mathematics grades. The independent variable, program type, consisted of two levels: played Prodigy (Treatment) and did not play Prodigy (Comparison). The second independent variable, gender, consisted of two levels: male and female.

Testing the Assumptions for Pre-Data

Pretest data were analyzed because in order to perform a two-way MANOVA the assumptions need to be tested. The independent variables were as follows: (a) program type, two levels: Played Prodigy and Did Not Play Prodigy and (b) gender, two levels: male and female. The dependent variables were pretest Winer STAR and T2 mathematics grades. Meyers et al.

(2006) stated the statistical assumptions and limitations for a MANOVA are "normality, linearity, and homoscedasticity" (p. 67).

Normality. The researcher first reviewed descriptive statistics to check the normality by looking at the skewness and kurtosis values, all within the ±1 range for gender and program type. The test of normality was assumed based on the acceptable values for the Kolmogorov-Smirnov and Shapiro-Wilk procedures.

Linearity. Next, to assess linearity, the researcher checked that the variables were related to each other in a linear manner and computed the Pearson correlation coefficient to assess the degree of the linear relationship observed (Meyers et al., 2006). To be statistically significant the computed correlation coefficient reaches the .05 alpha level which means that correlation occurs by chance 5% or less of the time if the null hypothesis is true (Meyers et al., 2006). The Pearson correlation coefficient indicates the strength of the relationship between two variables. If the value of the correlation coefficient is between .30 and .49 then there is a medium relationship, and if the value of the correlation coefficient is .50 to 1 it is a larger correlation (Pallant, 2016). The values in Tables 25 and 26 respectively for gender and program type. These correlations were calculated by splitting the sample to compare correlations coefficients for two groups of participants (Pallant, 2016). All correlations indicated a linear relationship among the variables.

Table 25

Correlations by Gender Pretest Data (n = 205)

Gender Identity		T2 Grades
Male (*n* = 105)	Winter STAR	.656***
Female (*n* = 100)	Winter STAR	.493***

Note. ***$p < .001$; Bonferroni adjustment, $p < 3/.05 = .017$

Table 26

Correlations by Program Type Pretest Data (n = 205)

Program Type		T2 Grades
Comparison (*n* = 118)	Winter STAR	.547***
Treatment (*n* = 87)	Winter STAR	.491***

Note. ***$p < .001$; Bonferroni adjustment, $p < 3/.05 = .017$

Calculating all the correlations without splitting the groups produced the correlations below in Table 27.

Table 27

Summary of Correlations for all Pretest Data (n = 205)

Variables	Gender	Winter STAR	T2 Grades
Program Type	-.107	-.028	.018
Gender		-.126	.280**
Winter STAR			.519**

Note: For program type, comparison = 0 and treatment = 1. For gender, males = 0 and females = 1. **$p < .001$; Bonferroni adjustment, $p < 3/.05 = .017$

Homoscedasticity. Homoscedasticity, "refers to the assumption that dependent variable(s) exhibit equal levels of variance across the range of predictor variable(s)" (Hair et al., p. 74, 2010). To check homogeneity of variance violations, the Levene's test was used, which assessed the statistical hypothesis of equal variances across the levels of the independent variable (Meyers et al., 2006). When there is more than one quantitative dependent variable the Box's Test of Equality of Covariance Matrices tests for equality of variance-covariance matrices, which is used to test for homoscedasticity (Meyers et al., 2006). Both analyses were conducted for the pretest scores. The Box's M, using a more conservative alpha level of .001, was found not to be significant (Box's M = 11.822, $p > 0.237$) and the Levene's Test, using the significance value of

.05, for Winter STAR was not significant ($p = 0.834$), but there was a significant Levene's value for T2 math grades ($p = 0.016$) (Pallant, 2016).

To address this violation the researcher created an equal spread of variance across the independent variable gender by making the male ($n = 105$) and female ($n = 100$) sample sizes the same. The researcher randomly eliminated five males, two from the comparison group and three from the treatment group, one at a time to check for any outliers. Once that was completed, the researcher tested the assumptions.

The updated comparison of means for the pretest data for STAR scores can be found below in Table 28.

Table 28

Summary of Comparison Means for Pretest STAR Scores with equal gender (n = 200)

			Pretest Winter STAR		Pretest T2 Grades	
Program	Gender	*n*	Mean	*SD*	Mean	*SD*
Played Prodigy	Male	47	795.17	75.198	83.83	9.574
	Female	37	758.78	76.349	88.08	6.689
Total		84	779.14	77.411	85.70	8.644
Did Not Play Prodigy	Male	53	786.00	75.563	82.58	7.350
	Female	63	778.40	72.842	87.11	7.360
Total		116	781.87	73.871	85.04	7.665
Grand Mean			780.73	75.198	799.72	75.263

Revised Normality. Normality was rechecked because the sample size was reduced from 205 to 200, for the dependent variables and were found to be normally distributed. As

demonstrated in Tables 29 and 30 all kurtosis and skewness values were within the ±1 acceptable range (Hair et al., 2010).

Table 29

Skewness and Kurtosis Values by Gender Pretest Data (n = 200)

Variable	Skewness		Kurtosis	
	Male (n = 100)	Female (n = 100)	Male (n = 100)	Female (n = 100)
Winter STAR	-0.358	-0.238	-0.636	-0.464
T2 Grades	-0.342	-0.578	-0.354	-0.005

Table 30

Skewness and Kurtosis Values by Group Pretest Data (n = 200)

Variable	Skewness		Kurtosis	
	Comparison (n = 116)	Treatment (n = 84)	Comparison (n = 116)	Treatment (n = 84)
Winter STAR	-0.343	-0.215	-0.477	-0.673
T2 Grades	-0.467	-0.595	-0.145	-0.152

The test of normality was also reviewed using Kolmogorov-Smirnov and Shapiro-Wilk. The criterion is that the significance level should be higher than .001 (Meyers et al., 2007). Those values can be seen in Table 31 and 32 for gender and program type, respectively.

Table 31

Tests of Normality Pretest by Program Type (n = 200)

		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	*df*	Sig.	Statistic	*df*	Sig.
Winter STAR	Comparison	0.087	116	0.031	0.973	116	0.020
	Treatment	0.072	84	.200*	0.976	84	0.114
T2 Grade	Comparison	0.076	116	0.097	0.975	116	0.031
	Treatment	0.087	84	0.175	0.957	84	0.007

Note. *$p < .001$

Table 32

Tests of Normality Pretest by Gender (n = 200)

		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	*df*	Sig.	Statistic	*df*	Sig.
Winter STAR	Male	0.084	100	0.081	0.971	100	0.026
	Female	0.050	100	.200*	0.986	100	0.390
T2 Grade	Male	0.076	100	0.171	0.975	100	0.057
	Female	0.090	100	0.046	0.966	100	0.011

Note. *$p < .001$

Histograms were generated to visually check normality using SPSS v.27 and shown in Figures 4 through 11.

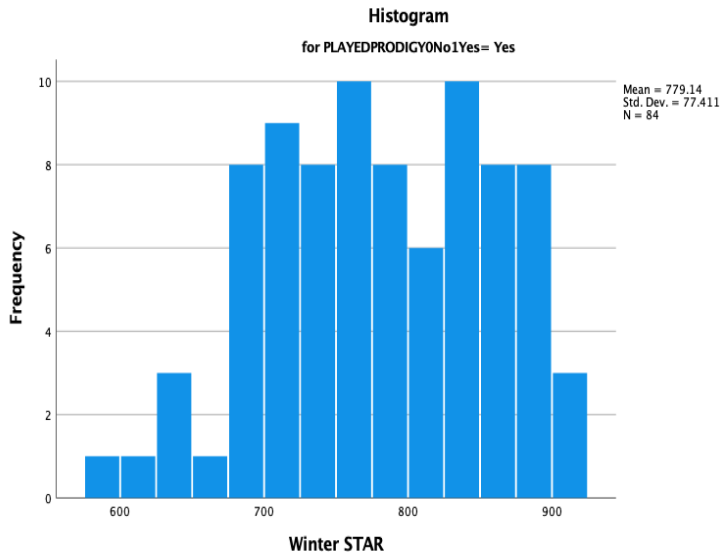

Figure 4. Histogram of the mean pretest Winter STAR scores for treatment group.

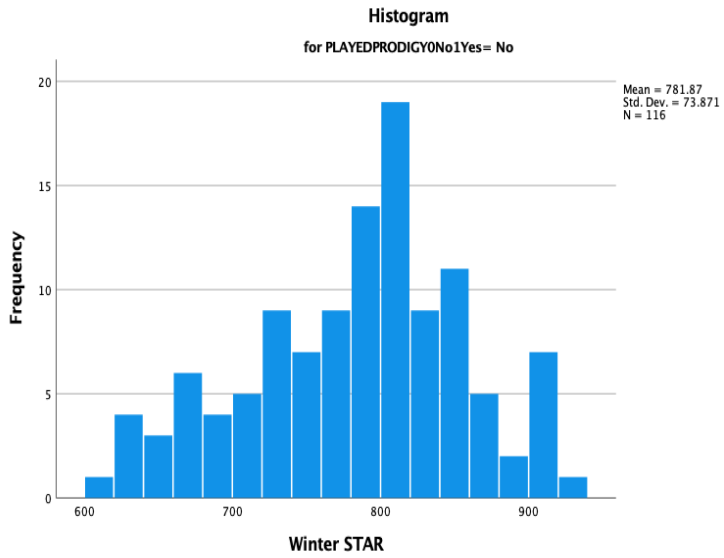

Figure 5. Histogram of the mean pretest Winter STAR scores for comparison group.

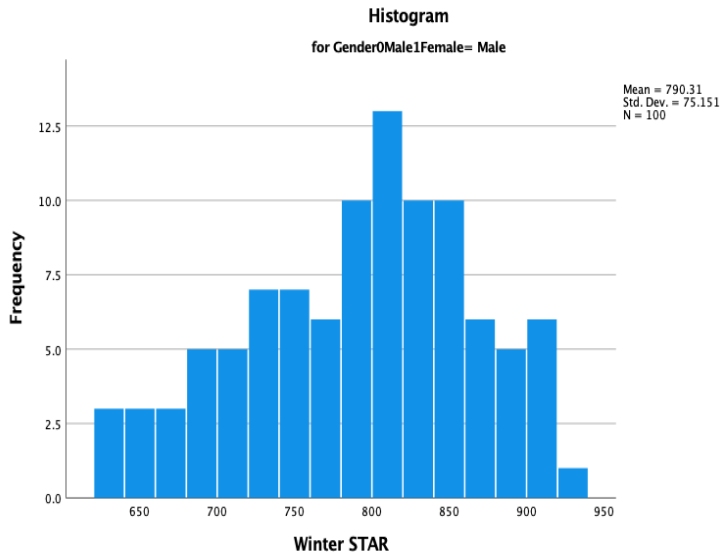

Figure 6. Histogram of the mean pretest Winter STAR scores for males.

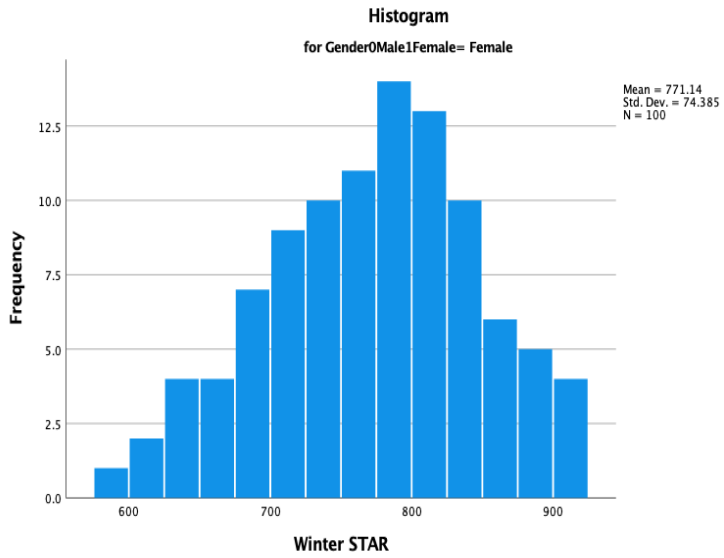

Figure 7. Histogram of the mean pretest Winter STAR scores for females.

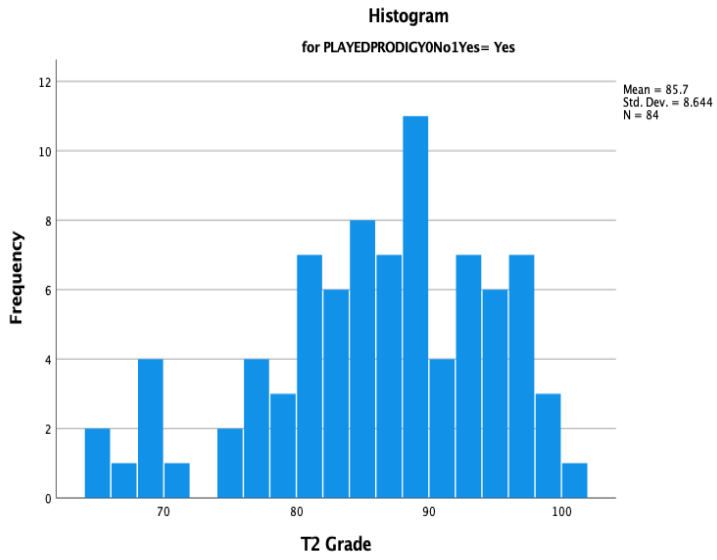

Figure 8. Histogram of the mean pretest T2 Grade for treatment group.

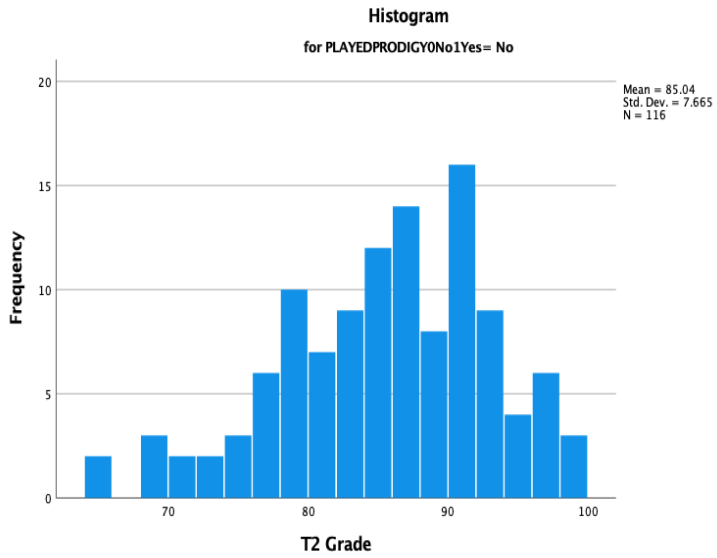

Figure 9. Histogram of the mean pretest T2 Grades for comparison group.

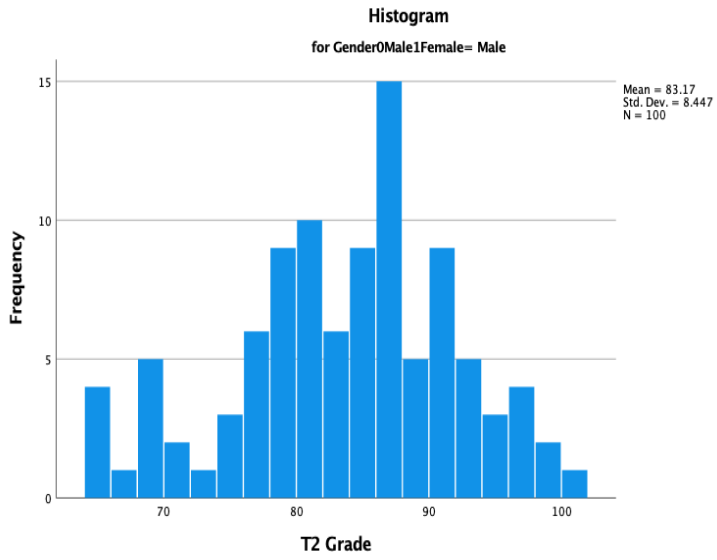

Figure 10. Histogram of the mean pretest T2 Grades for males.

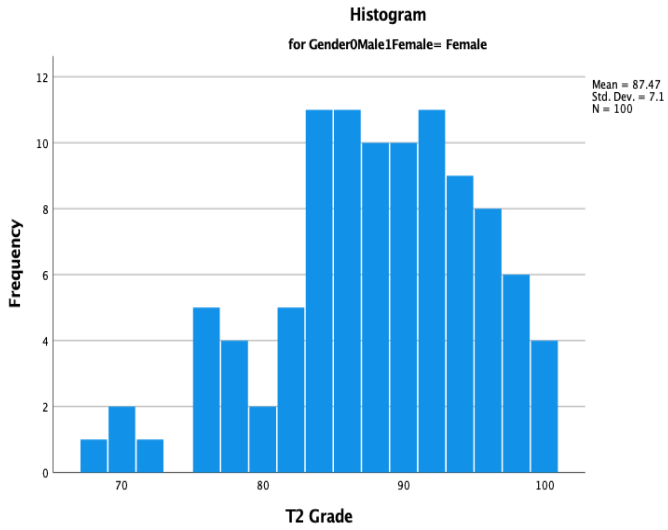

Figure 11. Histogram of the mean pretest T2 Grade for females.

Revised Linearity. To check for linearity the researcher calculated the Pearson correlation coefficient to verify the correlations (Meyers et al., 2006). These correlations were calculated by splitting the sample to compare correlations coefficients for two groups of participants (Pallant, 2016). The correlations are shown in Tables 33 and 34.

Table 33

Revised Correlations by Pretest Data (n = 200)

Gender Identity		T2 Grades
Male (*n* = 100)	Winter STAR	.647***
Female (*n* = 100)	Winter STAR	.493***

Note. ***$p < .001$; Bonferroni adjustment, $p < 3/.05 = .017$

Table 34

Revised Correlations by Program Type Pretest Data (n = 200)

Program Type		T2 Grades
Comparison (*n* = 116)	Winter STAR	.554***
Treatment (*n* = 84)	Winter STAR	.474***

Note. ***$p < .001$; Bonferroni adjustment, $p < 3/.05 = .017$

Calculating all the correlations without splitting the groups produced the correlations below in Table 35.

Table 35

Summary of Revised Correlations for all Pretest Data (n = 200)

Variables	Gender	Winter STAR	T2 Grades
Program Type	-.101	-.018	.040
Gender		-.128	.267**
Winter STAR			.516**

Note: For program type, comparison = 0 and treatment = 1. For gender, males = 0 and females = 1. **$p < .001$

Additionally, scatterplots were created to visually inspect that the variables were related to each other in a linear manner by showing that a straight line is the best fit function. The scatterplots were generated to check linearity using SPSS v.27 shown in Figures 12 through 15.

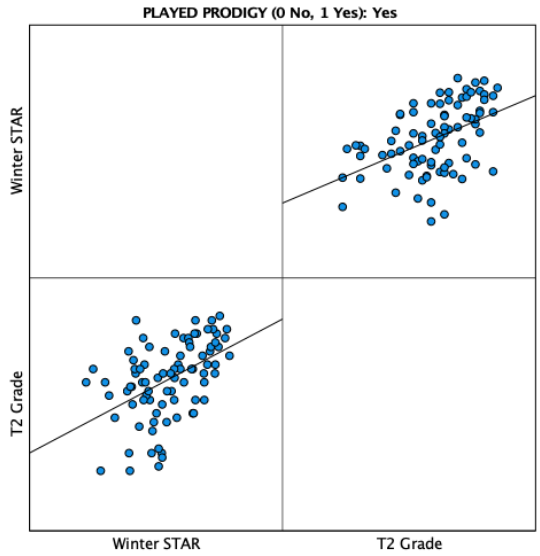

Figure 12. Scatterplot for pretest scores for treatment group.

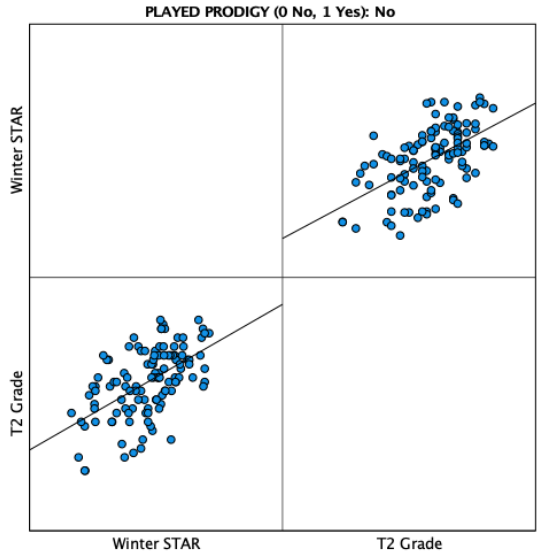

Figure 13. Scatterplot for pretest scores for comparison group.

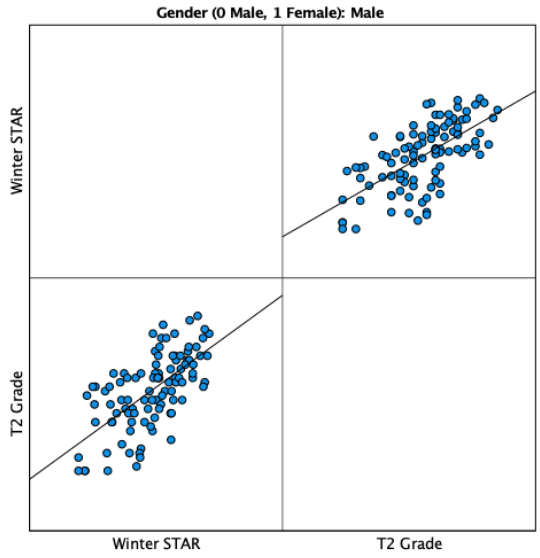

Figure 14. Scatterplot for pretest scores for males.

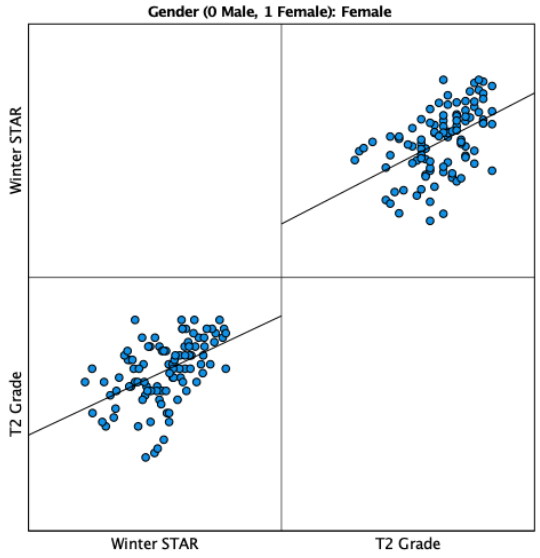

Figure 15. Scatterplot for pretest scores for females.

Revised Homoscedasticity. The homoscedasticity needed to be rechecked after 5 males were randomly eliminated from the study. To check the homoscedasticity assumption, the Box's Test of Equality of Covariance Matrices was generated. The value was not statistically significant (Box's M = 10.047, $p > 0.362$), as it was greater than .001, indicating that the

dependent variables' covariance matrices were equal across the levels of the independent variables (Meyers et al., 2006).

The Levene's test for Winter STAR was not significant ($p = 0.901$) but an examination of the T2 mathematics grades produced a significant ($p = 0.037$) result, however, since the number of males and females was equal the researcher proceeded with the analysis.

Results for Pre-Data MANOVA

A two-way between-groups multivariate analysis of variance (MANOVA) was performed to investigate achievement. A MANOVA was the appropriate test to use as per Hair et al. (2010): "a MANOVA is useful when the researcher designs an experimental situation to test hypotheses concerning the variance in group responses on two or more metric dependent variables" (p. 18). Two dependent variables were used: Winter STAR and T2 mathematics grades. The independent variables were program type (treatment or comparison) and gender (male or female). The assumptions of normality, linearity, univariate and multivariate outliers, homogeneity of variance-covariance matrices, and multicollinearity were checked with no serious violations noted (Pallant, 2016). To determine the significance of the MANOVA the Wilk's Lambda was used as it is "commonly used for testing overall significance between groups in a multivariate situation" (Hair et al., 2010, p. 369). There was no significant main effect for program type, $F(2, 195) = 1.317$, $p = 0.270$ Wilks' Lambda = 0.99; partial eta squared = 0.01. There was a statistically significant difference between males and females on the main effect for gender, $F(2, 195) = 21.457$, $p = 0.000$; Wilks' Lambda = 0.98; partial eta squared = 0.02. When the results of the dependent variables were considered separately, T2 mathematics grades were statistically significant, $F(3, 196) = 15.160$, $p = 0.000$, partial eta squares = 0.07, as were the Winter STAR scores for gender, $F(3, 196) = 4.168$, $p = 0.043$, partial eta squared = 0.02 for

gender. There was no significant interaction. Refer to Table 36 for the descriptive statistics, Table 37 for the multivariate tests, and Table 38 for pretest two-way MANOVA results.

Table 36

Descriptive Statistics for the Pretest Scores

	Group	Gender	Mean	Std. Deviation	N
Winter STAR	Comparison	Male	786.00	75.563	53
		Female	778.40	72.842	63
		Total	781.87	73.871	116
	Treatment	Male	795.17	75.198	47
		Female	758.78	76.349	37
		Total	779.14	77.411	84
	Total	Male	790.31	75.151	100
		Female	771.14	74.385	100
		Total	780.73	75.198	200
T2 Grade	Comparison	Male	82.58	7.350	53
		Female	87.11	7.360	63
		Total	85.04	7.665	116
	Treatment	Male	83.83	9.574	47
		Female	88.08	6.689	37
		Total	85.70	8.644	84
	Total	Male	83.17	8.447	100
		Female	87.47	7.100	100
		Total	85.32	8.076	200

Table 37

Multivariate Tests for the Pretest Scores

Effect		Value	*F*	Hypothesis *df*	Error *df*	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	0.993	13833.555	2.000	195.000	0.000	0.993
	Wilks' Lambda	0.007	13833.555	2.000	195.000	0.000	0.993
	Hotelling's Trace	141.883	13833.555	2.000	195.000	0.000	0.993
	Roy's Largest Root	141.883	13833.555	2.000	195.000	0.000	0.993
Program	Pillai's Trace	0.013	1.317	2.000	195.000	0.270	0.013
	Wilks' Lambda	0.987	1.317	2.000	195.000	0.270	0.013
	Hotelling's Trace	0.014	1.317	2.000	195.000	0.270	0.013
	Roy's Largest Root	0.014	1.317	2.000	195.000	0.270	0.013
Gender	Pillai's Trace	0.180	21.457	2.000	195.000	0.000	0.180
	Wilks' Lambda	0.820	21.457	2.000	195.000	0.000	0.180
	Hotelling's Trace	0.220	21.457	2.000	195.000	0.000	0.180
	Roy's Largest Root	0.220	21.457	2.000	195.000	0.000	0.180
Program* Gender	Pillai's Trace	0.012	1.209	2.000	195.000	0.301	0.012
	Wilks' Lambda	0.988	1.209	2.000	195.000	0.301	0.012
	Hotelling's Trace	0.012	1.209	2.000	195.000	0.301	0.012
	Roy's Largest Root	0.012	1.209	2.000	195.000	0.301	0.012

Note: $n = 200$

Table 38

Two-Way MANOVA Results for Mean Pretest Scores

Source	Dependent Variable	Type III Sum of Square	*df*	Mean Square	F	Sig.	Partial Eta Squared
Program	Winter STAR	1313.183	1	1313.183	0.235	0.628	0.001
	T2 Grade	59.072	1	59.072	0.965	0.327	0.005
Gender	Winter STAR	23301.769	1	23301.769	4.168	0.043	0.021
	T2 Grades	927.749	1	927.749	15.160	0.000	0.072
Program* Gender	Winter STAR	9976.279	1	9976.279	1.784	0.183	0.009
	T2 Grades	0.910	1	0.910	0.015	0.903	0.000

Note: $n = 200$

Mean pretest scores for T2 mathematics grades and Winter STAR scores were statistically significant. On the Winter STAR for males ($M = 790.31$) were significantly higher than those for females ($M = 771.14$), while for T2 grades the mean for females ($M = 87.47$) was statistically significantly higher than that for males ($M = 83.17$). Based on these results both, T2 grades and Winter STAR scores, served as covariates to adjust for these initial differences from the pretest analysis. As Hair et al. (2010) stated, "covariates are effective ways of controlling for external influences on the dependent variables" (p. 364). When using covariates, "the researcher can interpret the actual effect of the covariates on the dependent variate and their impact on the actual statistical test of the treatments" (p. 374).

Testing the Assumptions for Post-Data

The researcher tested all assumptions for the posttest data before conducting the MANOVA procedures. Meyers et al. (2006) stated the statistical assumptions and limitations for a MANOVA are "normality, linearity, and homoscedasticity" (p. 67).

Normality. The assumption of normality was addressed with an analysis of the skewness and kurtosis values, along with a visual inspection of the histograms. The values of skewness and kurtosis are in Tables 39 and 40 and were within the acceptable limits of ±1.00 (Hair et al., 2010).

Table 39

Skewness and Kurtosis Values by Program Type for Posttest Scores

Variable	Skewness Comparison (n = 116)	Skewness Treatment (n = 84)	Kurtosis Comparison (n = 116)	Kurtosis Treatment (n = 84)
Spring STAR	-0.546	-0.353	-0.210	-0.545
T3 Grades	-0.263	-0.403	-0.619	-0.261

Table 40

Skewness and Kurtosis Values by Gender for Posttest Scores

Variable	Skewness Male (n = 100)	Skewness Female (n = 100)	Kurtosis Male (n = 100)	Kurtosis Female (n = 100)
Spring STAR	-0.534	-0.456	-0.283	-0.407
T3 Grades	-0.147	-0.477	-0.475	-0.368

The test of normality was also reviewed by inspecting the results of the Kolmogorov-Smirnov and Shapiro-Wilk procedures. The criterion for this test is that the significance level should be higher than .001 (Meyers et al., 2006). Those values can be seen in Tables 41 and 42 for program type and gender, respectively.

Table 41

Tests of Normality Posttest Scores by Program Type (n = 200)

		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	*df*	Sig.	Statistic	*df*	Sig.
Spring STAR	Comparison	0.092	116	0.018	0.972	116	0.015
	Treatment	0.106	84	0.021	0.975	84	0.101
T3 Grade	Comparison	0.070	116	0.200	0.980	116	0.075
	Treatment	0.067	84	0.200	0.977	84	0.142

Table 42

Tests of Normality Posttest Scores by Gender (n = 200)

		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	*df*	Sig.	Statistic	*df*	Sig.
Spring STAR	Male	0.079	100	0.128	0.967	100	0.013
	Female	0.093	100	0.032	0.971	100	0.025
T3 Grade	Male	0.052	100	0.200	0.986	100	0.402
	Female	0.082	100	0.094	0.969	100	0.017

A visual inspection of normality was conducted using histograms. The histograms were generated using SPSS v.27 and are presented in Figures 16 through 23.

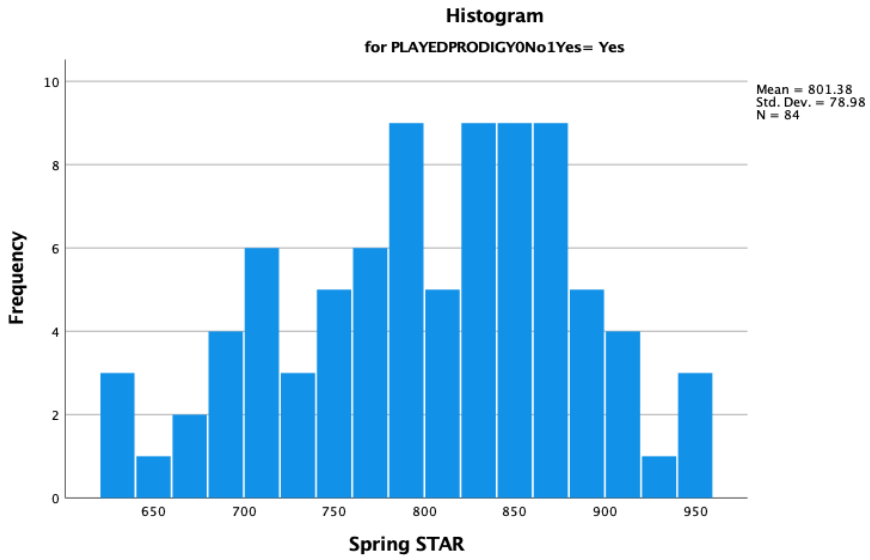

Figure 16. Histogram of the mean posttest Spring STAR scores for treatment group.

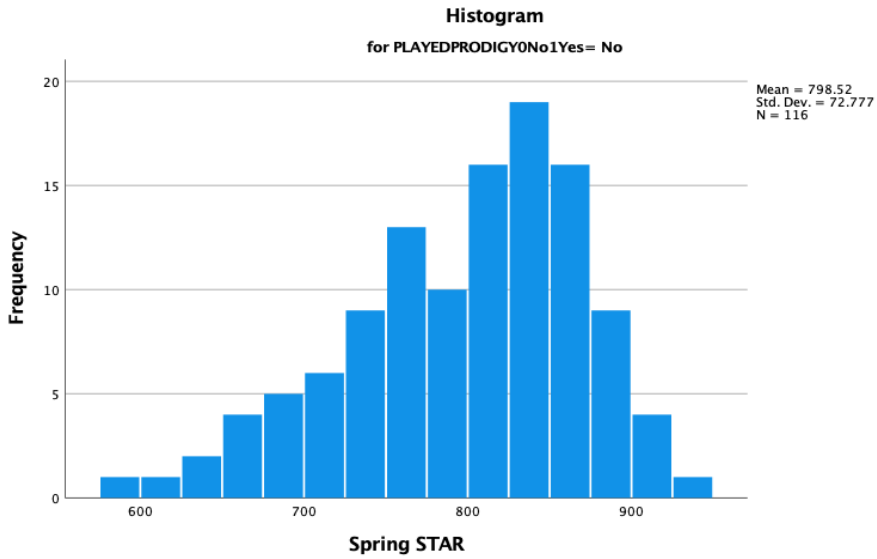

Figure 17. Histogram of the mean posttest Spring STAR scores for comparison group.

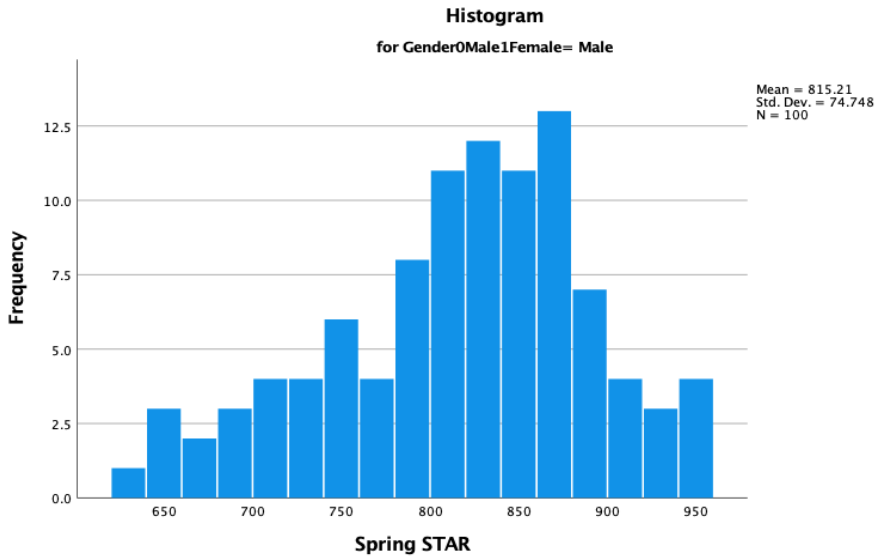

Figure 18. Histogram of the mean posttest Spring STAR scores for males.

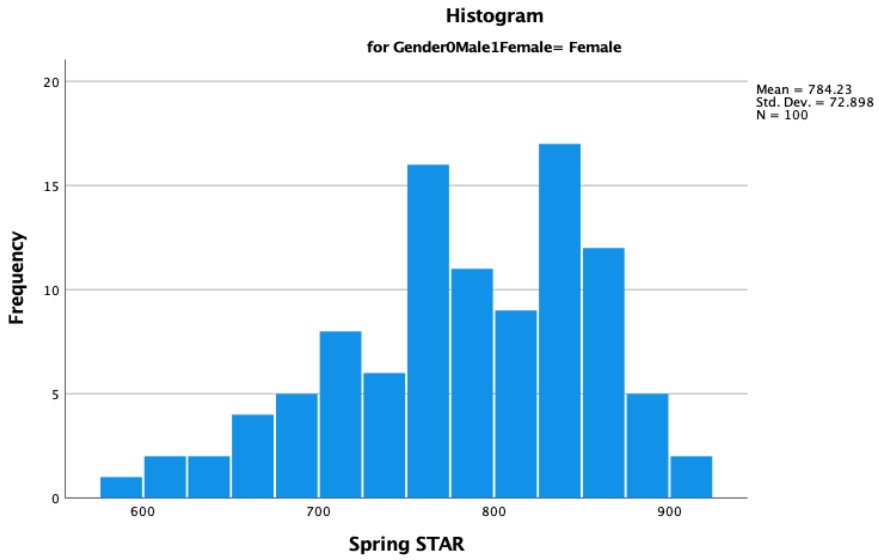

Figure 19. Histogram of the mean posttest Spring STAR scores for females.

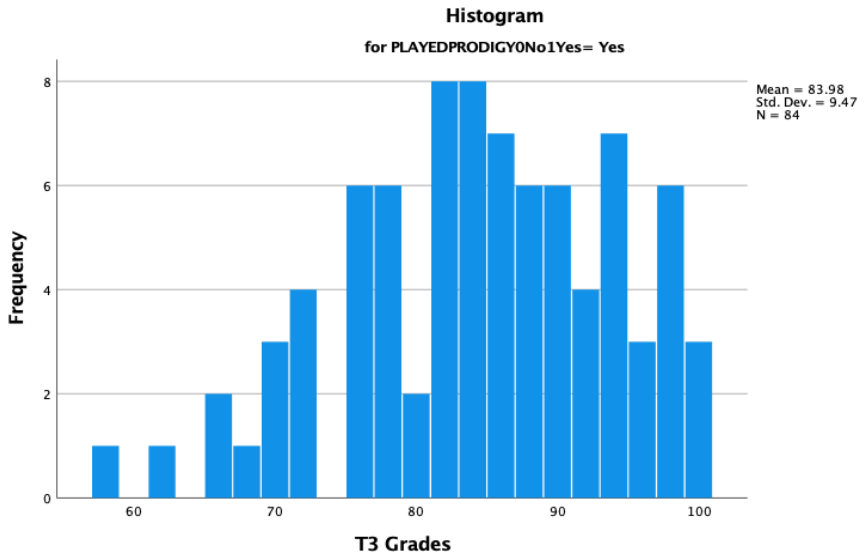

Figure 20. Histogram of the mean posttest T3 Grades for treatment group.

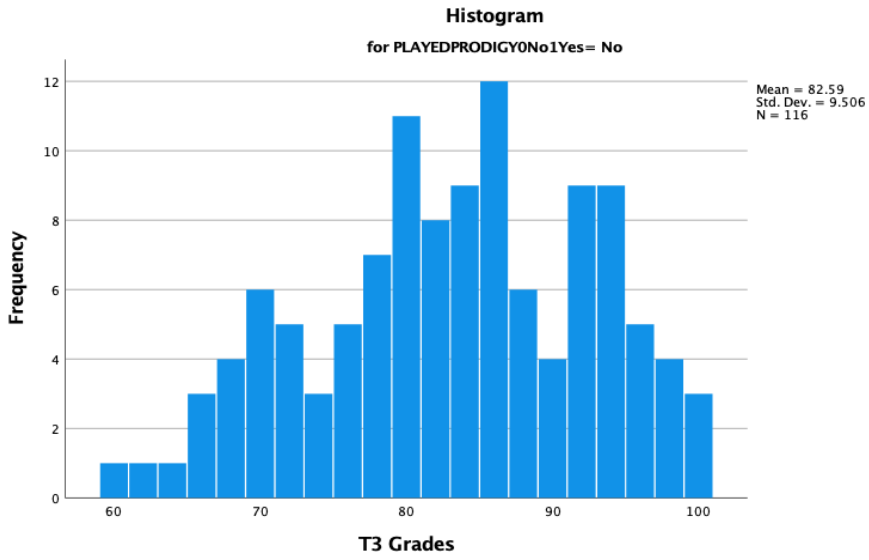

Figure 21. Histogram of the mean posttest T3 Grades for comparison group.

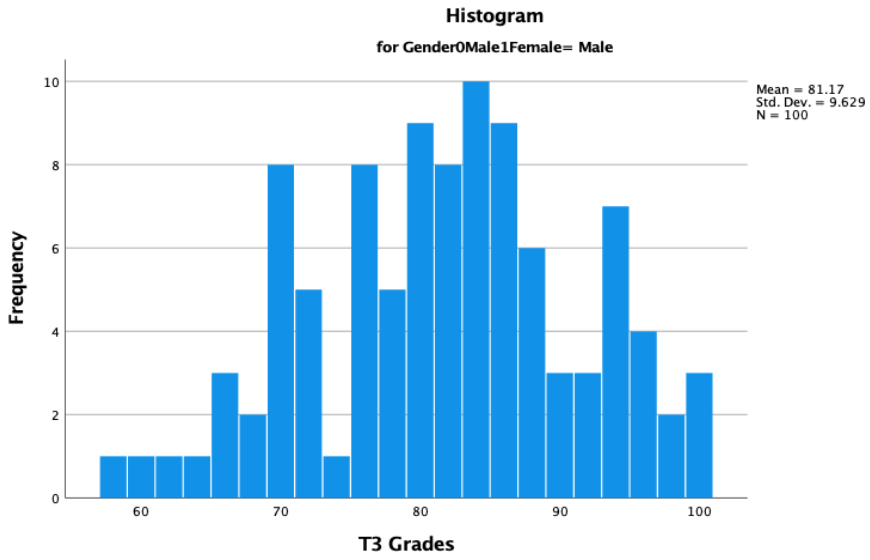

Figure 22. Histogram of the mean posttest T3 Grades for males.

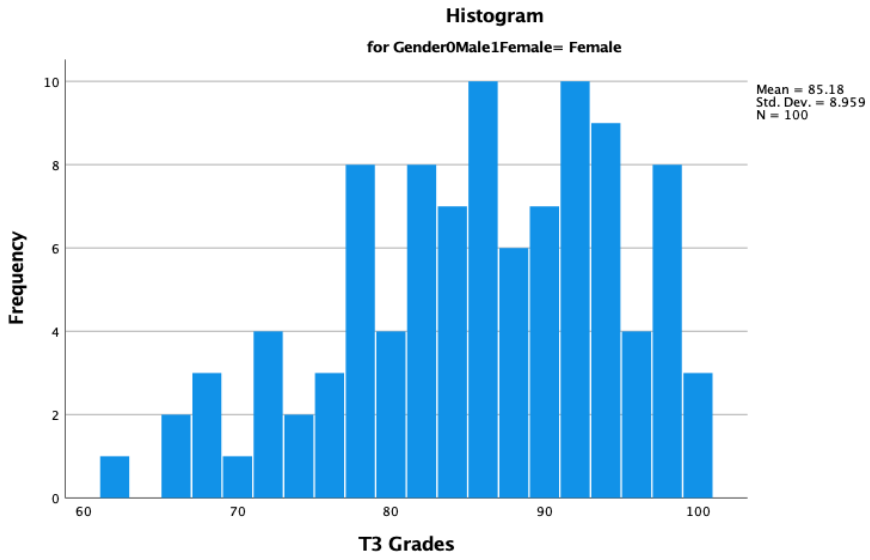

Figure 23. Histogram of the mean posttest T3 Grades for females.

Linearity. The researcher calculated a series of Pearson correlation coefficients to verify the degree of linear relationship between the variables (Meyers et al., 2006). Tables 43 and 44 show the correlations between Spring STAR and T3 mathematics grades. These correlations were calculated by splitting the sample to compare correlations coefficients for two groups of

participants (Pallant, 2016). The coefficients represented low to medium values, which were acceptable linear relationships.

Table 43

Correlations by Gender for Posttest (n = 200)

Gender Identity		T3 Grades
Male (*n* = 100)	Spring STAR	.523***
Female (*n* = 100)	Spring STAR	.471***

Note. ***$p < .001$; Bonferroni adjustment, $p < 3/.05 = .017$

Table 44

Correlations by Program Type for Posttest (n = 200)

Program Type		T3 Grades
Comparison (*n* = 116)	Spring STAR	.520***
Treatment (*n* = 84)	Spring STAR	.322***

Note. ***$p < .001$; Bonferroni adjustment, $p < 3/.05 = .017$

Calculating all the correlations without splitting the groups produced the correlations below in Table 45.

Table 45

Summary of Correlations for all Posttest Data (n = 200)

Variables	Gender	Spring STAR	T3 Grades
Program Type	-.101	.019	.072
Gender		-.206**	.212**
Spring STAR			.433**

Note: For program type, comparison = 0 and treatment = 1. For gender, males = 0 and females = 1. **$p < .001$

To check linearity, scatterplots were created to visually inspect that the variables were related to each other in a linear manner by showing that a straight line is the best fit function. The scatterplots to check linearity were generated using SPSS v.27 and are presented in Figures 24 through 27.

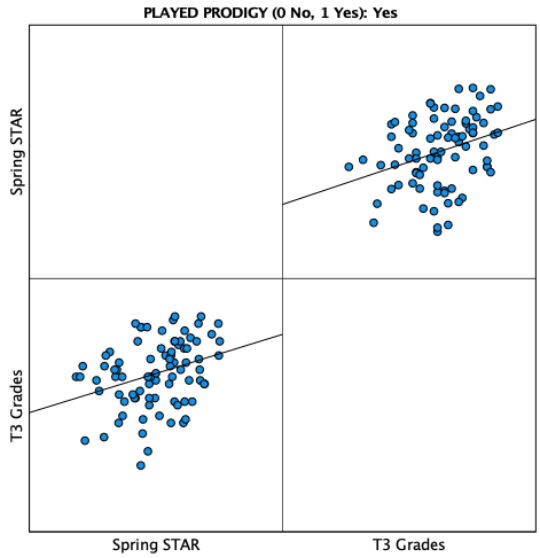

Figure 24. Scatterplot for posttest scores for treatment group.

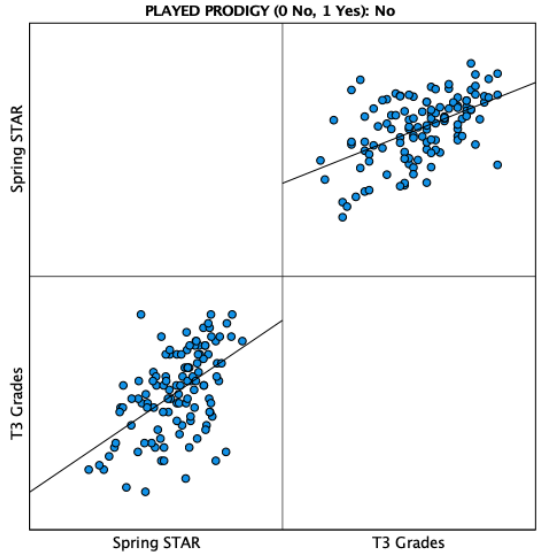

Figure 25. Scatterplot for posttest scores for comparison group.

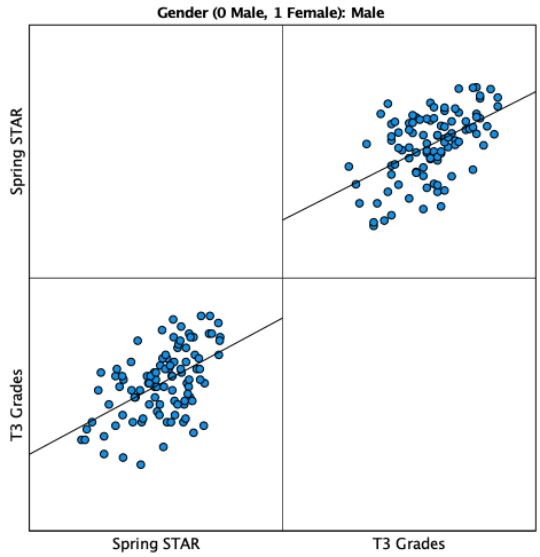

Figure 26. Scatterplot for posttest scores for males.

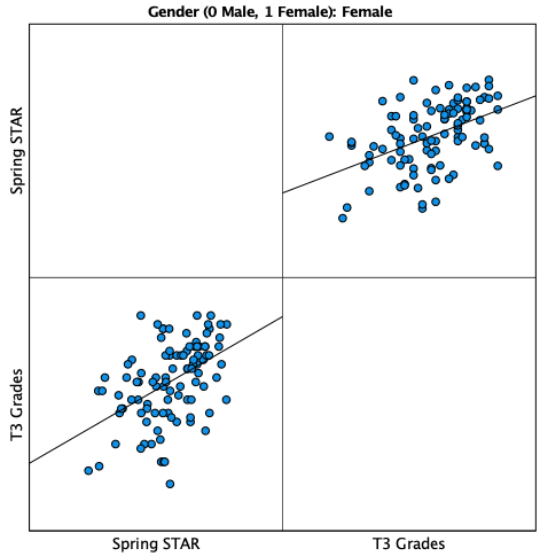

Figure 27. Scatterplot for posttest scores for females.

Homoscedasticity. To check the homoscedasticity assumption, both the Box's Test of Equality of Covariance Matrices and Levene's Test of Equality of Error Variances were used. The Box's M was not statistically significant (Box's M = 6.118, $p > 0.740$), indicating that the dependent variables' covariance matrices are equal across the levels of the independent

variables. Levene's showed no significant values for Spring STAR scores (p = -.099) or T3 mathematics grades (p = 0.771).

Results for Research Question One

A two-way between-groups multivariate analysis of variance (MANOVA) was performed to investigate achievement between groups. Two dependent variables were used: Spring STAR and T3 mathematics grades. The independent variables were as follows: (a) Program type, two levels: GBL enhanced mathematics program (treatment) and traditional mathematics program (comparison) and (b) gender, two levels: male and female. Preliminary assumption testing was conducted to check for normality, linearity, univariate and multivariate outliers, homogeneity of variance-covariance matrices, and multicollinearity, with no serious violations noted (Pallant, 2016). Since T2 math grades and Winter STAR for gender were statistically significant, they were used as covariates for the posttest analysis. The MANOVA indicated there was a statistically significant difference on gender between males and females, F (2, 193) = 10.726, p = 0.000; Wilks' Lambda = 0.900; partial eta squared = 0.10. Follow up analysis indicated that this difference related to the dependent variable, Spring STAR, F(1, 194) = 19.235, p = 0.000, partial eta squared = 0.090. An inspection of the mean scores indicated that males had significantly higher Spring STAR scores (M = 814.32, SD = 4.538) than females (M = 784.19, SD = 4.729). There was no significant difference between the two levels of program type, F (2, 193) = 0.556, p = 0.575 Wilks' Lambda = 0.994; partial eta squared = 0.006 and no significant interaction. Refer to Table 46 for the descriptive statistics for posttest scores including adjusted means are in Table 47 for the multivariate tests, Table 48 for posttest two-way MANOVA results, and Table 49 for the estimated marginal means for the covariates.

Table 46

Descriptive Statistics for the Posttest Scores

	Group	Gender	Mean	Std. Deviation	Adjusted Means	N
Spring STAR	Comparison	Male	806.94	77.534	809.872	53
		Female	791.43	68.340	788.795	63
		Total	798.52	72.777	799.334	116
	Treatment	Male	824.53	71.144	818.769	47
		Female	771.97	79.533	779.584	37
		Total	801.38	78.980	799.177	84
	Total	Male	815.21	74.748	814.320	100
		Female	784.23	72.898	784.189	100
		Total	799.72	75.263	799.255	200
T3 Grade	Comparison	Male	80.30	9.683	82.295	53
		Female	84.52	8.984	83.206	63
		Total	82.59	9.506	82.751	116
	Treatment	Male	82.15	9.576	83.110	47
		Female	86.30	8.928	84.464	37
		Total	83.98	9.470	83.787	84
	Total	Male	81.17	9.629	82.703	100
		Female	85.18	8.959	83.835	100
		Total	83.18	9.492	83.269	200

Table 47

Multivariate Tests for the Posttest Scores

Effect		Value	F	Hypo-thesis *df*	Error *df*	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	0.993	13833.555	2.000	195.000	0.000	0.993
	Wilks' Lambda	0.007	13833.555	2.000	195.000	0.000	0.993
	Hotelling's Trace	141.883	13833.555	2.000	195.000	0.000	0.993
	Roy's Largest Root	141.883	13833.555	2.000	195.000	0.000	0.993
Program	Pillai's Trace	0.006	.556	2.000	193.000	0.575	0.006
	Wilks' Lambda	0.994	.556	2.000	193.000	0.575	0.006
	Hotelling's Trace	0.006	.556	2.000	193.000	0.575	0.006
	Roy's Largest Root	0.006	.556	2.000	193.000	0.575	0.006
Gender	Pillai's Trace	0.100	10.726	2.000	193.000	0.000	0.100
	Wilks' Lambda	0.900	10.726	2.000	193.000	0.000	0.100
	Hotelling's Trace	0.111	10.726	2.000	193.000	0.000	0.100
	Roy's Largest Root	0.111	10.726	2.000	193.000	0.000	0.100
Program* Gender	Pillai's Trace	0.011	1.114	2.000	193.000	0.330	0.011
	Wilks' Lambda	0.989	1.114	2.000	193.000	0.330	0.011
	Hotelling's Trace	0.012	1.114	2.000	193.000	0.330	0.011
	Roy's Largest Root	0.012	1.114	2.000	193.000	0.330	0.011

Table 48

Follow-up Results for Mean Posttest Scores

Source	Dependent Variable	Type III Sum of Square	*df*	Mean Square	*F*	Sig.	Partial Eta Squared
Program	Spring STAR	1.171	1	1.171	.001	0.980	0.000
	T3 Grade	51.052	1	51.052	1.097	0.296	0.006
Gender	Spring STAR	35841.878	1	35841.878	19.235	0.000	0.090
	T3 Grade	50.632	1	50.632	1.088	0.298	0.006
Program* Gender	Spring STAR	3899.753	1	3899.753	2.093	0.150	0.011
	T3 Grade	2.330	1	2.330	0.050	0.823	0.000

Table 49

Estimated Marginal Means for Covariate for Posttest Scores

Dependent Variable	Program Type	Gender	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Spring STAR	Comparison	Male	809.872	6.115	797.811	821.933
		Female	788.795	5.519	777.909	799.680
	Treatment	Male	818.769	6.428	806.091	831.446
		Female	779.584	7.424	764.942	794.226
	Total		799.255	3.111	793.119	805.391
T3 Grade	Comparison	Male	82.295	.966	80.390	84.201
		Female	83.206	.872	81.486	84.926
	Treatment	Male	83.110	1.016	81.107	85.113
		Female	84.464	1.173	82.151	86.778
	Total		83.269[a]	0.492	84.238	83.269

The original hypothesis stated that there would be a statistically significant difference in posttest between student achievement and program types. This hypothesis was partially met

since males scores higher than females on the STAR test. However, there was no statistically significance due to program type.

Research Question Two

Research question two was used to explore to what degree and in what manner gender (male, female), mathematics grade, and amount of time spent using the virtual mathematics game would predict student achievement on the Spring STAR Assessment. A multiple linear regression procedure was conducted to analyze research question two. The dependent variable was Spring STAR score, and the independent variables were gender, two levels (male, female), T3 mathematics grades, and hours spent using a virtual mathematics game.

Descriptive Statistics of Data

Using the cleaned data set described earlier, with a population of 87 students in the treatment group and 118 in the comparison group, the researcher investigated additional outliers from the independent variable of time spent using Prodigy. While analyzing box plots, regions beyond the inner fences are extreme scores and a data point that is farther than ±1.5 IQRs (inter quartile range) but less than ±3.0 IQRs is labeled as an outlier by SPSS (Meyers et al., 2006). Once the outliers were visually detected in this way, the researcher also used the rule by Hair et al. (2010) that outliers are defined as cases with scores of ±2.5 standard deviations from the mean.

There was a total of three rounds to identify and eliminate outliers in the data set. A total of 4 outliers were eliminated in this way. Table 50 shows the students who were identified as outliers from the box and whisker plots. If the student values for any indicator were 2.5 standard deviations below or above the mean they were eliminated from the analysis.

Table 50

Investigation of Potential Outliers for Research Question Two

Case	IV	DV	Case Mean	Group Mean	*SD*	Criteria 2.5 SD Away Low	Criteria 2.5 *SD* Below Mean	Criteria 2.5 *SD* Above Mean
				Round 1				
94	Male	Time	20.11	6.82	4.43	-4.24	17.89	Yes
113	Female	Time	14.98	5.03	4.01	-4.99	15.06	No
288	Female	Time	14.40	5.03	4.01	-4.99	15.06	No
114	Female	Time	15.00	5.03	4.01	-4.99	15.06	No
118	Female	Time	14.25	5.03	4.01	-4.99	15.06	No
				Round 2				
251	Male	Time	17.57	6.55	4.03	-3.52	16.63	Yes
113	Female	Time	14.98	5.03	4.01	-4.99	15.06	No
288	Female	Time	14.40	5.03	4.01	-4.99	15.06	No
114	Female	Time	15.00	5.03	4.01	-4.99	15.06	No
118	Female	Time	14.25	5.03	4.01	-4.99	15.06	No
378	Female	T3 Grades	62.00	86.30	8.93	63.98	108.62	Yes
				Round 3				
316	Male	Time	16.72	6.32	3.74	-3.02	15.66	Yes
113	Female	Time	14.98	5.07	4.06	-5.09	15.22	No
288	Female	Time	14.40	5.07	4.06	-5.09	15.22	No
114	Female	Time	15.00	5.07	4.06	-5.09	15.22	No
118	Female	Time	14.25	5.07	4.06	-5.09	15.22	No

After Round 3 of checking the data, the researcher inspected the box and whisker plot for Total Time with regards to gender and the same female outliers from round 3 were still visible above the upper fence, see Figure 28. While not eliminated in Round 3 the values were so close to the criterion that they showed up as outliers and were therefore eliminated. Therefore, a total of eight outliers were eliminated, for a total population for the analysis of 79, with 47 males and 32 females.

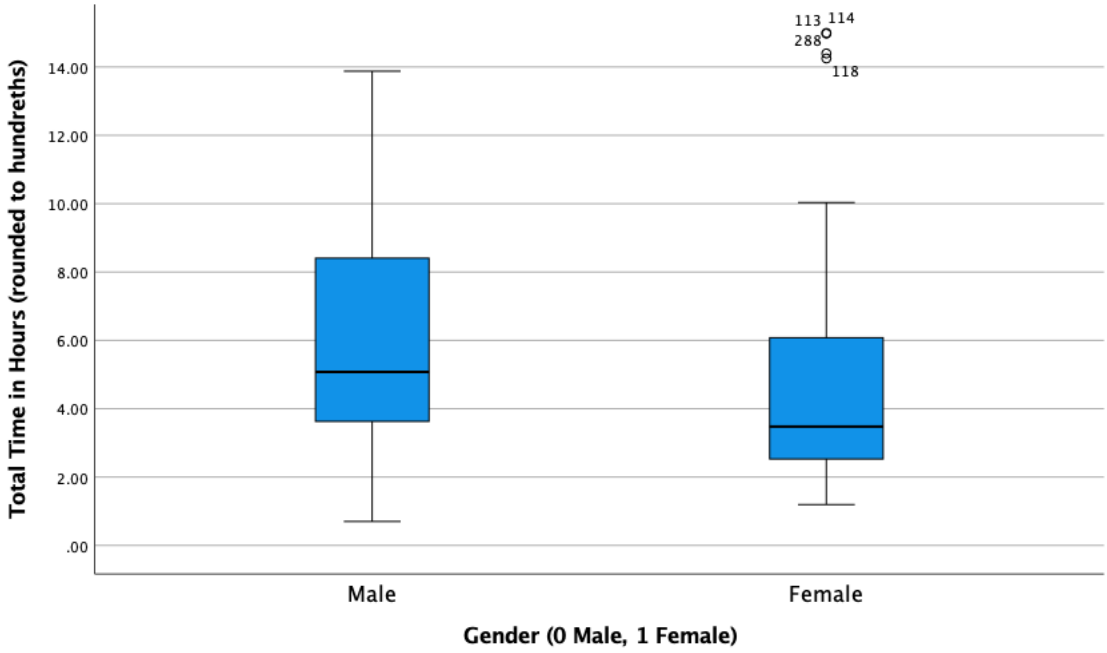

Figure 28. Box and Whisker Plot After Three Rounds of Cleaning Data

Table 51 presents the descriptive statistics, the frequency, mean and standard deviations for the variables.

Table 51

Descriptive Statistics for Posttest of Independent Variables with Respect to Each Dependent Variable for Multiple Regression

Gender			Statistic	Std. Error
Total Time in Hours	Male	Mean	6.103	0.502
		Std. Deviation	3.444	
		Minimum	0.702	
		Maximum	13.881	
		n = 47		
	Female	Mean	3.869	0.406
		Std. Deviation	2.296	
		Minimum	1.196	
		Maximum	10.034	
		n = 32		
T3 Grades	Male	Mean	81.380	1.409
		Std. Deviation	9.663	
		Minimum	58.000	
		Maximum	100.000	
		n = 47		
	Female	Mean	87.88	1.328
		Std. Deviation	7.512	
		Minimum	70.000	
		Maximum	100.000	
		n = 32		
Spring STAR	Male	Mean	815.040	11.325
		Std. Deviation	77.642	
		Minimum	613.000	
		Maximum	944.000	
		n = 47		
	Female	Mean	777.34	13.922
		Std. Deviation	78.757	
		Minimum	620.000	
		Maximum	909.000	
		n = 32		

Testing the Assumptions

The assumptions for a multiple regression include an examination of the multivariate outliers, normality, linearity and homoscedasticity (Meyers et al., 2006).

Normality. The researcher checked the skewness and kurtosis and found every measure except females and time were within the acceptable range, with the skewness and kurtosis values within an acceptable range of ±1 as suggested by Meyers et al. (2006). The female and time values showed a slight skewness of the data, however, they were within ±2.0, which George and Mallery (2003) state is also acceptable. The values are shown in Table 52.

Table 52

Skewness and Kurtosis for Achievement Regression Dependent Variables

			Statistic	Std. Error	*n*
Total Time in Hours	Male	Skewness	0.688	0.347	47
		Kurtosis	-0.461	0.681	47
	Female	Skewness	1.288	0.414	32
		Kurtosis	1.084	0.809	32
T3 Grades	Male	Skewness	-0.083	0.347	47
		Kurtosis	-0.351	0.681	47
	Female	Skewness	-0.368	0.414	32
		Kurtosis	-0.461	0.809	32
Spring STAR	Male	Skewness	-0.536	0.347	47
		Kurtosis	-0.039	0.681	47
	Female	Skewness	-0.175	0.414	32
		Kurtosis	-0.894	0.809	32

The test of normality was also reviewed using Kolmogorov-Smirnov and Shapiro-Wilk. The criterion for this test is that the significance level should be higher than .001 (Meyers et al., 2006). Those values can be seen in Table 53 for gender.

Table 53

Tests of Normality for Achievement Regression Dependent Variables

	Gender	Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	*df*	Sig.	Statistic	*df*	Sig.
Total Time in Hours	Male	0.137	47.000	0.027	0.932	47	0.009
	Female	0.199	32.000	0.002	0.862	32	.001
T3 Grades	Male	0.059	47.000	0.200	0.987	47	0.88
	Female	0.079	32.000	0.200	0.970	32	0.488
Spring STAR	Male	0.097	47.000	0.200	0.971	47	0.290
	Female	0.123	32.000	0.200	0.969	32	0.484

Multivariate Outliers. The researcher first checked for multivariate outliers with the Mahalanobis distance. This distance is measured using a chi-square criterion, in this case with three degrees of freedom, the number of independent/predictor variables, at alpha = $p < .001$, resulted in a chi-square value of 16.266 (Meyers et al., 2006). No values exceeded these chi-squared criterion, therefore, this assumption was met. The results from the Mahalanobis distance can be found in Table 54.

Table 54

Mahalanobis Distance for Regression Assumptions

Mahalanobis Distance	Extreme Values	Case Number	Value
Highest	1	361	8.99181
	2	211	8.08558
	3	354	8.08337
	4	49	7.98141
	5	47	7.32714
Lowest	1	63	0.75570
	2	179	0.75865
	3	108	0.76832
	4	128	0.94740
	5	129	1.20709

Another check for outliers is found through examination of residuals in a scatterplot. The case wise diagnostic produced no outliers that exceeded three standard deviations units (Meyers et al., 2006). In addition, the scatterplot showed a rectangularity with the data that indicated that the residuals were normally distributed (Meyers et al., 2006). This scatterplot can be seen in Figure 29.

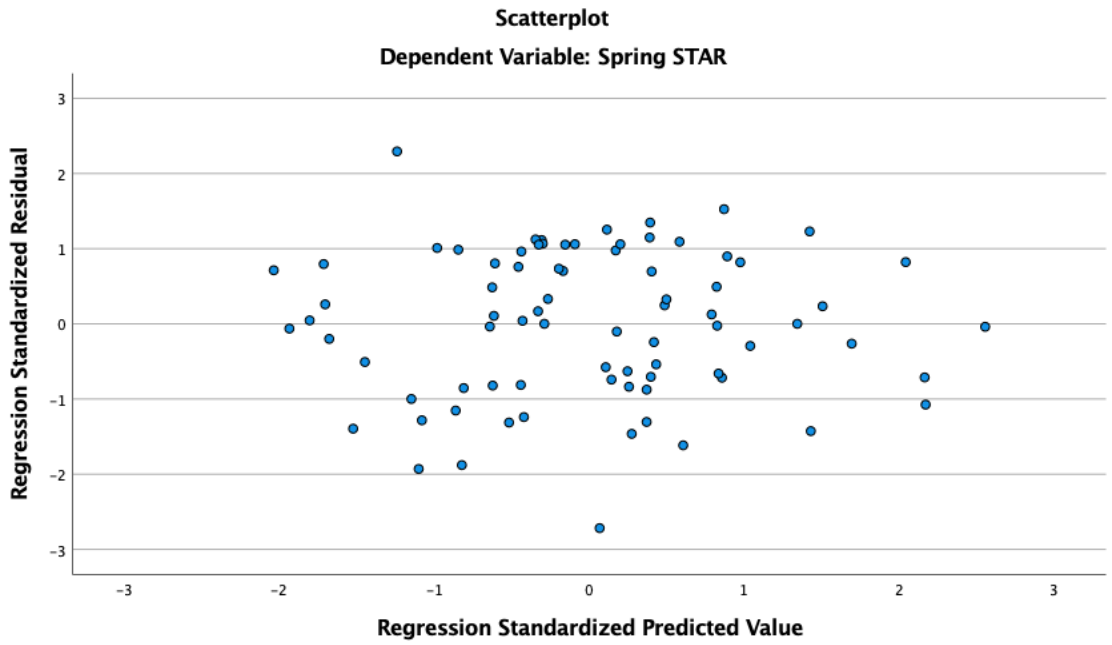

Figure 29. Scatterplot for multivariate outliers.

In addition, the scatterplot of the residuals shows points that were equally distributed and therefore the data were homoscedastic and passed the assumption of homoscedasticity.

Linearity. The researcher calculated the Pearson correlation coefficient to verify the degree of linear relationship between the variables (Meyers et al., 2006). The correlations can be seen in Table 55, and the scatterplot matrix between each pair of variables can be seen in Figure 30, all scatterplots were made using SPSS v.27.

Table 55

Simple Correlations Among All Variables in the Regression Procedure

Variables	Gender	Total Time	T3 Grades
Spring STAR	-.233*	0.170	.304**
T3 Grades	.324*	-0.135	
Total Time	-.344**		

Note: For gender, males = 0 and females = 1. *Correlations significant at the 0.05 level, **Correlation is significant at the 0.01 level

To check linearity the researcher examined the scatterplots of the variables to identify any nonlinear patterns in the data (Hair et al., 2010). Although some plots showed a weak correlation, none of the plots showed non-linearity or zero correlations, therefore, the assumption of linearity was satisfied (George & Mallery, 2003). In addition, because the residuals were normally distributed and homoscedastic, linearity was also met (Meyers et al., 2006).

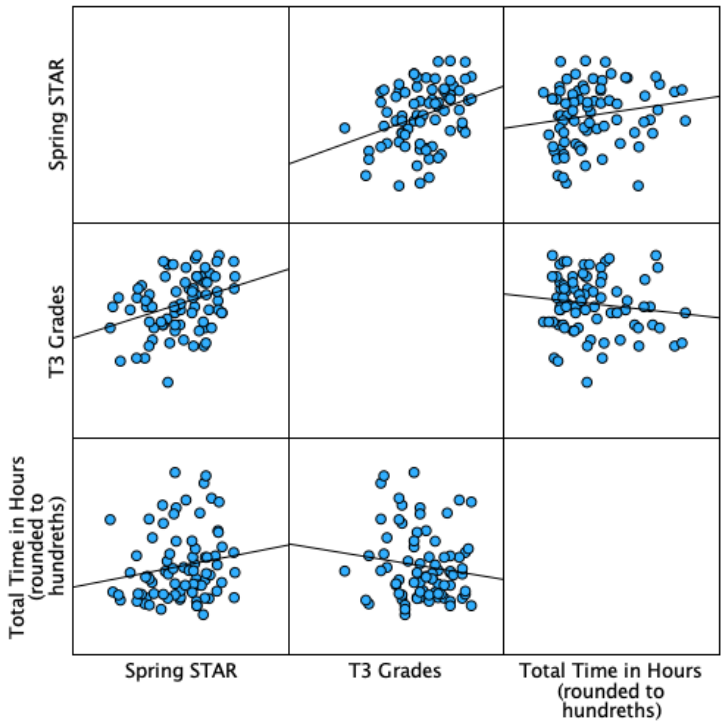

Figure 30. Scatterplot for variables in multiple regression for gender.

Results for Research Question Two

A stepwise multiple regression was used to identify possible predictors of the outcome Spring STAR using the following variables: Gender, Time Spent and T3 mathematics grades. At each step, variables were added based on *p*-values.

At step one of the analyses, T3 mathematics grades entered into the regression equation. The multiple correlation coefficient was $R = 0.340$. Consequently $R^2 = 0.115$ indicating 11.5% of the variance of the Spring STAR could be accounted for by T3 mathematics grades. At step two

of the analysis, gender entered into the regression equation. The multiple correlation coefficient was 0.504[b] indicating 25.4% of the variance of the Spring STAR could be accounted for by T3 mathematics grades and gender, with 74.6% of the variance not accounted for. Time spent using Prodigy was not a significant predictor of Spring STAR scores. Students with higher mathematics grades had higher Spring STAR scores and overall females scored higher in achievement. Tables 56 and 57 show the model summary and ANOVA results, respectively.

Table 56

Model Summary for Achievement Linear Regression

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Sig. F. Change	Durbin-Watson
1	.340[a]	0.115	0.104	75.534	.002	
2	.504[b]	0.254	0.234	69.823	<.001	2.115

a Predictors: (Constant), T3 Grades
b Predictors: (Constant), T3 Grades, Gender (0 Male, 1 Female)
c Dependent Variable: Spring STAR

Table 57

ANOVA Results for Achievement Linear Regression

Model		Sum of Squares	*df*	Mean Square	*F*	Sig.
1	Regression	57322.254	1.000	57322.254	10.047	.002
	Residual	439315.645	77.000	5705.398		
	Total	496637.899	78.000			
2	Regression	126122.059	2.000	63061.029	12.935	.000[c]
	Residual	370515.840	76.000	4875.208		
	Total	496637.899	78.000			

a Dependent Variable: Spring STAR
b Predictors: (Constant), T3 Grades
c Predictors: (Constant), T3 Grades, Gender (0 Male, 1 Female)

The original hypothesis stated that gender, mathematics grade and amount of time spent using the virtual game will predict student achievement on STAR assessment. This hypothesis was partially met since two out of the three predictors, T3 grades and gender, significantly predicted Spring STAR scores. However, time spent on the game did not enter regression as a predictor.

Research Question Three

Research question three was used to explore whether or not there was a significant difference in mathematics class engagement between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female). To answer this question a two-way multivariate analysis of variance (MANOVA) was used. The dependent variables consisted of three posttest scores on Student Engagement in Mathematics Scale (SEMS): Cognitive, Social, and Emotional. The independent variables were, (a) program type, two levels: a GBL enhanced mathematics program (treatment), a traditional mathematics program (comparison) and (b) gender, two levels: male and female.

Data Cleaning

Before the researcher conducted the two-way MANOVA, data cleaning needed to be conducted based on missing data. The method that was used to handle missing data was listwise, which involved deleting all of the cases that had missing data (Meyers et al., 2006). As a result of cleaning the data, 108 cases were deleted from a total of 230. After this was completed, the researcher analyzed the outliers. The new data set started with a sample size of 122 students. The researcher identified outliers by viewing the box and whisker plots and then used the rule by Hair et al. (2010) that outliers are defined as cases with scores of $\pm$ 2.5 standard deviations away from the mean.

There was only one round needed to identify and eliminate outliers in the data set. Table 58 shows data per case identified as outliers from the box and whisker plots. If a student mean value was ± 2.5 standard deviations from the mean, they were eliminated from the analysis.

Table 58

Analyzing Outliers for Engagement Analysis

						Criteria		
Case	IV	DV	Case Mean	Group Mean	*SD*	2.5 *SD* Below Mean	2.5 *SD* Above Mean	Eliminated
325	Male	Emotional	1	3.022	0.690	1	5	Yes
293	No	Social	1	2.634	0.775	1	5	Yes
223	No	Social	1	2.634	0.775	1	5	Yes
325	Yes	Emotional	1	2.985	0.743	1	5	Yes
246	Yes	Emotional	1	2.985	0.743	1	5	Yes
43	Yes	Emotional	1	2.985	0.743	1	5	Yes

There was a total of five outliers identified. Eliminating these outliers created a clean data set of 117 cases. The frequency and percent of cases for gender can be found in Table 59.

Table 59

Gender of Student Participants in Treatment and Comparison Groups after Data Cleaning for SEMS

Group	Comparison		Treatment		Total	
	n	Percent	*n*	Percent	*n*	Percent
Male	23	44.2	39	60	62	53.0
Female	29	55.8	26	40	55	47.0
Total	52	100.0	65	100	117	100.0

Descriptive Statistics

Once the data were cleaned, the researcher used a two-way MANOVA using program type (treatment or comparison) and gender (male or female) as the independent variables and the three subscales mean posttest scores for the Student Engagement in Mathematics Scale (SEMS); Cognitive, Social, and Emotional. Only posttest data were analyzed. Descriptive statistics are presented in Table 60.

Table 60

Descriptive Statistics for Posttest of Independent Variables with Respect to Each Dependent Variable (cognitive, social and emotional) for SEMS

Independent Variable	Dependent Variable	Mean	*SD*
Program Type - Comparison (*n* = 52)	Cognitive	3.385	0.599
	Social	2.697	0.718
	Emotional	2.938	0.709
Program Type - Treatment (*n* = 65)	Cognitive	3.492	0.547
	Social	2.685	0.912
	Emotional	3.077	0.620
Gender – Male (*n* = 62)	Cognitive	3.411	0.585
	Social	2.698	0.851
	Emotional	3.071	0.618
Gender – Female (*n* = 55)	Cognitive	3.482	0.557
	Social	2.682	0.810
	Emotional	2.953	0.709

Testing the Assumptions

The statistical assumptions that need to be checked for a MANOVA are normality, linearity, and homoscedasticity (Meyers et al., 2006).

Normality. The researcher first did a visual inspection of the histograms to check the normality (Hair et al., 2010). To further check that normality was within an acceptable range, the researcher examined descriptive statistics and determined the skewness and kurtosis values were within an acceptable range of ± 2.0 which George and Mallery (2003) state an acceptable criterion in Tables 61 and 62.

Table 61

Skewness and Kurtosis Values by Program Type for Posttest Scores for Engagement

Variable	Skewness Comparison (n = 52)	Skewness Treatment (n = 65)	Kurtosis Comparison (n = 52)	Kurtosis Treatment (n = 65)
Cognitive	-0.689	-0.635	-0.502	-1.018
Social	-0.280	-0.189	-0.237	-0.834
Emotional	-0.186	-0.295	-0.740	-0.704

Table 62

Skewness and Kurtosis Values by Gender for Posttest Scores for Engagement

Variable	Skewness Male (n = 62)	Skewness Female (n = 55)	Kurtosis Male (n = 62)	Kurtosis Female (n = 55)
Cognitive	-0.435	-0.992	-1.089	.037
Social	-0.114	-0.373	-0.807	-0.251
Emotional	-0.149	-0.308	-0.947	-0.658

The test of normality was also reviewed using the Kolmogorov-Smirnov and Shapiro-Wilk procedures. Those values can be seen in Table 63 and 64 for program type and gender, respectively. Significant levels at $p < .001$ and above are acceptable (Meyers et al., 2006).

Table 63

Tests of Normality Posttest Engagement Scores by Program Type (n = 117)

		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	*df*	Sig.	Statistic	*df*	Sig.
Cognitive	Comparison	0.175	52	0.000	0.881	52	0.000
	Treatment	0.235	65	0.000	0.831	65	0.000
Social	Comparison	0.126	52	0.040	0.969	52	0.191
	Treatment	0.079	65	0.200	0.944	65	0.006
Emotional	Comparison	0.097	52	0.200	0.956	52	0.053
	Treatment	0.114	65	0.035	0.961	65	0.039

Table 64

Tests of Normality Posttest Engagement Scores by Gender (n = 117)

		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	*df*	Sig.	Statistic	*df*	Sig.
Cognitive	Male	0.246	62	0.000	0.854	62	0.000
	Female	0.230	55	0.000	0.845	55	0.000
Social	Male	0.091	62	0.200	0.957	62	0.030
	Female	0.138	55	0.010	0.951	55	0.025
Emotional	Male	0.111	62	0.058	0.953	62	0.018
	Female	0.124	55	0.035	0.954	55	0.033

Since the variables of cognitive engagement failed the assumption, because it was statistically significant, for both program type and gender with respect to normality it was eliminated from the analysis. Table 65 provides and updated descriptive statistic.

Table 65

Descriptive Statistics for Posttest of Independent Variables with Respect to Each Dependent Variable (social and emotional) for SEMS

Independent Variable	Dependent Variable	Mean	*SD*
Program Type - Comparison (*n* = 52)	Social	2.697	0.718
	Emotional	2.938	0.709
Program Type - Treatment (*n* = 65)	Social	2.685	0.912
	Emotional	3.077	0.620
Gender – Male (*n* = 62)	Social	2.698	0.851
	Emotional	3.071	0.618
Gender – Female (*n* = 55)	Social	2.682	0.810
	Emotional	2.953	0.709

An observation of the histograms showed that several indicated negatively skewed data was not problematic since none were beyond the acceptable level of ± 2 (George & Mallery, 2003). Figures 31 through 38 show the histograms of all the independent variables on the two dependent variables. The histograms were made on SPSS v.27.

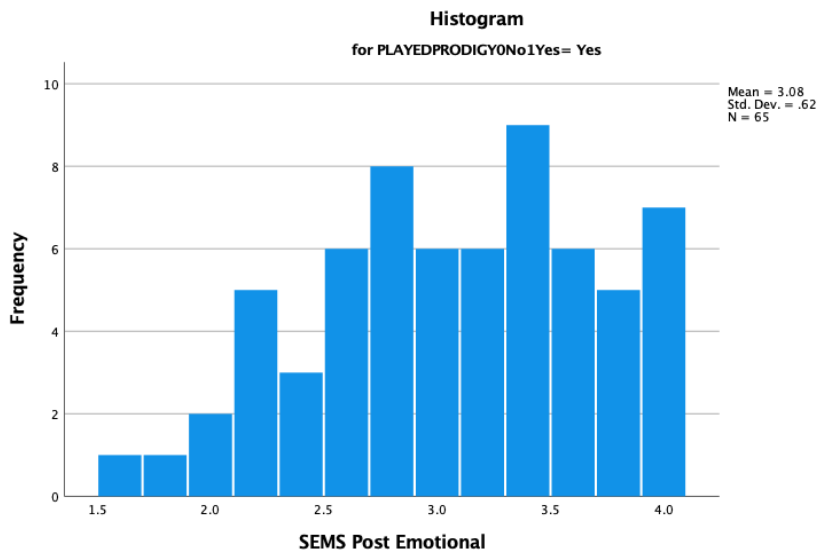

Figure 31. Histogram of the mean posttest scores of the Emotional for treatment group.

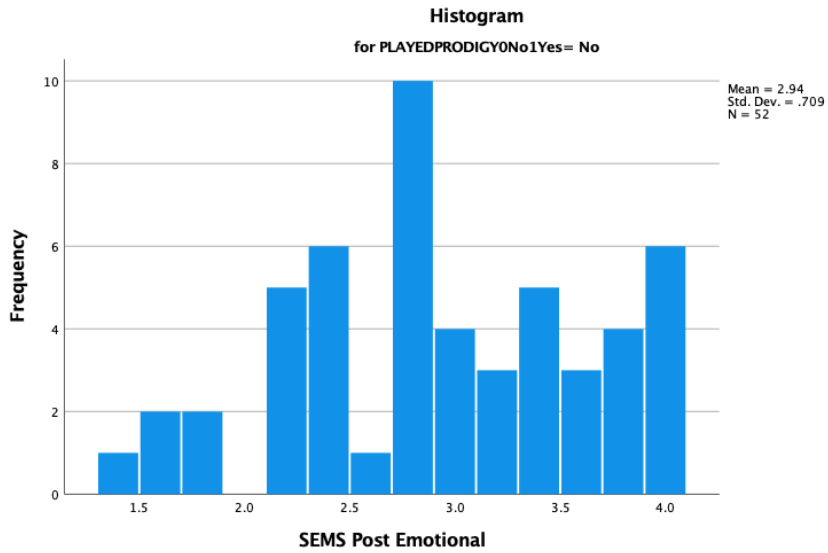

Figure 32. Histogram of the mean posttest scores of the Emotional for comparison group.

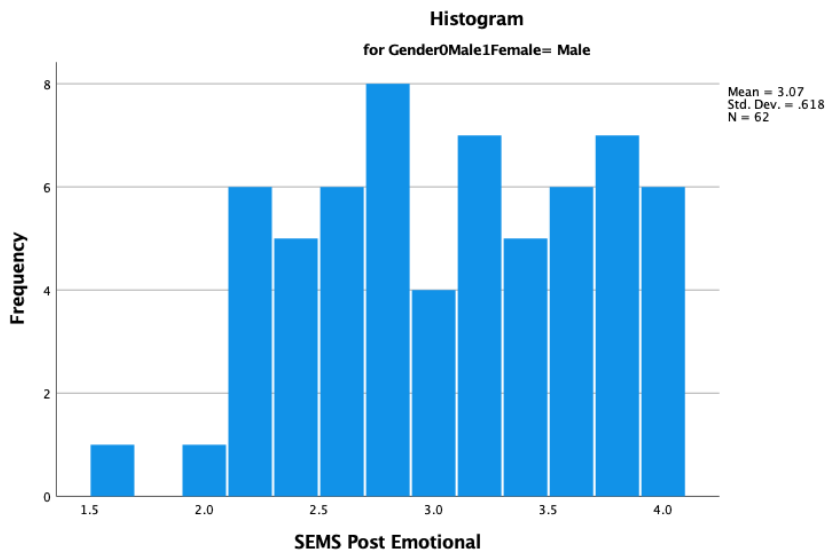

Figure 33. Histogram of the mean posttest scores of the Emotional for males.

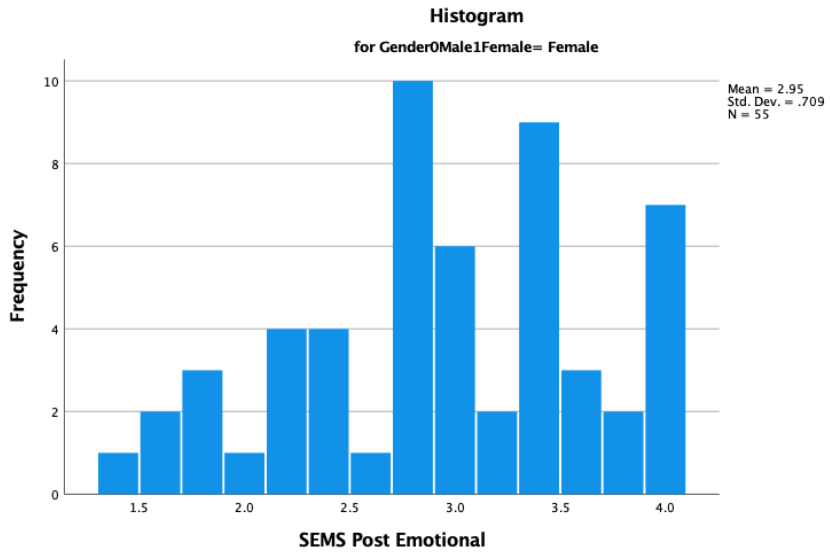

Figure 34. Histogram of the mean posttest scores of the Emotional for females.

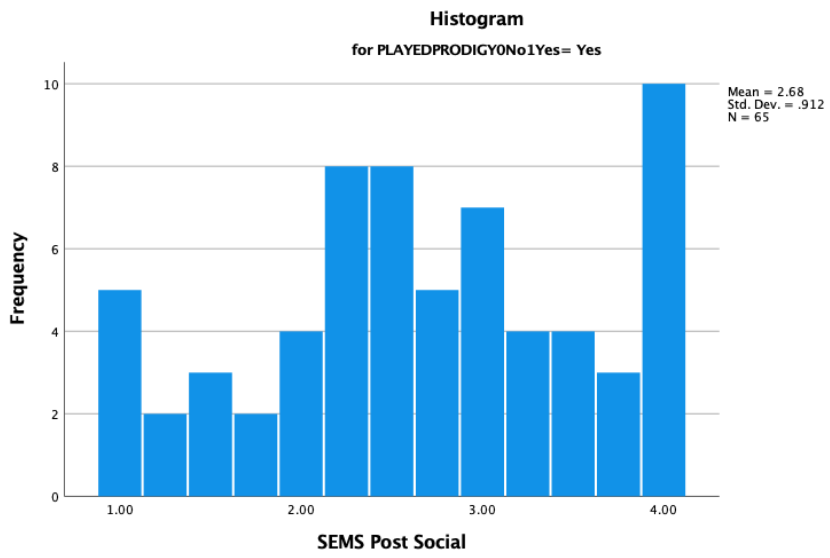

Figure 35. Histogram of the mean posttest scores of the Social for treatment group.

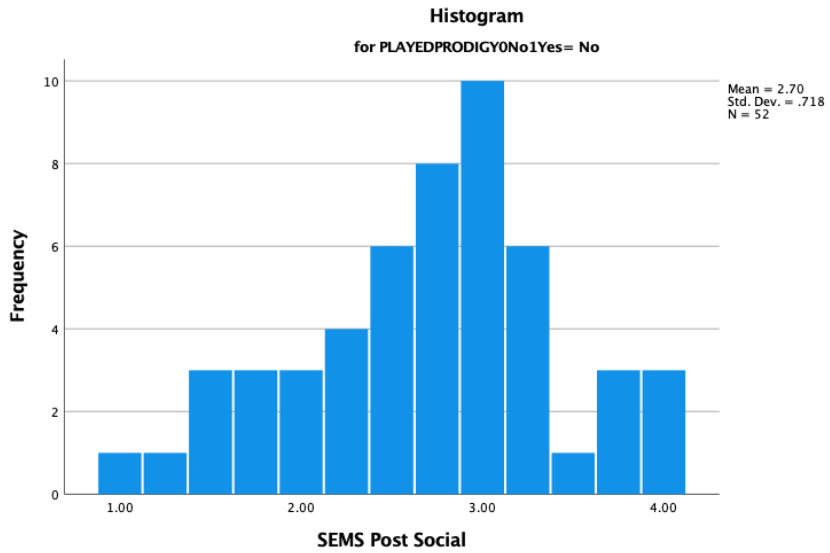

Figure 36. Histogram of the mean posttest scores of the Social for comparison group.

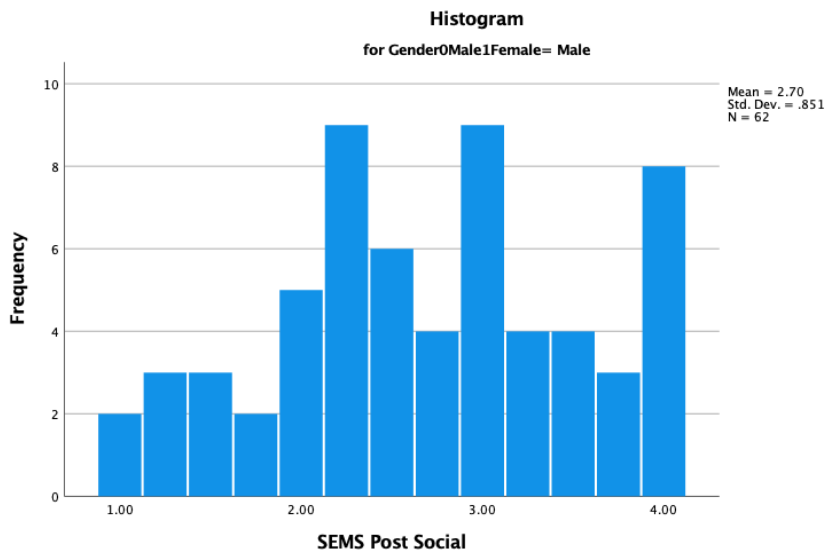

Figure 37. Histogram of the mean posttest scores of the Social for males.

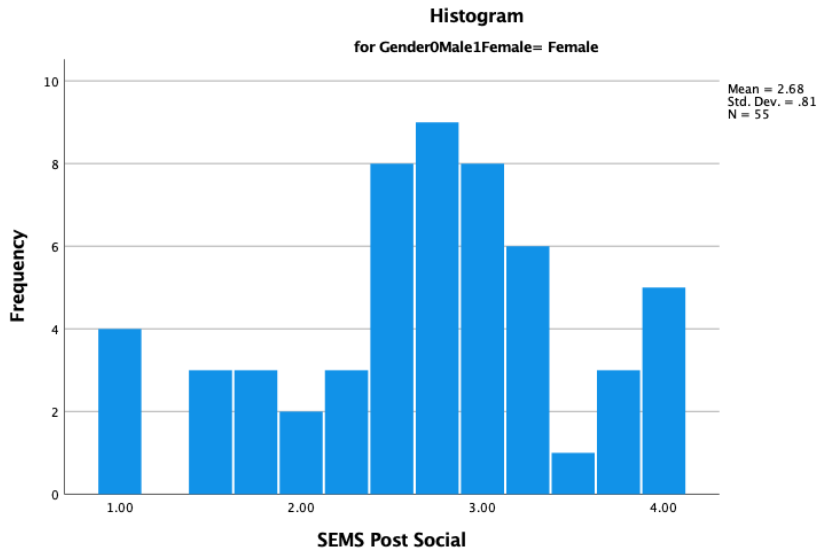

Figure 38. Histogram of the mean posttest scores of the Social for females.

Linearity. To check linearity, the researcher examined the scatterplots of the variables to identify any nonlinear patterns in the data (Hair et al., 2010). None of the plots showed any non-linearity; therefore, the assumption of linearity was satisfied (Pallant, 2016). The scatterplot matrix between each pair of variables can be found in Figures 39 through 42, all scatterplots were made using SPSS v.27.

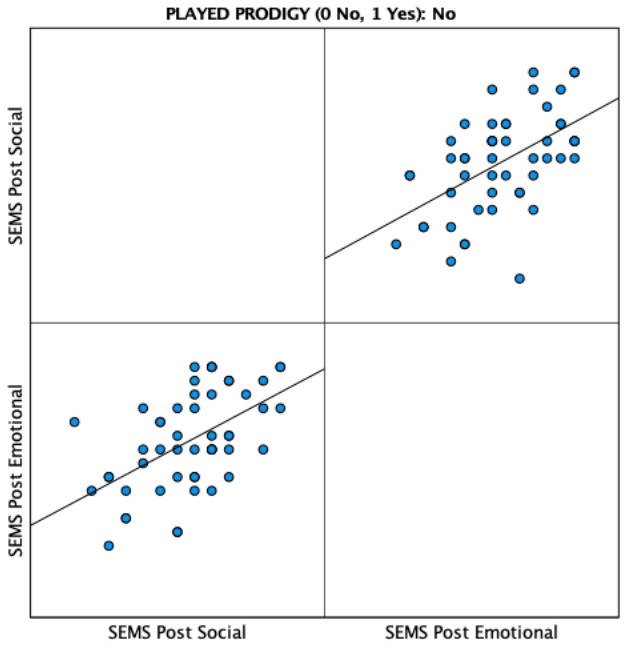

Figure 39. Scatterplot for posttest score for treatment group for engagement (SEMS).

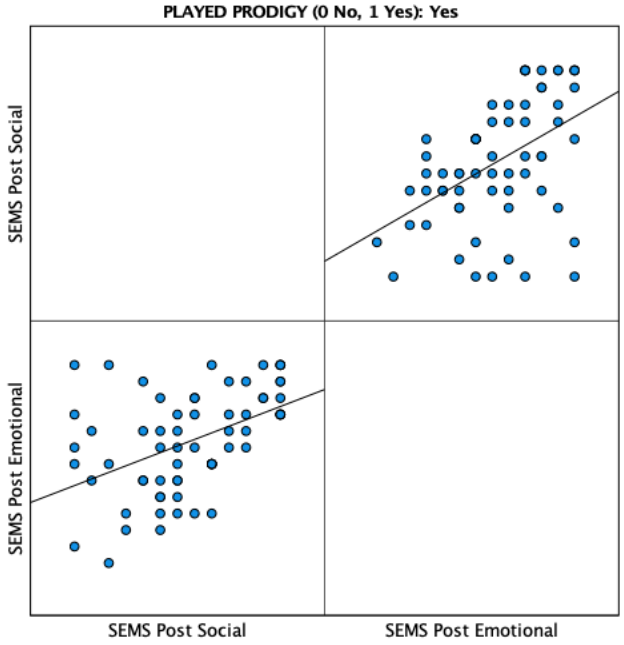

Figure 40. Scatterplot for posttest score for comparison group for engagement (SEMS).

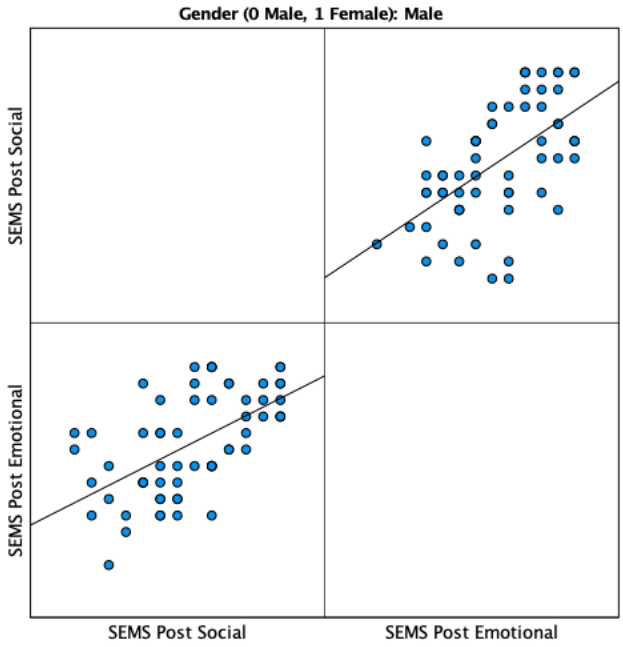

Figure 41. Scatterplot for posttest score for males for engagement.

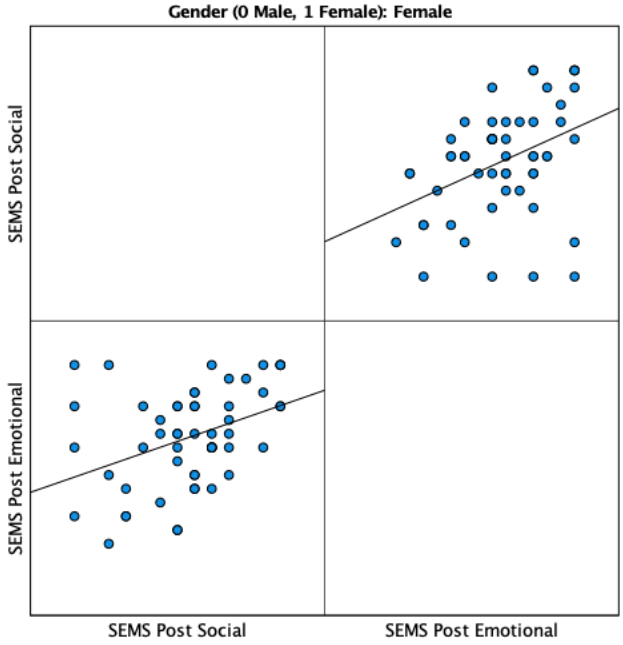

Figure 42. Scatterplot for posttest score for females for engagement.

Correlations were also calculated to check for linearity. Correlations are used to verify that variables are related to each other in a linear way and the Pearson correlation coefficient is

part of the calculations needed for multivariate statistical analysis (Meyers et al., 2006). In order to compare the correlations for two different groups the sample was split to look at specifically gender or program type. The correlations were significant at least at the $p < 0.001$ level for program type and gender as see in Table 66 and Table 67, respectfully.

Table 66

Summary of Correlations by Groups (Social and Emotional)

Program Type		Emotional
Comparison (n = 52)	Social	.538***
Treatment (n = 65)	Social	.470***

Note. ***$p < .001$; Bonferroni adjustment, $p < 3/.05 = .017$

Table 67

Summary of Correlations by Gender (Social and Emotional)

Gender Identity		Emotional
Male (n = 62)	Social	.581***
Female (n = 55)	Social	.396***

Note. ***$p < .001$; Bonferroni adjustment, $p < 3/.05 = .017$

Calculating all the correlations without splitting the groups produced the correlations below in Table 68.

Table 68

Summary of Correlations for all variables (Social, Emotional, Gender and Program Type)

Variables	Gender	Emotional Engagement	Social Engagement
Program Type	-.157	.104	-.008
Gender		-.090	-.010
Emotional Engagement			.487**

Note: For program type, comparison = 0 and treatment = 1. For gender, males = 0 and females = 1. **$p < .001$

Homoscedasticity. To check homoscedasticity, Box's M test for equality of variance-covariance matrices was used (Meyers et al., 2006). The Box's M test was not significant (p = 0.127) meaning the variance of the dependent variables were equivalent across groups. The values for the Levene's Test of Equality of Error Variances also showed that the variances of the dependent variables were equivalent across groups; Social $F(3, 113) = 1.691$, $p = 0.173$ and Emotional $F(3, 113) = 0.232$, $p = 0.874$. Table 69 show the values for the Levene's Test of Equality of Error Variance.

Table 69

Levene's Test of Equality or Error Variances for Engagement

		Levene Statistic	*df*1	*df*2	Sig.
Social	Based on Mean	1.692	3	113	0.173
	Based on Median	1.438	3	113	0.235
	Based on Median and with adjusted *df*	1.438	3	105.293	0.236
	Based on trimmed mean	1.661	3	113	0.180
Emotional	Based on Mean	0.232	3	113	0.874
	Based on Median	0.191	3	113	0.902
	Based on Median and with adjusted *df*	0.191	3	102.368	0.902
	Based on trimmed mean	0.244	3	113	0.865

Results for Research Question Three

The researcher produced a two-way multivariate analysis of variance (MANOVA) using program type (Treatment or Comparison) and gender (male or female) as the independent

variables and the mean posttest scores of the two dimensions on the Student Engagement in Mathematics Scale (Social, Emotional) as the dependent variables.

Results of the two-way MANOVA for the posttest scores showed there was no significant main effect for the independent variable, program type, $F(2, 112) = 0.844$, $p = 0.433$, Wilks' Lambda = 0.985. There was also no significant main effect for the independent variable, gender, $F(2, 112) = 0.689$, $p = 0.504$, Wilks' Lambda = 0.988. There was a statistically significant interaction effect between program type and gender $F(2, 112) = 4.565$, $p = 0.012$, Wilks' Lambda = 0.925. When the results of the dependent variables were considered separately, the only difference to reach statistical significance was emotional, $F(1, 113) = 3.474$, $p = .004$, partial eta squared = 0.070. An inspection of the mean scores indicated that for the emotional engagement scale females who played Prodigy had higher scores ($M = 3.208$, $SD = 0.6202$) than females who did not play Prodigy ($M = 2.724$, $SD = 0.7140$). Table 70 displays the multivariate test results and Table 71 displays the test between subject effects.

Table 70

Multivariate Test for Engagement

Effect		Value	F	Hypothesis *df*	Error *df*	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	0.959	1305.995	2.00	112.000	0.000	0.959
	Wilks' Lambda	0.041	1305.995	2.000	112.000	0.000	0.959
	Hotelling's Trace	23.321	1305.995	2.000	112.000	0.000	0.959
	Roy's Largest Root	23.321	1305.995	2.000	112.000	0.000	0.959
Program	Pillai's Trace	0.012	0.689	2.000	112.000	0.504	0.012
	Wilks' Lambda	0.988	0.689	2.000	112.000	0.504	0.012
	Hotelling's Trace	0.12	0.689	2.000	112.000	0.504	0.012
	Roy's Largest Root	0.012	0.689	2.000	112.000	0.504	0.012
Gender	Pillai's Trace	0.015	0.844	2.000	112.000	0.433	0.015
	Wilks' Lambda	0.985	0.844	2.000	112.000	0.433	0.015
	Hotelling's Trace	0.015	0.844	2.000	112.000	0.433	0.015
	Roy's Largest Root	0.015	0.844	2.000	112.000	0.433	0.015
Program* Gender	Pillai's Trace	0.07	4.565	2.000	112.000	0.012	0.075
	Wilks' Lambda	0.925	4.565	2.000	112.000	0.012	0.075
	Hotelling's Trace	0.082	4.565	2.000	112.000	0.012	0.075
	Roy's Largest Root	0.082	4.565	2.000	112.000	0.012	0.075

Table 71

Tests of Between-Subject Effects for Engagement

Source	Dependent Variable	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.	Partial Eta Squared
Corrected Model	Social	0.326[a]	3	0.109	0.155	0.926	0.004
	Emotional	4.307[b]	3	1.436	3.488	0.018	0.085
Intercept	Social	820.172	1	820.172	1169.387	0.000	0.912
	Emotional	1035.758	1	1035.758	2516.693	0.000	0.957
Program	Social	0.004	1	0.004	0.006	0.937	0.000
	Emotional	0.493	1	0.493	1.198	0.276	0.011
Gender	Social	0.023	1	0.023	0.032	0.857	0.000
	Emotional	0.500	1	0.500	1.216	0.273	0.011
Program* Gender	Social	0.313	1	0.313	0.446	0.506	0.004
	Emotional	3.474	1	3.474	8.441	0.004	0.070
Error	Social	79.255	113	0.701			
	Emotional	46.506	113	0.412			
Total	Social	926.313	117				
	Emotional	1114.640	117				
Corrected Total	Social	79.581	116				
	Emotional	50.812	116				

a. *R* Squared = .004 (Adjusted *R* Squared = -.022)

b. *R* Squared = .085 (Adjusted *R* Squared = .060)

Figure 43 displays the interaction of the posttest means for the emotional engagement scale for gender and program type.

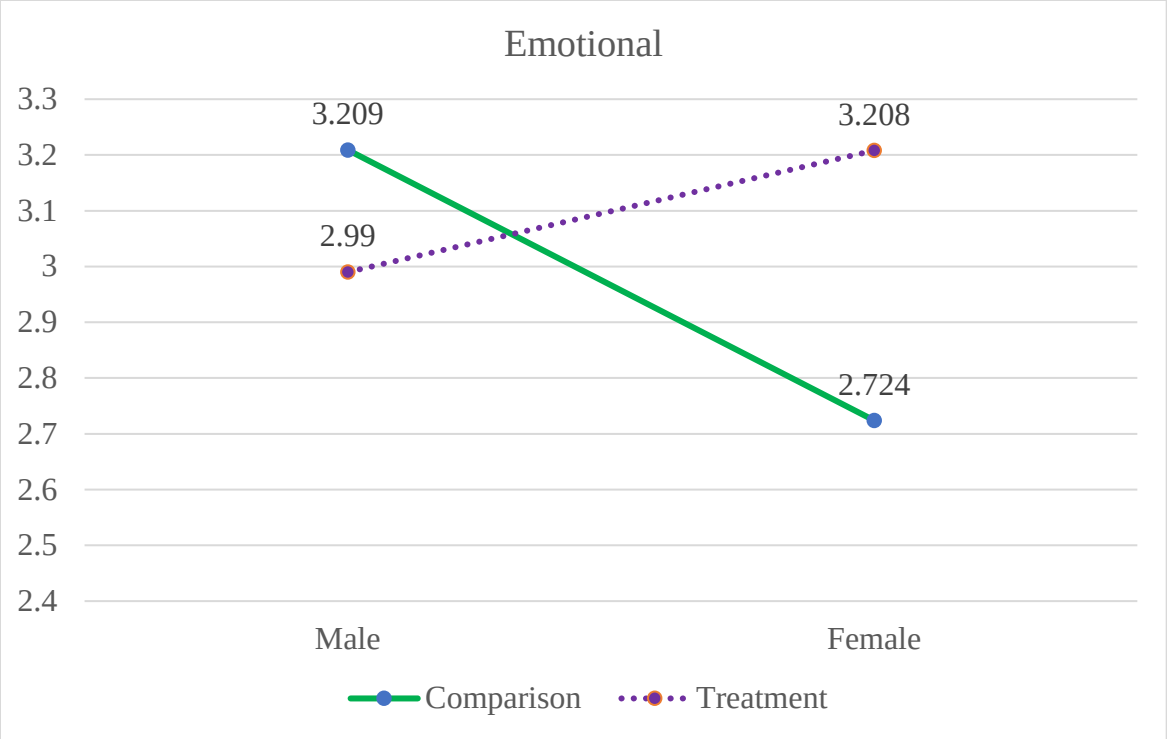

Figure 43. Graph of the Interaction of Gender and Program Type for Posttest Means Regarding Emotional Engagement Scores

The original hypothesis stated that there would be a statistically significant difference in mathematics engagement between gender and program type. This hypothesis was partially met since there was a significant interaction between program type and gender. However, there was no significant main effect for program type and gender.

Summary of Chapter Four

In chapter four, the data and analysis were presented for the three research questions posed by the researcher. For the first research question, a MANOVA was performed to investigate achievement between groups. Two dependent variables were used: Spring STAR and T3 mathematics grades. The independent variables were program type and gender. Mean pretest scores for T2 mathematics grades and Winter STAR scores were statically significant for gender and were therefore used as covariates for the posttest analysis. The posttest results revealed that

there was a statistically significant main effect for gender, $F(2, 193) = 10.726$, $p = 0.000$; Wilks' Lambda = 0.900; partial eta squared = 0.10. Follow up results indicated that, Spring STAR scores showed a statistically significant difference for gender, $F(1, 194) = 19.235$, $p = 0.001$, partial eta squares = 0.090. However, there were no statistically significant difference for T3 mathematics grades, $F(1, 194) = 1.088$, $p = 0.296$, partial eta squared = 0.006. An inspection of the mean scores indicates that males had higher Spring STAR scores ($M = 814.32$, $SD = 4.538$) than females ($M = 784.19$, $SD = 4.729$). There was no significant difference between Program type for any of the dependent variables.

For the second research question, the researcher explored the relationship between Spring STAR scores and the predictors of Gender, T3 mathematics grades, and Time spent on the game. A stepwise regression procedure indicated that, T3 mathematics grades and gender entered into the regression equation to predict Spring STAR scores $F(2, 76) = 12.935$, $p < 0.001$. $R^2 = .254$ indicating that 25.4% of the variance of the Spring STAR could be accounted for by T3 mathematics grades and gender and 74.6% of the variance not accounted for.

For the third research question, a two-way multivariate analysis of variance (MANOVA) using program type (treatment or comparison) and gender (male or female) as the independent variables and the mean posttest scores of the three dimensions on the Student Engagement in Mathematics Scale (cognitive, social, emotional) as the dependent variables. Cognitive engagement was removed from the analysis based on failing the test for normality. Results of the two-way MANOVA for the posttest scores showed there was no significant main effect for the independent variables, program type or gender. There was a statistically significant interaction effect $F(2, 112) = 4.565$, $p = 0.012$, Wilks' Lambda = 0.925. Follow up revealed there were statistically significant difference for females with respect to program type for

emotional engagement, $F(1, 113) = 1.216$, $p = .004$, partial eta squared = 0.70. An inspection of the mean scores indicated that females who played prodigy had significantly higher scores on emotional engagement ($M = 3.208$, $SD = 0.6202$) than females who did not play Prodigy ($M = 2.724$, $SD = 0.7140$).

CHPATER FIVE: SUMMARY AND CONCLSUIONS

The purpose of this study was to investigate the impact of a virtual mathematics game and gender on students' achievement and engagement. This chapter is divided into three sections: (a) an overview and discussion of the study, (b) an explanation of the study's validity, and (c) conclusions.

Overview and Discussion of the Study

The study included a sample of convenience of seventh and eighth grade student (n = 413) and teacher (n = 7) participants. Teachers chose to be a part of either the treatment or comparison group and decided which of their classes participated. Students in the treatment group played Prodigy 1-5 times per week during their enrichment block, while nothing changed during the enrichment block for the control group. All students participated in the regular mathematics curriculum.

Research Questions

Research question one. Using a systematic approach, the following research questions were addressed:

1. For grade 7 and 8 students, is there a significant difference in post student achievement (STAR Assessment, mathematics grade) between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female)?
    a. Is there a statistically significant difference in student achievement (STAR Assessment, mathematics grade) between program types (GBL enhanced mathematics program, traditional mathematics program)?

b. Is there a statistically significant difference in student achievement (STAR Assessment, mathematics grade) between males and females?

c. Is there a statistically significant interaction between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female) for student achievement (STAR Assessment, mathematics grade)?

Non-Directional Hypothesis: There will be a statistically significant difference in post student achievements between gender and program type.

Research question one was analyzed using a quasi-experimental comparison group pretest, posttest design. Data were collected for achievement using students' Trimester 2 (T2) and Trimester 3 (T3) mathematics grades along with students' Winter and Spring STAR scores. Two covariates were applied to the analysis using the pretest data, T2 and Winter STAR, because they had significance while testing the assumptions. A MANOVA procedure was used to investigate if there was a significant difference in post student achievement (Spring STAR Assessment, T3 mathematics grade) between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female) for middle school students.

Results. There was a statistically significant difference between males and females, F (2, 193) = 10.726, p = 0.000; Wilks' Lambda = 0.900; partial eta squared = 0.10. Follow up results revealed a statistically significant effect for gender regarding Spring STAR scores, F (1, 194) = 19.235, p = 0.000, partial eta squared = 0.090. An inspection of the mean scores indicated that males had significantly higher Spring STAR scores (M = 814.320, SD = 4.538) than females (M = 784.189, SD = 4.729). There were not statistically significant differences related to T3 mathematics grades, F(1, 194) = 1.088, p = 0.296, partial eta squared = 0.006. There was no

statistically significant difference for Program type, F (2, 193) = 0.556, p = 0.575 Wilks' Lambda = 0.994; partial eta squared = 0.006 and no statistically significant interaction, F (2, 193) = 1.114, p = 0.330 Wilks' Lambda = 0.989.

Limitations. Based on the preferences of the teacher participants this study could not use random sampling. Instead, the researcher used a sample of convenience of intact groups. In addition, while Prodigy selected items for the students to work on based on the pre-diagnostic test, the content was not necessarily specific to what the students were learning in their mathematics classes.

Relationship to the theoretical foundation. Vygotsky's ZPD is the theoretical framework to investigate the relationship between achievement and the use of games. Vygotsky viewed play as a way to create the ZPD, as play serves as a scaffolding function to support children learning new concepts (Homer et al., 2019). In addition, recognizing the current level of cognition within the ZPD would be the ideal learning situation (Hung, 2001).

Related research. There were four studies that showed an increase in student performance as a result of using games to supplement classroom content (Ke & Grabowski, 2007, Kebritchi et al., 2010; Ku, Chen, Wu, Lao & Chan, 2014; Miller & Robertson, 2011). Ku, Chen, Wu, Lao and Chan (2014) showed that students who played mini games that focused on the math content had a gain in achievement that was significantly greater than that of the comparison group. Kim and Chang (2010) compared two levels of gender, showing that computer games had greater positive effects on males than they did on females.

Implications for educators. For this sample, using Prodigy did not result in a statistically significant difference between treatment and comparison group or interaction between program

type and gender. Therefore, using Prodigy did not distract from the curriculum. It can be used without the game affecting students' achievement in a negative way.

Potential future research. Researchers could investigate the use of the planning feature on Prodigy to see if gains can be seen from pretest to posttest design on a specific topic or standard. Researchers should also investigate the effectiveness of using Prodigy for ELL students.

Research question two. Using a systematic approach, this researcher addressed the following research question:

2. For grade 7 and 8 students, to what degree and in what manner do gender (male, female), mathematics grade, and amount of time spent using the virtual mathematics game predict student achievement on the STAR Assessment?

Non-Directional Hypothesis: Gender, mathematics grades, amount of time spent using the virtual game, will predict student achievement on the STAR assessment.

Research question two used a correlation design. The data collected for the regression analysis revealed the predictive relationship between students' Trimester 3 (T3) mathematics grades, Spring STAR scores, and Total time spent using the virtual mathematics game. A stepwise multiple regression was conducted with Spring STAR as the dependent variable and Gender, Time Spent, and T3 mathematics grades as independent variables.

Results. Using a step-wise linear regression, at step one of the analysis, T3 mathematics grades entered into the regression equation and were significantly related to spring STAR scores $F(1, 78) = 10.047$, $p < 0.002$. The multiple correlation coefficient was 0.34. The adjusted R^2 indicated that 10.4% of the variance of the spring STAR could be accounted for by T3 mathematics grades. At step two of the analysis, gender entered the equation, contributing to the

prediction of spring STAR scores, $F(2, 78) = 12.935$, $p < 0.001$. Females were more likely to have statistically significantly higher T3 mathematics grades, while males were more likely to have statistically significantly higher spring STAR scores. The multiple correlation coefficient was 0.504. The adjusted R^2 indicated 25.4% of the variance of the Spring STAR could be accounted for by T3 mathematics grades and gender, with 74.6% of the variance not accounted for. Total Time playing Prodigy did not enter the equation. Two out of the three of the predictors entered the equation supporting the hypothesis.

Limitations. Experimental mortality was a limitation with this analysis since some participants stopped using Prodigy at different times over the 8-week period. The researcher addressed this by grouping the students into Group 0, comparison group students, who never played Prodigy and Group 3, treatment group students, who played for 7 weeks.

Relationship to the theoretical foundation. Activity theory is the framework to consider everything happening in an activity, the interactions between the subjects and the object necessary to complete the goal. Carvalho and colleagues (2015) developed the Activity Theory-based Model of Serious Games (ATMSG). The ATMSG uses the concepts of activity theory and includes a serious games-component taxonomy. In addition, one can use this model as a tool for the conceptual design of serious games. Serious games, games with a purpose, have been shown to have instructional potential (Carvalho et al., 2015).

Related research. Miller and Robertson (2011) conducted a randomly controlled trial study, with 635 primary school children ages 10 and 11, where the experimental group played a game for 20 minutes a day for nine weeks. All students, regardless of group, had an increase in accuracy of computation and speed of processing. The experimental group displayed a statically significant improvement of 50% in content accuracy as compared to the control group. In

comparison, Kim and Chang (2010), conducted a multi-stage probability study and found that ELL students who played mathematics computer games daily had a higher mathematics performance compared to non-ELL students who did not play games. However, non-ELL male students who played the same games daily performed lower in math.

Implications for educators. The current research showed that the amount of time spent playing Prodigy was not a significant predictor of achievement. This result is seen by other studies indicating that the research is conflicted about the amount of time that is beneficial for students to play academic games (Kim & Chang, 2010; Miller & Robertson, 2011).

Potential future research. Researchers should examine which variable has the greatest impact on achievement: the amount of time students play the game per week or how many weeks they play the game. Since only 25.4% of the variance in Spring STAR scores were predicted by T3 grades and gender, future research should be designed to investigate other variables that predict this criterion.

Research question three. Using a systematic approach, this researcher addressed the following research question:

3. For grade 7 and 8 students, is there a statistically significant difference in mathematics class engagement (social, emotional, cognitive) between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female)?
    a. Is there a statistically significant difference in mathematics class engagement between males and females?

b. Is there a statistically significant difference in student engagement between program types (GBL enhanced mathematics program, traditional mathematics program)?

c. Is there a statistically significant interaction between program type (a GBL enhanced mathematics program, a traditional mathematics program) and gender (male, female) for student mathematics class engagement?

Non-Directional Hypothesis: There will be a statistically significant difference in mathematics engagement between gender and program type.

Research question three used a quasi-experimental comparison group posttest design. Data were collected for engagement using the Student Engagement in Mathematics Scales (Leis, Schmidt, & Rimm-Kaufmann, 2015). This produced three mean scores, one each for cognitive, social, and emotional engagement. The researcher produced a two-way multivariate analysis of variance (MANOVA) using program type (Treatment or Comparison) and gender (male or female) as the independent variables and the mean posttest scores for two of the three dimensions on the Student Engagement in Mathematics Scale (Cognitive, Social, Emotional) as the dependent variables. The Cognitive subscale did not pass the normality assumption and was not used in the analysis.

Results. The two-way MANOVA for the posttest scores showed there was no significant main effect for the independent variables, program type or gender. However, there was a statistically significant interaction effect between program type and gender, $F(2, 112) = 4.565$, $p = 0.012$, Wilks' Lambda = 0.925. When the results of the dependent variables were considered separately, the only mean difference that was statistically significant was emotional engagement, $F(1, 113) = 1.216$, $p = .004$, partial eta squared = 0.70. An inspection of the mean scores

indicated that females who played prodigy had statistically significantly higher scores (M = 3.208, SD = 0.6202) than females who did not play Prodigy (M = 2.724, SD = 0.7140). There was no significant difference between males who played Prodigy or did not play, with respect to their emotional engagement.

Limitations. The sampling procedure for the study was not done with a random sample. Data were missing due to data collection issues.

Relationship to the theoretical foundation. Schlechty's (2002) theory of engagement suggests eight qualities: product focus, clear product standards, protection from adverse consequences for initial failures, affirmation, affiliation, choice, novelty and variety, and authenticity. According to Schlechty (2002, 2011), engagement is based on the tasks assigned to students, which need to be the central concern for the teacher. As stated by Schlechty (2011), "the primary goal of teachers is that of designing engaging work for students and guiding them to the sources of instruction they need to do this work successfully" (p. 58). Authentic engagement requires that the task, activity, or work must be associated with a result or outcome that has a clear meaning and relatively immediate value to the student being engaged (Schlechty, 2002). Engagement paves the way for learning so "assessing engagement is a means of ensuring the possibility of preventing deficiencies in learning" (Schlechty, 2002, p. 41). The results of this study did show that females who played Prodigy were emotionally engaged.

Related research. In a review of three different studies about affect and computer use in the classroom, two showed positive outcomes in student attitudes and confidence (Ku, Chen, Wu, Lao & Chan, 2014; Ke & Grabowski, 2007) and one study showed no change in motivation (Kebritchi, Hirumi, & Bai, 2010). This study showed a positive outcome on females' engagement in the subscale of emotional engagement for those who used Prodigy.

Implications for educators. Playing Prodigy helped female students feel more connected to mathematics. Therefore, it is hypothesized that if females play Prodigy, then they will be more engaged in mathematics. As Burke (2014) discussed engagement happens at an emotional level because people are more invested in achieving and reaching the next level when they are striving to win at a game.

Potential future research. A researcher could conduct an experimental design with random group assignment to treatment or comparison group. Another study could focus on investigating different types of games in classrooms to see if there is a type of game that engages females more often, leading to increased achievement. In addition, observations of students playing Prodigy could help educators to understand what aspects of the game engaged students the most. Lastly, a study can be conducted to examine how engaged ELL students are compared to their non-ELL peers when playing Prodigy.

Study Validity

To address limitations, the researcher needs to consider how validity. Internal validity is, the extent to which the extraneous variables have been controlled and external validity refers to how any variable that is not controlled can affect the experimental outcome of a study (Gall et al., 2003). This section explains the types of threats to or limitations of this study.

Internal Validity

History represents the specific events happening throughout a study (Campbell & Stanley, 1963). This concept may become a threat if an unanticipated event affects the responses of the participants (Gall et al., 2003). The researcher used a large sample size (n = 413) of students from two middle schools in the same school district. During the treatment, the school was closed for Spring Break. All students experienced the same amount of time out of school.

In addition, the researcher met with teachers before the study to discuss end-of-year field trips and other events to make sure they did not impact the study. Because the researcher took the proactive steps, the field trips did not affect the timeline of the treatment or testing and therefore history was deemed a small threat.

Maturation means that participants can have changes physically or psychologically due to natural development **(**Gall et al., 2003). This is especially true with the population that was studied, middle school students. The researcher concluded that the threat of maturation was small since a comparison group was employed.

Testing can become a limitation if students show an improvement on the posttest measure simply as an effect of their experience with the pretest **(**Gall et al., 2003). Regarding the STAR adaptive test, the pretest was given in January and the posttest was given in June, the items were different for each student, depending on how they answered each question, each time they completed the test. Regarding grades and engagement, similar grading guidance was used to maintain validity of student mathematics grades and a valid and reliable tool was used to assess engagement. Testing as a threat using these measures was deemed to be small by the researcher.

The data gathered were grades, time spent using Prodigy, the STAR assessment, and Student Engagement in Mathematics Scale. In addition, there were no pretest data on the Student Engagement in Mathematics Scale. The instrument for the engagement scale had high reliability and validity and since there was no pretest, "the extraneous variable of instrumentation is not a threat to its internal validity" (Gall et al., 2003, p. 371). Therefore, overall testing as a threat using these instruments was deemed to be small.

Instrumentation is a limitation if the measuring instrument does not have adequate validity and reliability and if it is not administered correctly (Meyers et al., 2006). Addressing

the validity and reliability the researcher gave all the teachers detailed instructions regarding how to administer the data instruments used. The STAR and grades assessment are administered by the district guidelines.

Experimental mortality occurs when the "participants might be lost from the experimental or control group because they drop out of the study, miss pretesting or post testing, or are absent during some sessions" (Gall et al., 2003, p. 372). This was experienced in the study when some of the students in the treatment group did not play the virtual mathematics game for the duration of the study. To address this, the students were placed into one of four groups; Group 0 included the comparison group, students who never played Prodigy; Group 1 students played for 1-3 weeks; Group 2 students played for 4-6 weeks; and Group 3 students played for 7+ weeks. There were 28 students who were in the treatment group, but never played the game and they were eliminated from the study. Regarding the analyses for all research questions, it was decided to use Group 0, the comparison group (n = 137), and Group 3, as the treatment group (n = 93). The researcher deemed the threat of experimental mortality to be medium since students were eliminated from the study.

Selection-maturation interaction specifically examines confounding variables (Meyers et al., 2006). The researcher selected participants from two middle schools in the same school district with equivalent demographics. The participants were from the same grade level bands in both schools. Therefore, the researcher deemed the threat of selection-maturation to be low.

Experimental treatment diffusion occurs when the treatment is perceived as highly desirable and then the comparison group may seek access (Gall et al., 2003). The researcher planned for the treatment to occur during an additional block of time in the school schedule to help eliminate contact between the experimental and comparison groups. The threat of

experimental treatment diffusion was perceived as a moderate threat based on the contact being minimized but the students were in the same building.

Compensatory rivalry and demoralization of the control group refers to the comparison group perceiving that the experimental group is receiving a desirable treatment and the demoralization of the comparison group refers to the comparison group feeling left out because they are not getting the treatment and therefore many perform worse (Gall et al., 2003). To address this, a large sample size was used (n = 385). During a weekly check-in, teachers were asked how everything was going and if they noticed anything different with the students' behaviors, such as students in the comparison group asking about the game or discussions about the game outside of the enrichment class. Teachers did inform the researcher that students in the treatment group were excited to play the game and were eager to go to enrichment class. Teachers from the comparison group were given information about the treatment if they wanted to use it in the future. Compensatory rivalry and demoralization were perceived to be low threats as the teachers never indicated that they saw or heard any changed behaviors from students who were in the comparison group, although the students were in the same building.

External Validity

External validity refers to the extent to which the findings of the experiment can be generalizable (Campbell & Stanley, 1963; Gall et al., 2003). Population validity is the "extent to which the results of an experiment can be generalized from the sample that was studied to a specified, larger group" (Gall et al., 2003, p. 374). Readers could generalize the findings only for populations that are demographically similar to the experimental sample employed by the researcher.

Ecological validity, according to Gall et al. (2003), is the ability to generalize the findings from one set of environmental conditions to different environmental conditions. To help generalize the findings in this research, the researcher made sure to give explicit descriptions of the experimental treatment so that it could be replicated by any other researcher.

The Hawthorne effect, as stated by Gall et al. (2003), refers to "any situation in which the experimental conditions are such that the mere fact that individuals are aware of participating in an experiment, are aware of the hypothesis, or are receiving special attention improves their performance" (p. 376). The researcher made sure that no one group received special treatment.

Novelty effect occurs when the treatment might be effective because it is different from what participants normally receive (Gall et al., 2003). Using a virtual game was not something new or novel because students had been exposed to virtual math games in the past. Therefore, the researcher deemed the threat of novelty low.

Experimenter effect occurs when the treatment is effective or not because of who administered the treatment (Gall et al., 2003). The researcher provided each teacher with the timeline, sent out weekly reminders, and checked in with each teacher weekly. From the weekly check in all teachers indicated that they were following the prescribed research plan.

Conclusion

The researcher found there was no significant difference between students who played Prodigy and those who did not with respect to student achievement. In addition, the multiple regression showed that time spent on the virtual mathematics game was not a predictor of students' STAR math scores. The researcher concluded that Prodigy does not distract from the curriculum and can be used without affecting student achievement in a negative way. The researcher suggests that future research be conducted to target specific mathematical content or

standards using Prodigy to see if a more targeted use, focusing on specific content or standards, will increase student achievement.

This study did show that there was a statistically significant interaction effect between program type and gender with the emotional subscale reaching statistical significance for females who played Prodigy. The researcher concluded that females should be given the opportunity to play the virtual mathematics game as the game helped the students feel more connected and engaged in this mathematical activity. Future research should be done with a randomized sample, comparing different types of games, and observing females playing Prodigy.

In conclusion, "student engagement is a necessary element for improving student outcomes" (Reschly & Christenson, 2012, p. 3) and if a virtual mathematics game can increase that engagement then that is something to invest in, for all students. As one teacher exclaimed during the study, they had "never seen their students so excited for their enrichment block than on the days they got to play Prodigy."

References

Ainley, M. (2012). Students' interest and engagement in classroom activities. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 283–302). Springer. https://doi.org/10.1007/978-1-4614-2018-7

Anastasiadis, T., Lampropoulos, G., & Siakas, K. (2018). Digital game-based learning and serious games in education. *International Journal of Advances in Scientific Research and Engineering*, *4*(12), 139–144. http://doi.org/10.31695/IJASRE.2018.33016

Boykin, A. W., & Noguera, P. (2011). *Creating the oppurtunity to learn*. ASCD.

Buckley, P., Doyle, E., & Doyle, S. (2017). Game on! Students' perceptions of gamified learning. *Journal of Educational Technology & Society*, *20*(3), 1–10. http://www.jstor.org/stable/26196115

Burke, B. (2014). *Gamify: How gamification motivates people to do extaordinary things*. Bibliomotion.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally & Company.

Carvalho, M. B., Bellotti, F., Berta, R., De Gloria, A., Sedano, C., Hauge, J., Hu, J., & Rauterberg, M. (2015). An activity theory-based model for serious games analysis and conceptual design. *Computers & Education, 87*, 166–181. https://doi.org/10.1016/j.compedu.2015.03.023

Connecticut State Department of Education (CSDE). (2018). *Strategic school profiles*. Retrieved from https://public-edsight.ct.gov

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). *From game design elements to gamefulness: Defining "gamification"*. ACM.

Dichev, C., & Dicheva, D. (2017). Gamifying education: what is known, what is believed and what remains uncertain: a critical review. *International Journal of Educational Technology in HIgher Education*, *14*(9). https://doi.org/10.1186/s41239-017-0042-5

Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gamification in education: A systematic mapping study. *Journal of Educational Technology & Society*, *18*(3), 75–88. http://www.jstor.org/stable/jeductechsoci.18.3.75

Engestrom, Y., Miettinen, R., Raija-Leena, P., & Punamaki, R. L. (1999). *Perspectives on activity theory* (1st ed.). Cambridge University Press.

Ernest, P. (1986). Games. A rationale for their use in the teaching of mathematics in school. *Mathematics in School*, *15*(1), 2–5. https://www.jstor.org/stable/30216298

Gall, M. D., Gall, J. P., & Borg, W. R. (2003). *Educational research: An introduction* (7th ed.). Pearson Education.

George, D., & Mallery, P. (2003). *SPSS for windows step by step: A simple guide and reference* (4th ed.). Pearson Education.

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Education.

Harrison, C., Alvermann, D. E., & Afflerbach, P. (2017). What is engagement, how is it different from motivation, and how can I promote it? *Journal of Adolescent & Adult Literacy*, *61*(2), 217–220. https://www.jstor.org/stable/26631102

Homer, B. D., Raffaele, C., & Henderson, H. (2019). Games as playful learning: Implication of developmental theory for game-based learning. In *Handbook of game-based learning* (pp. 25–52). The MIT Press.

Hung, D. (2001). Design principles for web-based learning: Implications from Vygotskian thought. *Educational Technology Publications, Inc.*, *41*(3), 33–41. https://www.jstor.org/stable/44428668

Hwa, S. P. (2018). Change in mathematics learning: Harnessing the power of digital game-based learning. *Journal of Eudcation Technology & Society*, *21*(4), 259–276. https://www.jstor.org/stable/10.2307/26511553

Jonassen, D. H., & Rohrer-Murphy, L. (1999). Activity theory as a framework for designing constructivist learning environments. *Educational Technology Research and Development*, *47*(1), 61–79. https://doi.org/10.1007/bf02299477

Jones, S. M. (1982). Fitting games into a mathematics curriculum. *National Council of Teachers of Mathematics*, *30*(4), 35–36. www.jstor.org/stable/41192104

Kapp, K. M., Blair, L., & Mesch, R. (2014). *The gamification of learning and instruction fieldbook*. Wiley.

Ke, F., & Abras, T. (2013). Games for engaged learning of middle school children with special learning needs. *British Journal of Educational Technology*, *44*(2), 225–242. https://doi.org/10.1111/j.1467-8535.2012.01326.x

Ke, F., & Grabowski, B. (2007). Gameplaying for maths learning: Cooperative or not? *British Journal of Educational Technology*, *38*(2), 249–259. https://doi.org/10.1111/j.1467-8535.2006.00593.x

Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effects of modern mathematics computer games on mathematics achievement and class motivation. *Computers & Education*, *55*, 427–443. https://doi.org/10.1016/j.compedu.2010.02.007

Kim, S., & Chang, M. (2010). Computer games for the math achievement of diverse students. *Journal of Educational Technology & Society*, *13*(3), 224–232. http://www.jstor.org/stable/jeductechsoci.13.3.224

Klein, J. D., & Freitag, E. (1991). Effects of using a instructional game on motivation and performance. *The Journal of Educational Research*, *84*(5), 303–308. http://www.jstor.org/stable/40539699

Ku, O., Chen, S. Y., Wu, D. H., Lao, A. C. C., & Chan, T.-W. (2014). The effects of game-based learning on mathematical confidence and performance: High ability vs. low ability. *Journal of Education Technology & Society*, *17*(3), 65–78. https://www.jstor.org/stable/10.2307/jeductechsoci.17.3.65

Lee, J. J., & Hammer, J. (2011). Gamification in education: what, how, why bother? *Academic Exchange Quarterly*, *15*(2). https://www.researchgate.net/publication/258697764_Gamification_in_Education_What_How_Why_Bother

Leis, M., Schmidt, K. M., & Rimm-Kaufman, S. E. (2015). Using the partial credit model to evaluate the student engagement in mathematics scale. *Journal of Applied Measurement*, *16*(3), 251–267. https://files.eric.ed.gov/fulltext/ED563451.pdf

Leonard, L. M., & Tracy, D. M. (1993). Using games to meet the standards for middle school students. *The Arithmetic Teacher*, *40*(9), 499–503. https://www.jstor.org/stable/41195843

Liu, T. Y., & Chu, Y. L. (2010). Using ubiquitous games in an English listening and speaking course: Impact on learning outcomes and motivation. *Computers & Education*, *55*, 630–643. https://doi.org/10.1016/j.compedu.2010.02.023

Mahimker, R. (2014). *Can game-based learning increase standardized test scores?* [White paper]. Prodigy Math Game. https://www.prodigygame.com

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation*. Sage Publications.

Microsoft Corporation. (2021). *Microsoft Excel*. Retrieved from https://office.microsoft.com/excel

Miller, D. J., & Robertson, D. P. (2011). Educational benefits of using game consoles in a primary classroom: A randomized controlled trial. *British Journal of Educational Technology, 42*(5), 850–864. https://doi.org/10.111/j.1467-8535.2010.01114.x

Morrison, J. R., Risman, K. L., Reilly, J., & Eisinger, J. M. (2020). *An evolution of prodigy: A case-study approach to implementation and student achievement outcomes*. John Hopkins University.

National Council of Teachers of Mathematics. (2020). *Catalyzing change in middle school mathematics: Initiating critical conversations*. The National Council of Teachers of Mathematics.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards Mathematics*.

Pallant, J. (2016). *SPSS survival manual* (6th ed.). McGraw Hill Education.

Perrotta, C., Featherstone, G., Aston, H. and Houghton, E. (2013). *Game-based learning: latest evidence and future directions* (NFER Research Programme: Innovation in Education). Slough: NFER.

Peters, A. & Mahimker, R. (2011) Prodigy (Online version) [video game]. www.prodigygame.com

Plass, J. L., Mayer, R. E., & Homer, B. D. (Eds.). (2019). *Handbook of game-based learning*. The MIT Press.

Renaissance. (2014). *The research foundation for star assessments: The science of star* [White paper]. Wisconsin Rapids, WI: Renaissance Learning.

Renaissance Learning Inc. (n.d.). Renaissance: See every student. Retrieved May 5, 2021, from https://www.renaissance.com

Reschly, A. L., & Christenson, S. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 3–19). Springer. https://doi.org/10.1007/978-1-4614-2018-7

Sáez-López, J. M., Miller, J., Vázquez-Cano, E., & Domínguez-Garrido, M. C. (2015). Exploring application, attitudes and integration of video games: MinecraftEdu in middle school. *Journal of Educational Technology & Society*, *18*(3), 114–128. https://www.jstor.org/stable/jeductechsoci.18.3.114

Schlechty, P. C. (2002). *Working on the work*. Jossey-Bass.

Schlechty, P. C. (2011). *Engaging students: The next level of working on the work*. Jossey-Bass.

Schwartz, R. N., & Plass, J. L. (2019). Types of engagement in learning with games. In *Handbook of game-based learning* (pp. 53–82). The MIT Press.

Shapiro, E. S., Dennis, M. S., & Fu, Q. (2015). Comparing computer adaptive and curriculum-based measures of math in progress monitoring. *School Psychology Quarterly*, *30*(4), 470-487. https://psycnet.apa.org/record/2015-11187-001

IBM Corp. Released 2020. IBM SPSS Statistics for Macs, Version 27.0. Armonk, NY: IBM Corp.

Sun-Lin, H.-Z., & Chiou, G.-F. (2019). Effects of gamified comparison on sixth graders' algebra word problems solving and learning attitudes. *Journal of Educational Technology & Society*, *22*(1), 120–130. https://www.jstor.org/stable/10.2307/26558833

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wohlwend, K., & Peppler, K. (2015). All rigor and no play is no way to improve learning. *Phi Delta Kappan, 96*(8), 22–26. https://doi.org/10.1177/0031721715583957

Appendix A

IRB approval

Carol O'Connor

Tue 2/19/2019 1:10 PM

To: Amanda Peterson <peterson081@connect.wcsu.edu>;

Cc: WCSU IRB <wcsuirb@wcsu.edu>; Marcia Delcourt <delcourtm@wcsu.edu>; Kristy Zaleta <zaletak@wcsu.edu>;

Hello Amanda Peterson,

I am pleased to inform you that your I.R.B. protocol number 1819-118 has been approved by full review. This email is documentation of your official approval to start your research. If you need a copy of this official approval for funding purposes, please let me know oconnorc@wcsu.edu. The WCSU I.R.B. wishes you the best with your research.

You have 1 year from the date of this email to complete your research; if you are still conducting that date, you will need to fill out a renewal application. When are you finished with your study please fill out and return via email a Termination/Completion Report (available here: http://wcsu.edu/irb/forms.asp) so we know your study is complete.

Finally – and most importantly! – we have recently learned that current BOR technology policies do *not* guarantee privacy of *any* info stored on work computers physically, remotely, or otherwise (i.e., laptop, Dropbox, etc.). As such, to maintain the truth of any anonymity or confidentiality promises you make to participants (consent form, for example), you will need to store all electronic data obtained from those human subjects on a system/computer/file *not* connected to any CSU system. It is your responsibility as the primary researcher to make sure personal data of participants remains securely private – something not guaranteed in the currently existing CSU system. *Rest assured,(because it's ridiculous to expect faculty to store work-related research on non-work-related systems and/or to conduct research where participants are not guaranteed anonymity/confidentiality), we are working to gain an exception for research purposes to this policy. But until then, it's technically and legally possible for anyone in the system office to access your participants' data at any time – without your consent or knowledge before doing so… which makes any guarantees made on research documents (e.g., consent forms) deceptive unless info is stored elsewhere.*

Thanks,
Jessica Eckstein, Ph.D.
Chair, Institutional Review Board
Western Connecticut State University
www.wcsu.edu/irb

Appendix B

February 2019

Dear (Superintendent),

My name is Amanda A. Peterson. I am a doctoral candidate for the EdD in Instructional Leadership program at Western Connecticut State University and am implementing a dissertation research study. I have chosen to investigate the impact of a virtual mathematics game program, called Prodigy, on student achievement and mathematics classroom engagement. I have developed a 13-week intervention aimed at increasing student mathematics achievement scores and engagement in mathematics. I believe that the results of this investigation will demonstrate to teachers and administrators that the Prodigy virtual game program can provide students with a way to obtain intervention and practice of their mathematics skills in an engaging way. It is proposed that grade 7 and 8 teachers and students at the School A and B [Rogers Park Middle School and Broadview Middle Schools] will be asked to participate in the study. This study will be available to all mathematics teachers. They may either voluntarily choose to be a part of the treatment group, comparison group, or to opt out completely. The treatment will take place during the student enrichment blocks and will not interfere with the school curriculum. Teachers and students in the intervention classes at each middle school will also be invited to participate in the study. There will be no comparison group for the intervention classes.

As part of the study, I will be collecting five types of student information. I plan to administer a demographic survey for general background information and a pre- and post-survey of the Student Engagement in Mathematics Scale (SEMS). Each of these will take approximately 10 minutes to complete during a class activity time via Google Forms. I will request students' STAR Mathematics scores and mathematics grades from trimester two and three. These data will be provided by the STEM Department Head or a school administrator. Finally, the Prodigy program computes detailed summary data regarding each student's skill development and time on task.

I will ask the teachers to complete a brief online demographic survey via Google Forms and maintain a student attendance log to record students who were absent on treatment days and amount of time assigned to using Prodigy.

Participation in this study is completely voluntary and participants may withdraw at any time. Teachers who consent to participate will obtain training in the virtual mathematics game program and receive a weekly check-in and coaching as needed. Parents will be asked if their students may participate using an opt out consent process. The students will have two weeks to hand in the parent opt out form if they do not want their student to be included in the study. Students will also be asked for their assent to participate. Participation in this study will not affect a student's grade. If requested, the Prodigy program will be offered to all students at the end of the study. Privacy will be protected for all participants (district, school, teacher, student) by numerical coding. All identities will be maintained in a secure location to protect confidentiality. Results will only be reported in aggregate form.

The research project has been reviewed and approved by the WCSU Institutional Review Board. If you have questions concerning the right of subjects involved in this research study, please email the WCSU Assurances Administration at irb@wcsu.edu and mention Protocol

Number (1819-118). This study is valid until (1 year from approval date). If you have any questions, please feel free to contact me or one of my advisors.

Sincerely,
Amanda A. Peterson
Western Connecticut State University Doctoral Student

Dr. Kristy Zaleta, Principal
Principal Rogers Park Middle School

Dr. Marcia A. B. Delcourt
Coordinator, EdD in Instructional Leadership Program
Western Connecticut State University

I agree to allow this research to be conducted in (district name):

_____	_____	_____
Superintendent Name (Print)	Superintendent Signature	Date

Appendix C

February 2019

Dear (Principal),

My name is Amanda A. Peterson. I am a doctoral candidate for the EdD in Instructional Leadership program at Western Connecticut State University and am implementing a dissertation research study. I have chosen to investigate the impact of a virtual mathematics game program, called Prodigy, on student achievement and mathematics classroom engagement. I have developed a 13-week intervention aimed at increasing student mathematics achievement scores and engagement in mathematics. I believe that the results of this investigation will demonstrate to teachers and administrators that the Prodigy virtual game program can provide students with a way to obtain intervention and practice of their mathematics skills in an engaging way. It is proposed that grade 7 and 8 teachers and students at the School A and B [Rogers Park Middle School and Broadview Middle Schools] will be asked to participate in the study. This study will be available to all mathematics teachers. They may either voluntarily choose to be a part of the treatment group, comparison group, or to opt out completely. The treatment will take place during the student enrichment blocks and will not interfere with the school curriculum. Teachers and students in the intervention classes at each middle school will also be invited to participate in the study. There will be no comparison group for the intervention classes.

As part of the study, I will be collecting five types of student information. I plan to administer a demographic survey for general background information and a pre- and post-survey of the Student Engagement in Mathematics Scale (SEMS). Each of these will take approximately 10 minutes to complete during a class activity time via Google Forms. I will request students' STAR Mathematics scores and mathematics grades from trimester two and three. These data will be provided by the STEM Department Head or a school administrator. Finally, the Prodigy program computes detailed summary data regarding each student's skill development and time on task.

I will ask the teachers to complete a brief online demographic survey via Google Forms and maintain a student attendance log to record students who were absent on treatment days and amount of time assigned to using Prodigy.

Participation in this study is completely voluntary and participants may withdraw at any time. Teachers who consent to participate will obtain training in the virtual mathematics game program and receive a weekly check-in and coaching as needed. Parents will be asked if their students may participate using an opt out consent process. The students will have two weeks to hand in the parent opt out form if they do not want their student to be included in the study. Students will also be asked for their assent to participate. Participation in this study will not affect a student's grade. If requested, the Prodigy program will be offered to all students at the end of the study. Privacy will be protected for all participants (district, school, teacher, student) by numerical coding. All identities will be maintained in a secure location to protect confidentiality. Results will only be reported in aggregate form.

The research project has been reviewed and approved by the WCSU Institutional Review Board. If you have questions concerning the right of subjects involved in this research study, please email the WCSU Assurances Administration at irb@wcsu.edu and mention Protocol

Number (1819-118). This study is valid until (1 year from approval date). If you have any questions, please feel free to contact me or one of my advisors.

Sincerely,
Amanda A. Peterson
Western Connecticut State University Doctoral Student

Dr. Kristy Zaleta, Principal
Principal Rogers Park Middle School

Dr. Marcia A. B. Delcourt
Coordinator, EdD in Instructional Leadership Program
Western Connecticut State University

I agree to allow this research to be conducted in (district name):

_____	_____	_____
Principal Name (Print)	Principal Signature	Date

Appendix D

February 2019

Dear (Participating Teacher),

My name is Amanda Peterson and I am a doctoral candidate for the Educational Instructional Leadership at Western Connecticut State University and am implementing a dissertation research study. I have chosen to investigate the impact of a virtual mathematics game program, called Prodigy, on student achievement and mathematics classroom engagement. I have developed a 13-week intervention aimed at increasing student mathematics achievement scores and engagement in mathematics. It is my hope that the results of this investigation will enable students and teachers to have a program that can provide students with a way to obtain intervention, extension, and practice in an engaging way in their mathematics classes.

As part of the study, I will be administering the Student Engagement in Mathematics Scale (SEMS) twice to the students and using the students' STAR Mathematics scores and the students' mathematics grades from trimester two and three. All students will complete the assessments. The Student Engagement in Mathematics Scale (SEMS) will take no more than 10 minutes to administer to the students via Google Forms. I will also be collecting demographic information from the students and teachers. I will ask that you maintain a student attendance log during the time of the study.

As a teacher you will have a choice to be a part of the treatment group, the comparison group or to opt out completely. Teachers who choose to be a part of the treatment or comparison will administer the SEMS twice one pre and post, the student demographic survey and complete a teacher demographic survey. The teachers who choose to be a part of the treatment group will be asked to have their students use Prodigy at least twice a week for 15 minutes with a total minimum of 30 minutes per week and keep a student attendance log.

Your participation in this study, as well as that of the students, is completely voluntary and participants may withdraw at any time. Parents will be asked if their students may participate using an opt out consent process. The students will have two weeks to hand in the parent opt out form if they do not want their student to be included in the study. Students will also be asked for their assent to participate. Participation in this study cannot be used as part of a student's mathematics grade. Privacy will be protected for all participants (district, school, teacher, student) by numerical coding. All identities will be maintained in a secure location to protect confidentiality. Results will only be reported in aggregate form.

The research project has been reviewed and approved by the WCSU Institutional Review Board. If you have questions concerning the right of the subject involved in research studies, please email the WCSU Assurances Administration at irb@wcsu.edu and mention Protocol Number (1819-118). This study is valid unit (1 year from approval date). If you have any questions, please feel free to contact me.

Sincerely,
Amanda A. Peterson
Danbury High School
Western Connecticut State University Doctoral Student

Dr. Kristy Zaleta, Principal
Principal Rogers Park Middle School

Dr. Marcia A. B. Delcourt
Coordinator, EdD in Instructional Leadership Program
Western Connecticut State University

I agree to allow this research to be conducted in (district name):

Teacher Name (Print)	Teacher Signature	Date

Appendix E

February 2019
Dear Parents or Guardians,

My name is Amanda Peterson and I am a doctoral candidate for the Educational Instructional Leadership at Western Connecticut State University and am implementing a dissertation project. The purpose of this 13-week study is to determine if a virtual mathematics game program, called Prodigy, has an effect on student mathematics achievement and engagement.

I would like your child to participate in this study. At the beginning and end of the project, I would like to collect (1) your student's STAR Mathematics scores and (2) the mathematics grades. As part of the study, I will ask your student to: (1) answer a brief survey regarding his or her experience in math class using the Student Engagement in Mathematics Scale (SEMS) at the beginning and end of the study, (2) complete a demographic survey about information such as current grade level and frequency of playing electronic games. All students in the project will participate in the regular mathematics curriculum and will not miss any of their mathematics classes. Some classrooms will be using a virtual mathematics game program called Prodigy for at least 30 minutes per week during enrichment time, while the remaining students will not.

Participation in this study is completely voluntary. You are free to withdraw your child from the study at any time, students who agree to participate will submit all survey information to me, the researcher, using Google Forms. Privacy will be protected, as students' names will be a unique number. All student identities will be maintained in a secure location to protect confidentiality. Results will only be reported in aggregate form. Participation or non-participation will not affect a student's grade.

The research project has been reviewed and approved by the WCSU Institutional Review Board. If you have questions concerning the right of the subject involved in research studies please email the WCSU Assurances Administration at irb@wcsu.edu and mention Protocol Number (1819-118). This study is valid unit (1 year from approval date). If you have any questions or concerns, please feel free to contact me.

Sincerely,
Amanda A. Peterson
Danbury High School
Western Connecticut State University Doctoral Student

Dr. Kristy Zaleta, Principal
Principal Rogers Park Middle School

Dr. Marcia A. B. Delcourt
Coordinator, EdD in Instructional Leadership Program
Western Connecticut State University

DENIAL OF PERMISSION

Please complete this form and return to your child's teacher ONLY if you do NOT wish your child to be a part of this study.

Student Name

_____

School _____

I **DO NOT** give permission for my child, named above, to participate in this study.

Parent/Guardian Name

_____

Parent/Guardian Signature _____ Date _____

Appendix F

February 2019

Dear Student,

I am a doctoral candidate for the Educational Instructional Leadership at Western Connecticut State University. I am doing a research study and I would like you to be a part of it. I will send a permission slip home with you for your parents or guardians to look over.

The study is about how a virtual mathematics game program, called Prodigy, may have an effect on student mathematics achievement and engagement. Some of you will be asked to use the game for at least 30 minutes per week during your enrichment period and some of you will not use the game during the study, but can have the choice of using it after my study is over.

Two times during this study I will ask you to complete the Student Engagement in Mathematics Scale (SEMS), each will take about 10 minutes to complete. I will also ask you to complete a demographic survey with general information such as your current grade level and specific information about how often you play electronic games. I will also collect information about your school achievement and math grades. For those of you who use Prodigy during the study, I will be collecting the data about your progress in the game. This study will help me investigate how virtual mathematics programs can help you be engaged in math and improve your achievement.

Your participation in the study is completely voluntary. If you do decide to participate, I will not use your name in the study. I will use numbers instead of names. You will not be graded on this for your math class. All the information will be kept private.

If you have any questions, please ask me. If you would like to be in the study, please print and sign your name below.

Thank you,
Ms. Peterson
Danbury High School
Western Connecticut State University Doctoral Student

Dr. Kristy Zaleta, Principal
Principal Rogers Park Middle School

Dr. Marcia A. B. Delcourt
Coordinator, EdD in Instructional Leadership Program
Western Connecticut State University

Student Name (Print)	Student Signature	Date

Appendix G

Teacher Demographic Survey[1]

1. Name
2. Gender
    a. Male
    b. Female
    c. Choose not to answer
3. What is your ethnicity? Please choose one of the following:
    a. African American
    b. Asian – Pacific Islander
    c. Hispanic
    d. Native American
    e. White – non – Hispanic
    f. Two or more
    g. Prefer not to answer
4. What middle school do you teach at?
    a. RPMS (School A)
    b. BMS (School B)
5. What grade level that you primarily teach?
    a. 7
    b. 8
6. I teach in cluster:
    a. A/1
    b. B/2
    c. C/3
7. How many years have you been teaching?
8. How many years have you been teaching math?
9. How many years have you been teaching in this school district?
10. What type of academic games do you play in your mathematics class? (please check all that apply)
    a. Bingo
    b. Board Games
    c. Card based games
    d. Kahoot!
    e. Moby Max
    f. Minecraft
    g. Prodigy
    h. IXL
    i. Quizlet
    j. Other
11. How often do you play academic games in your mathematics class within a unit?
    a. 0

[1] This will be made into a Google Form and administered digitally.

b. 1
c. 2
d. 3
e. 4
f. 5+

12. Do you play academic games in your mathematics class every day?
    a. Yes
    b. No

Appendix H

Student Demographic Survey[2]

1. Name
2. Gender
    a. Male
    b. Female
    c. Choose not to answer
3. What is your ethnicity? Please choose one of the following:
    a. African American
    b. Asian – Pacific Islander
    c. Hispanic
    d. Native American
    e. White – non – Hispanic
    f. Two or more
    g. Prefer not to answer
4. Which Middle School do you attend?
    a. RPMS (School A)
    b. BMS (School B)
5. What is your grade level?
    a. 7
    b. 8
6. My cluster is:
    a. A/1
    b. B/2
    c. C/3
7. I have math class period/block:
    a. 71 or 81
    b. 72 or 82
    c. 73 or 83
    d. 74 or 84
    e. 75 or 85
8. I like math class . . .
    a. Very much
    b. Somewhat
    c. A little
    d. Not at all
9. What type of academic games do you play in mathematics class? (please check all that apply)
    a. Bingo
    b. Board Games
    c. Card based games
    d. Kahoot!
    e. Moby Max

[2] This will be made into a Google Form and administered digitally.

f. Minecraft
g. Prodigy
h. IXL
i. Quizlet
j. Provide the name of any others

10. How many academic games do you play in mathematics class every week?
    a. 0
    b. 1
    c. 2
    d. 3
    e. 4 or more

10. Do you play academic games in mathematics class every day?
    a. Yes
    b. No

11. How often do you play any digital games (on a phone, on a system, on a computer) in one day?
    a. 0 (I do not play digital games)
    b. 1 (I play a game 1 time during the day)
    c. 2 – 3 (I play a game(s) 2 to 3 times a day)
    d. 4 – 5 (I play a game(s) 4 to 5 times a day)
    e. 6 + (I play a game(s) 6 or more times a day)

12. How many hours in a given day do you play digital games?
    a. I do not play digital games
    b. Less than an hour (< 1)
    c. Between 1 – 2 hours
    d. Between 2 – 3 hours
    e. More than 3 hours

13. On what device do your primarily play digital games?
    a. Phone
    b. Gaming System (PS4, Xbox, Wii, etc.)
    c. Computer
    d. Other
    e. I do not play games on a device

Dissertation Registration Form

Student: Amanda A. Peterson Date: 06/19/2024

Dissertation Title: THE IMPACT OF VIRTUAL MATHEMATICS GAMES ON STUDENT ACHIEVEMENT AND ENGAGEMENT IN THE MATHEMATICS CLASSROOM

Dissertation Committee Members: See attached Dissertation Approval Page

For Office Use Only.

Marcia A. B. Delcourt, PhD	*Marcia A. B. Delcourt*	06/19/2024
Dissertation Committee Co-Chair	Signature	Date

Marcia A. B. Delcourt, PhD	*Marcia A. B. Delcourt*	06/19/2024
Program Coordinator	Signature	Date

Joan Palladino, EdD	Joan S. Palladino	7/16/2025
Dean, School of Professional Studies	Signature	Date