12-2023

Using Gamification to Foster Student Resilience and Motivation to Learn, And Using Games to Teach Significance Testing Concepts in the Statistics Classroom

Todd Partridge
Utah State University, todd.partridge@usu.edu

USING GAMIFICATION TO FOSTER STUDENT RESILIENCE AND MOTIVATION TO LEARN, AND USING GAMES TO TEACH SIGNIFICANCE TESTING CONCEPTS IN THE STATISTICS CLASSROOM

by

Todd Partridge

A dissertation submitted in partial fulfilment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Statistics & Instructional Technology and Learning Sciences

Approved:

Kady Schneiter, Ph.D. Major Professor	Jody Clarke-Midura, Ed. D. Committee Member
John Stevens, Ph.D. Committee Member	Rebecca Bayeck, Ph.D. Committee Member
Jürgen Symanzik, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2023

ABSTRACT

Using Gamification to Foster Student Resilience and Motivation to Learn, and Using Games to Teach Significance Testing Concepts in the Statistics Classroom

by

Todd Partridge, Doctor of Philosophy

Utah State University, 2023

Major Professor: Dr. Kady Schneiter
Department: Mathematics & Statistics

This dissertation comprises two projects. The first used gamification techniques gleaned from several Super Mario Bros. video games to inform a new grading structure in a statistics classroom in an effort to remove barriers to student motivation and resilience when faced with difficulty or failure in the classroom. Evidence was shown that some barriers to motivation were removed, and that students were able to take full advantage of the materials and resources provided them without getting disheartened and doing less than their best. An evaluation of the gamified grading structure was performed, and strategies for ensuring that future iterations are even more effective were suggested. The second project explored the beliefs and strategies participants developed while playing "Your Average Game," a unique game developed by Partridge with the aim of providing a hands-on way for students to construct and discover the steps of a hypothesis test for a mean on their own. Participants displayed many behaviors and thought processes relevant to learning and understanding hypothesis testing as taught in introductory statistics courses while playing. A case is made for utilizing the game as a teaching tool in such courses.

(82 pages)

PUBLIC ABSTRACT

Using Gamification to Foster Student Resilience and Motivation to Learn, and Using Games to Teach Significance Testing Concepts in the Statistics Classroom

Todd Partridge

Two studies are outlined in this dissertation.

In the first study, elements of Super Mario Bros. videos games were used to change the way college students in a beginners' statistics course were graded on their work. This was part of an effort to help students remain optimistic in the face of challenging coursework and even failure on assignments and tests. The study shows that the changes made to the grading structure did help students to keep trying and to use the materials given to them by their professor until they achieved their desired grade in the course, and suggests ways to make the gamified grading structure even more effective in future uses of the program.

In the second study, an online activity was created where players engage in a game of deception against each other, and the tools of the game encourage players to naturally perform steps of a hypothesis test as taught in beginners' statistics courses in order to determine whether their opponent is lying to them. The study shows that players of the game naturally began to take actions and ask questions that foster an effective environment for learning about the more formal steps of performing a hypothesis test, and that this game may be a useful tool for educators to use to help their students learn about these complicated processes in a fun and natural way.

ACKNOWLEDGMENTS

I give special thanks to my wife, Tiana, who made sure I didn’t skip too many meals and who was proud of me even on the days that I felt like I accomplished next to nothing. I also want to thank my mother, Keturah, who convinced me several times over to keep going when I was ready to move on without finishing my doctorate. I must also thank my best friend, Ben, who gave me constant encouragement and fixed up my house when I didn’t have the time or energy to. The outpouring of love and support I received from these three is unparalleled and made these projects possible.

My brother, Ty, also deserves a big thank you, since without his hard work, tutelage, and patience, “Your Average Game” never would have become a functioning online application.

There are several members of my doctoral committee who have significantly impacted the creation of this paper, including Dr. John Stevens, who guided me through my very first teaching assignment at the university, Dr. Jürgen Symanzik, who taught me most of what I know about R programming and who spent hours marking and re-marking my papers for precision and accuracy, and Dr. Jody Clarke-Midura, who really opened my eyes to the beauty and utility of Instructional Technology and Learning Sciences.

Finally, I would like to thank my major professor, Dr. Kady Schneiter, who supported and tutored me without judgment or disappointment, even when I missed deadlines or focused on my teaching rather than my research. Her expertise and versatility as an educator inspired a great deal of my dissertation work because I knew that her passion to help her students better understand statistics would drive these projects to the quality that I hoped for.

Todd Partridge

CONTENTS

Page

Abstract iii

Public Abstract iv

Acknowledgments v

Contents..vi

List Of Tables.. viii

List Of Figures ix

Chapter I – Dissertation Introduction 1

1.1 Introduction 1

1.1.1 Background and Context 1
1.1.2 Problem Statement 2
1.1.3 Research Question 3
1.1.4 Relevance and Importance of the Research 3

Chapter II – Project 1 – Evaluation of Effects of Gamifying Grading Strategies on Student Motivation and Resilience 5

2.1 Introduction 5

2.1.1 Background and Context 5
2.1.2 Problem Statement 6
2.1.3 Research Questions 6
2.1.4 Relevance and Importance of the Research 6

2.2 Literature Review 7

2.2.1 Key Concepts, Theories and Studies 7
2.2.2 Key Debates and Controversies 9
2.2.3 Gaps in Existing Knowledge 9

2.3 Research Design 10

2.3.1 Participants and Setting 10
2.3.2 Implementation 10
2.3.3 Evaluation Questions 14
2.3.4 Data Collection 15

2.4 Discussion 15

2.4.1 Quantitative Results 15

2.4.2 Student Response 18
2.4.3 Evaluation Results 20
2.4.4 Future Implementations 23

2.5 Conclusion 28

Chapter III – Project 2 – A Qualitative Analysis of the Utility of “Your Average Game” as a Tool for Introductory Statistics Students to Construct and Explore the Steps of a Hypothesis Test for a Mean 30

3.1 Introduction 30

3.1.1 Background and Context 30
3.1.2 Problem Statement 33
3.1.3 Research Questions 34
3.1.4 Relevance and Importance of the Research 35

3.2 Literature Review 36

3.2.1 Key Concepts, Theories and Studies 36
3.2.2 Key Debates and Controversies 38
3.2.3 Gaps in Existing Knowledge 39

3.3 Research Design 40

3.3.1 Participants and Setting 40
3.3.2 Research Interviews 41
3.3.3 Asynchronous Surveys 43
3.3.4 Practical Considerations 44
3.3.5 Qualitative Coding Methods 47

3.4 Discussion 50

3.4.1 Interview Results 50
3.4.2 Survey Results 52
3.4.3 Emerging Themes 54
3.4.4 Evidences of Conducting Hypothesis Tests 58
3.4.5 Misconceptions 61
3.4.6 Future Classroom Use 62

3.5 Conclusion 64

Chapter IV – Dissertation Implications and Conclusion 66

4.1 Implications and Contributions to Knowledge 66

4.1.1 Practical Implications 66
4.1.2 Theoretical Implications 66
4.1.3 Research Limitations 68
4.1.4 Future Work 68

References 70

Appendix: “Your Average Game” Rules and Example Playthrough........................... 76

LIST OF TABLES

Page

Table 1: Gamified grading elements as based off Super Mario Bros. elements 11

Table 2: Gamified grading logic model 14

Table 3: Gilbert's performance matrix 24

Table 4: Categories of possible useful interventions for the gamified grading structure, with programs, strategies, and tools labeled 26

Table 5: Emergent strategies and beliefs of interviewees as Poser 51

Table 6: Emergent strategies and beliefs of interviewees as Peeker 51

Table 7: Evidences of hypothesis steps being formally or informally taken by interviewees 52

Table 8: Emergent themes from asynchronous survey questions 52

LIST OF FIGURES

Page

Figure 1: Stars earned by students and their final grade 16

Figure 2: Assignment attempts made by students and their final grade....................... 17

Figure 3: Student unit test scores 22

Figure 4: Student total test attempts 22

Figure 5: Two game options on site homepage.. 30

Figure 6: Player 1's screen when playing as the Poser 31

Figure 7: Player 2's screen when playing as the Peeker 32

Figure 8: Warning of decreasing points when choosing to peek 32

Figure 9: Results screen after Peekers have made their decisions 33

CHAPTER I – DISSERTATION INTRODUCTION

1.1 Introduction

1.1.1 Background and Context

This dissertation is a combination of two different projects that utilize games and gamification techniques in an attempt to improve students' experience and understanding in college-level statistics courses, with some possible relevance for other types of courses as well.

The first project, discussed in Chapter II, involved utilizing gamification techniques from Nintendo's Super Mario Bros.$^{©}$ video games to alter the grading structure of a college-level statistics course. Encouraging evidence was found that this gamified approach to grading students' work not only gave students as many opportunities as they needed to master the material, but also removed some barriers to students' motivation to take advantage of those opportunities. This gamified grading approach could possibly be utilized in other types of courses as well, though it would probably be difficult to implement in a class where a majority of the coursework by necessity must be graded by a human, such as a creative writing course.

The second project, discussed in Chapter III, explored the results of playing a game called "Your Average Game" which has been developed by Partridge. Several undergraduate statistics students played the game while describing their thought processes and strategies throughout their participation. Though the participants had not yet been formally taught the steps of a hypothesis test for a mean in a statistics course, every one of them intuitively came up with most or all of the steps themselves as they played the game. Many participants, without coaching or instruction, began to recognize challenges of formal hypothesis testing within 30 minutes of playing the game, such as the need to balance the cost of obtaining a larger sample with the benefit that larger sample can give, and how difficult it can be to detect minute differences between a hypothesized mean and the true population mean. This is encouraging evidence that “Your Average Game” can be an effective tool for allowing introductory statistics students to construct and explore the steps of a formal hypothesis test for a mean themselves before they are taught through lecture or other similar classroom method, to promote a more intuitive understanding of the process and to help students see that the steps they are learning are the most natural way to approach the problem, rather than a series of overcomplicated rules developed by an old mathematician.

The dissertation as a whole pairs this example of an actual game being used in the teaching of statistics with an example of gamification being used to enhance the classroom experience, in order to show the merit that games and gamification have in fields of teaching and statistics, as well as add to the currently small pool of resources available for helping instructors to use these methods in their own classrooms.

1.1.2 Problem Statement

There is a host of studies showing that using games (Subhash & Cudney, 2018) and gamification (Ortiz et al., 2016) can be an effective way of teaching in higher education. A difficulty for instructors is finding specific gamification approaches which successfully address individual learning objectives or particular issues with student motivation in the classroom (Deif, 2017). Educational researchers and practitioners both struggle with identifying when, where, and how to use gamification design concepts (Huang et al., 2020).

The researchers desire to provide a new specific use of gamification in the classroom to help remove barriers to student motivation and resilience, as well as a new game designed to bolster student achievement in a specific learning objective taught in introductory statistics courses, along with evidence of their effectiveness and insight into where and when these new approaches may be most suitable in the classroom.

1.1.3 Research Question

The overarching question to be addressed through the combined dissertation comprising both projects is: How do the methods of utilizing games and gamification in these two projects contribute to creating an environment where students feel increased motivation to both learn and master statistical concepts and procedures?

1.1.4 Relevance and Importance of the Research

The researchers believe that any novel approach to teaching mathematics and statistics that allows students to explore the subject from a new or unique experience or perspective, and effectively helps students correctly understand and implement that material, is worth investigating.

The insights gained from the projects in this dissertation are relevant to introductory statistics instructors across high school and college student age groups, as well as to members of teams or councils who make decisions about how mathematics and statistics should or can be taught at their institutions. The dissertation not only provides a relevant game and gamification procedure designed to help students learn more effectively which instructors can immediately modify and employ in their own classrooms, but also provides useful and specific examples of how to develop and use gamification techniques that can be extended to new and innovative games and methods that cover other course concepts.

Many instructors are looking for distinctive and engaging ways to both teach and evaluate course material that will be encouraging, effective, and memorable. The primary aim of this dissertation is to provide more such materials for instructors to use and modify in order to make introductory statistics students' experience with the subject an even more constructive and enduring one.

CHAPTER II – PROJECT 1 – EVALUATION OF EFFECTS OF GAMIFYING GRADING STRATEGIES ON STUDENT MOTIVATION AND RESILIENCE

2.1 Introduction

2.1.1 Background and Context

While teaching a higher-level introductory statistics course at a university for several years, the researchers noticed that there were many students that would experience “burnout”, or loss of motivation to learn the material partway through the semester. It was also noted that, as the course progressed, more students’ questions would be focused on how they could raise their grade than on how to better understand the course material.

After reading many studies about gamification and its ability to promote engagement in the classroom, the researchers investigated to see how gamifying the way students’ coursework is graded might change the way they approached and pursued their grade.

This chapter is derived from an article published in the College Teaching journal on July 3, 2023, copyright Taylor & Francis, available online: http://www.tandfonline.com/10.1080/87567555.2023.2227985 (Partridge & Schneiter, 2023). Some wording from the article has been changed, and some sections rearranged, to conform to the voice and formatting of the dissertation. Furthermore, some of the literature reviewed in the article was omitted, and some added, to provide a more targeted review in Section 2.2.1, and additional literature was reviewed for the dissertation to provide insight into debates and controversies in Section 2.2.2, as well as gaps in current knowledge in Section 2.2.3.

2.1.2 Problem Statement

An intervention that removes physical, mental, or emotional barriers to the students' motivation to learn, foster greater resilience to "failures" during the learning process, and encourage students to focus more on the course material than on their current letter grade, would be instrumental in helping instructors overcome the issues the researchers observed in their own classrooms.

The problem explored in this study was whether the gamified approach to a classroom grading structure developed by the researchers could have a positive impact in any of these desired intervention areas.

2.1.3 Research Questions

The questions explored throughout this project pertained to the implementation of the researchers' gamified grading structure in a college-level introductory statistics course of men and women with typical ages between 20 and 26:

Research Question 1: Does this gamification approach increase student motivation to learn and engagement in course material?

Research Question 2: Does this gamification approach increase student resilience to failure in the classroom?

Research Question 3: Does this gamification approach help students focus more on understanding the material than on their course grade?

Research Question 4: How can this gamification approach be modified to better accomplish these goals in future iterations?

2.1.4 Relevance and Importance of the Research

This evaluation is important because, though this course is a common entry-level course, it is also one of the most difficult courses that a student will have participated in thus far in their college career. The increased difficulty of the material compared to other classes tends to significantly decrease students' motivation to learn or expose them to a stark "failure" in the work they turn in that is greater than they have experienced. Any methods that can remove barriers to student motivation and resilience are important to find and implement for such a class.

The stakeholders in this program and evaluation are the professors of the course, their teaching assistants, and the hundreds of students that take the class each semester. This evaluation will help professors and teaching assistants to learn more about the nature of the barriers to student motivation and resilience and how to remove them in a way that is not too time consuming while still being effective. It will also help students by ensuring that they experience a system that removes traditional pressures of "the grade" in a way that allows them to confidently and consistently learn the important material in the class more effectively than they might have otherwise.

2.2 Literature Review

2.2.1 Key Concepts, Theories and Studies

Very few things have changed at the core of education in the last few decades. Smart investors put their money into things that are changing and adapting with the times. Yet, education may be one of the most static, stagnant systems that the general society continues to invest in. (Hebert, 2018)

Through many extensive studies, it has been shown that there are several elements of "classic" education often become huge barriers to a student's willingness to engage with the education process, such as that grades are assigned to every piece of work a student participates in (Kohn, 2013), failures throughout the learning process can be high-stakes risks due to grading procedures (Jones et al., 2003), and the learning mechanisms for the course and its gradebook lack room for student autonomy (Currie, 2014).

Creating entirely new classroom structures is not always feasible due to limited resources and administrative restraints. Educators looking for useful change consequently search for alternative adjustments that can be simply executed in order to enhance learning. Gamification is one of these adjustments that may present itself as a useful, cost-effective, and efficient tool for educators to improve learning outcomes. (Sanchez et al., 2020)

The most widely accepted definition of "gamification" was given by Deterding et al. (2011) as "the use of design elements characteristic for games (rather than game-based technology, full-fledged games, or even playfulness) in non-game contexts."

There are a great number of effective game elements that keep players trying at a difficult task over and over again, even in the face of a large number of failures. These game elements have the potential to help students approach their learning experience with the same tenacity, and also combat some of the barriers listed above, generally giving the player autonomy to approach a task in different ways, and low-stakes risks as there is always the opportunity to try again.

Yu-Kai Chou (2019), a gamification pioneer and researcher, claimed that good gamification doesn't start with game elements. It starts with how it motivates our core drives. He has spent years developing a theory called the "Octalysis" framework, which outlines how different game elements can utilize eight different core drives all humans share in order to keep them motivated to do something. When utilized correctly in the classroom, gamification techniques based on these core drives can help students feel a visceral motivation to learn course material for reasons other than "I have to do it for the grade."

2.2.2 Key Debates and Controversies

There are not many dissenters to the idea that gamification in the classroom can be effective, but there are a few researchers expressing warnings or hesitations on the subject. Von Ahn & Dabbish (2008) stated that most gamification practices are predicated on the belief that games are inherently fun, whereas some students may not share this belief.

Referring to the large number of empirical studies which have been done on the effects of gamification of education, Majuri et al. (2018) observed, "In terms of the results of the reviewed studies, a considerable majority of the studies reported mainly positively oriented results. However, while the results seem promising, there is also a significant amount of research with null or mixed results. As pointed out in the analysis, the reports of qualitative results often indicate very varying experiences and outcomes even when the general tendency of the findings would be positively oriented. Consequently, the findings regarding the considerable majority of research reporting positively leaning results should be considered with caution."

2.2.3 Gaps in Existing Knowledge

Notwithstanding the extensive use and expanding investigation into gamification, the results of gamification, as well as its academic foundations, still lack understanding (Landers et al., 2018).

While many researchers have shown the effects of various gamification techniques on students' final grades at the end of a unit or course, very little research has been done on the effects of applying gamification techniques to the grading structure itself. *The Multiplayer Classroom* by Sheldon (2011) and a select few follow-up studies such as Gressick & Langston's (2017) have done some encouraging research on the use of a “level-up” mechanic within a course grading system, but this looks to be about the extent of the literature in this area.

2.3 Research Design

2.3.1 Participants and Setting

The researchers implemented a gamified grading structure in a higher-level introductory statistics course of 136 college students primarily in their freshman and sophomore year. Students participated in a large lecture-style class twice a week, as well as in a recitation with about 30 of their classmates to review and practice the material introduced during the lectures. Many students had taken some introductory statistics at the high school level, but for many, this was their first statistics course.

The gamified grading scheme was implemented during the Fall semester of 2020. Due to issues with the COVID-19 pandemic such as remote course delivery and unique distractions and difficulties for each student, it is difficult to identify which changes came from the new grading scheme, and which came from the ramifications of a worldwide pandemic on students’ experiences.

2.3.2 Implementation

Several of the new Super Mario Bros.$^{©}$ video games released by Nintendo were examined, since these video games have remained in the top charts of video game sales across all player ages for decades (Richter, 2020), and common game mechanics that could be incorporated into a grading structure were identified. Outlined in Table 1 are game elements from Super Mario Bros.$^{©}$ and the corresponding adjustments to the grading structure that were made:

Table 1: Gamified Grading Elements as Based Off Super Mario Bros. Elements

Super Mario Bros. Element	**Grading Structure Adjustment**
Levels can be replayed infinitely, but have a rigid time limit. *(Octalysis: Scarcity and Impatience)*	All assigned material leading up to tests are infinitely resubmittable, but many materials have a "star" that can only be earned by turning in the first attempt by the soft due date.
Each level has several tokens you can earn. These tokens are later used to unlock world bosses. *(Octalysis: Development and Accomplishment)*	Assigned materials leading up to tests are not graded with a score, but rather with a number of "stars" that can be earned in different ways. A set number of "stars" earned unlocks the corresponding unit test for a student.
Not every level in a world must be completed to unlock the world boss. *(Octalysis: Ownership and Possession)*	Various assignments, tasks, quizzes and challenge prompts were assigned, each with "stars" available to earn. Double the "stars" needed to unlock a test are offered, so students can choose how they want to earn them.
Boss levels can be replayed infinitely, and are the primary means of saving the princess. *(Octalysis: Epic Meaning and Calling)*	A set number of additional "stars" earned by a student can unlock each retake of a unit test the student desires. Unit tests are the primary contributors to the final grade.

To summarize, students were provided with four different types of assignments throughout the course which they could submit in order to earn "stars". These different assignments could be resubmitted infinitely, but Assignments and Tasks each had one "star" that could only be earned by submitting a completed first attempt within a certain timeframe. The four different types of assignments were (1) Assignments, which assessed a basic understanding of definitions and applicable formulas through multiple-choice and free-response questions, where retakes were worded like the original Assignment, but numbers provided in the context would change, (2) Challenge Prompts, which were listed at the end of each Assignment, and required the student to do some prompted research and learning beyond what had been taught in lecture, and demonstrate that they understand and can apply the ideas presented in the outside research, which did not change between attempts, (3) Tasks, which required students to take ideas and processes which had been learned in class and demonstrate them in some kind of 'real-world' scenario through a personal project or study, where resubmissions just involved including more depth or effort to their responses, and (4) Quizzes, which presented students with opportunities to show more advanced and involved ways of using or interpreting formulas and algorithms that are used in class, where questions were pulled from question banks and could differ from retake to retake.

Students would receive personalized feedback with their scores on Tasks and Challenge Prompts since they had to be graded by hand. Assignment and Quiz scores did not come with personalized feedback as they were graded automatically, but students were encouraged to bring questions about any of the resubmittable assignments to recitations or to office hours for more clarification.

In an effort to establish formative assessments as low-stakes ways of discerning a student's progress, only summative assessments (unit tests) contributed to the final grade. All assignments, tasks, quizzes, and challenge prompts leading up to a test were used solely to give students opportunity to "unlock" the unit test. Unit 1 "stars" went toward unlocking the Unit 1 test, Unit 2 "stars" when to unlocking the Unit 2 test, and so forth. Each of the three unit tests required 16 "stars" to be unlocked, out of the 34 possible "stars" per unit. These 34 "stars" were broken down thus:

- Twelve stars for Assignments: three stars per Assignment, with four Assignments
    - One star for timely completion & submission
    - One star for completion with at least 75% accuracy
    - One star for completion with at least 90% accuracy
- Eight stars for Challenge Prompts: one star per Challenge Prompt, with two Challenge Prompts per Assignment
- Six stars for Tasks: three stars per Task, with two Tasks
    - One star for timely completion & submission
    - One star for completion with adequate work

  - One star for completion with superior work

- Eight stars for Quizzes: two stars per Quiz, with four Quizzes
  - One star for at least 75% correct
  - One star for at least 90% correct

For the average student, 16 “stars” generally looked like the timely submission of all four Assignments with 80% accuracy, as well as getting 80% of questions correct on each of the four quizzes, though earned “star” distribution varied from student to student.

Any retake of a Unit Test required three additional “stars”, so a student taking their first retake needed to have earned a total of 19 “stars,” for their second retake they needed to have a total of 22 “stars,” and so on. Each student received a slightly different test, as the online quiz pulled questions from several question banks, with each question bank designed to test for understanding of a particular learning objective for the course. Retakes, then, were also randomized and differed from a student’s original test, though they tested the same objectives at the same difficulty level. The highest score among a student’s attempts on a test was the recorded score for the final grade.

The goal in implementing this approach was that students would concentrate more on understanding the unit material than on getting a good grade on every assignment.

2.3.2.1 Logic model for the gamified grading structure

Table 2 illustrates the logic model for the evaluation of the gamified grading program, as defined by Mertens & Wilson (2019).

Table 2: Gamified Grading Logic Model

Inputs	Activities	Outputs	Outcomes	Impacts
- A clear syllabus with detailed outline of how to earn "points". - Preparation of 15 large question banks for testing understanding. - A reworking of the Canvas system to fit the unique grading system. - Many hours put in by teachers and teaching assistants for grading and regrading assignments. - Creation of 24 to 36 difficult yet attainable challenge prompts.	- 12 infinitely resubmittable assignments. - 24-36 infinitely resubmittable challenge prompts. - 12 infinitely retakeable Canvas quizzes. - 6 infinitely resubmittable application tasks/projects. - 3 Unit Tests that can be unlocked/retaken using points earned from other activities.	- Total number of hours teachers and teaching assistants spent grading. - Total number of quizzes and tests provided to students. - Total number of opportunities given to students to refine their answers to prompts.	- Proportion of students that pass the class. - Mean number of submissions and resubmissions per assignment. - Proportion of students with a positive attitude about statistics. - Mean number of points beyond the minimum requirement earned by students.	- Proportion of students voluntarily taking statistics-related classes in future semesters. - Proportion of students that experience an increased willingness to try in future college classes. - Proportion of students that experience a decrease in fear of failure during the learning process.

2.3.3 Evaluation Questions

Some formative evaluation questions addressed in this report about the 3 Unit Tests that could be unlocked/retaken using points earned from other activities (found in the “Activities” section of the Logic Model, in Figure 1) are:

- How difficult are students finding the test material?
- How many students require a retake of the test?
- Is the ease/difficulty students are experiencing a result of the test questions themselves, or a result of the program’s preparation (or the students’ lack of preparation) for the taking of those tests?

The main summative evaluation question addressed about the program stems from the "Outcomes" section of Table 2, which is the mean number of points *beyond* the minimum requirement earned by students:

- To what extent does this program increase students' willingness to try, and motivation to learn?

2.3.4 Data Collection

Unidentifiable records were obtained from the University that show students' assignment submission patterns and test grades. Voluntary anonymous surveys were sent to students at the beginning, the middle, and the end of the course to gauge students' understanding of and experience with the novel grading structure. Members of the Institutional Review Board determined that the data gathered for this study were exempt from review under exemption number 4, which exemption cites Family Educational Rights and Privacy Act (FERPA) law allowing for the use of educational records for research purposes on behalf of educational institutions to administer student aid programs or to improve instruction.

2.4 Discussion

2.4.1 Quantitative Results

From the University records on assignment submission patterns and test grades, very encouraging evidence was found that allowing for multiple avenues and opportunities to practice the material had a positive impact on students' grades.

In Figure 1, it is shown that the total number of "stars" earned by a student on assignments, quizzes, tasks, and challenge prompts combined had a 0.67 correlation with the student's final grade. Visually, it should be noted that this correlation is not higher than 0.67 because there are several students that don't need to see the material in as many different ways in order to master it – hence much of the upper-left area of the plot is filled. However, it should also be noted that the lower-right area of the plot is virtually empty, revealing that those students who needed to see more types of prompts and situations for the material in order to understand it were able to take advantage of the many different types of assignments offered to help them gain the desired understanding.

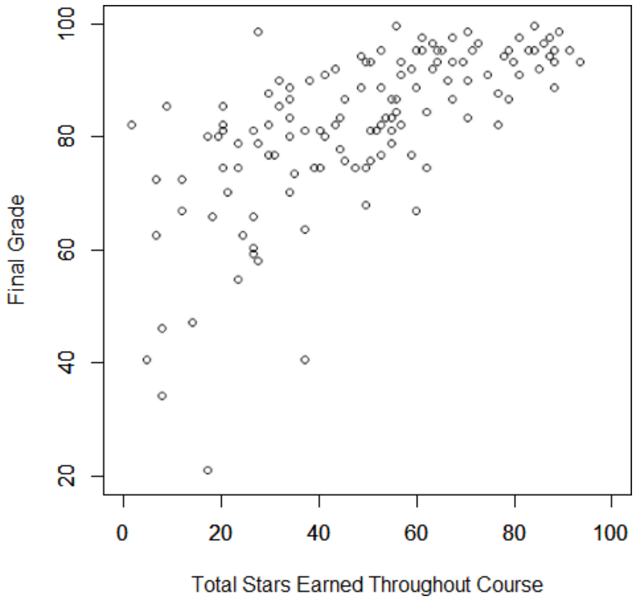


Figure 1: Stars Earned by Students and Their Final Grade

Figure 2 tells a very similar story, though the correlation is much lower. In Figure 2, there is depicted a 0.30 correlation between the total number of submissions students made of assignments, tasks, quizzes, and challenge prompts combined and their final grades. The filled upper-left area and the empty lower-right area show that while many students needed only make a few submissions in order to gain sufficient understanding of the material, but those students who needed more practice (even in the hundreds of submissions) were able to take advantage of the opportunity to do so to achieve the needed understanding.

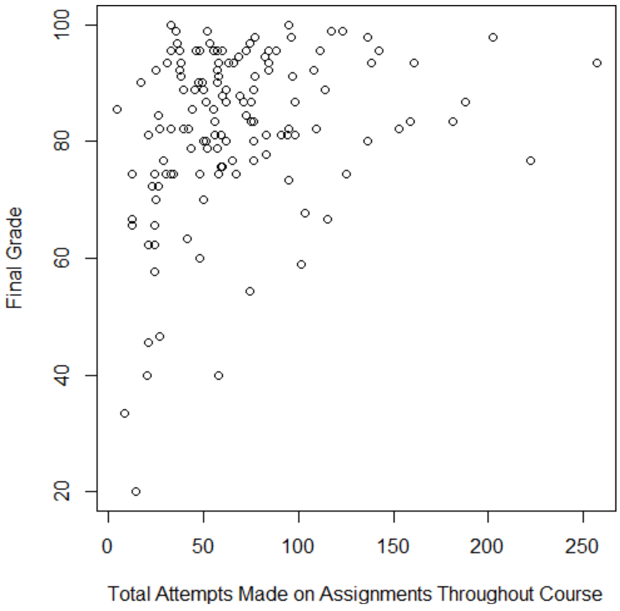


Figure 2: Assignment Attempts Made by Students and Their Final Grade

The reason the researchers equivocate final grades with “understanding” in Figures 1 and 2 is because neither the number of “stars” nor the number of attempts students make on assignments, quizzes, tasks, or challenge prompts have any numerical influence on the final grade. The final grade was purely composed of unit test scores, while any stars or attempts on non-test items were purely for students to prepare for the tests.

The data also show that while the point in the semester when a student submitted their first test attempt was correlated with their test score (those who procrastinated their first test attempt were much more likely to earn a poor ending test score), the point in the semester when a student turned in non-test material had almost no correlation with the material’s corresponding test score. This shows that while assignment due dates can be helpful for guiding student pacing and streamlining grading processes, forcing students to follow rigid deadlines for all assignments may not likely have a meaningful positive impact on the students’ ultimate understanding of the material.

Nearly identical tests were administered to students in the semester preceding the one in which this study took place. Scores between these two semesters may not be very comparable due to the COVID-19 shutdown halfway through the Spring 2020 semester, and the online structure of the Fall 2020 semester. However, we made some basic quantitative comparisons. The mean test scores were nearly a full percentage point higher with the gamified grading semester than in the previous semester, with a 69% chance of seeing this difference by chance. The proportion of students in the gamified grading semester who received an overall A on their test scores (a mean of 92% or greater) was 0.32, whereas the proportion in the previous semester with an overall A on test scores was 0.23, with a 12% chance of seeing this difference by chance.

2.4.2 Student Response

Using survey data, information was gathered on how students felt about the new grading scheme throughout the course. It is important to note that students notoriously, as a whole, are not very good at recognizing what types of class structure or assignments will help them learn best (Bowman & Seifert, 2011) or how well they have learned course material (Lew et al., 2010). However, the researchers still felt it important to know how students were feeling in order to find any major gaps in the system or any pain points that we could ease in future iterations of the grading structure.

Survey data show that, at the beginning of the semester, most of the students agreed that the gamified grading structure would help them be successful in the course, with only 6-10% of students not initially liking the new structure. However, as the semester went on, about 14-19% of students began to feel less confident that the grading structure would help them be successful. The majority of the reasons given for this were that they found they were likely to procrastinate their homework without a hard deadline, that they weren't getting enough feedback on their assignments and tests, and that they didn't like that they didn't get any final-grade-credit for non-test items they submitted.

From the survey data, it is estimated that 59-67% of the students liked the grading scheme from the beginning to the end of the course, and that another 12-18% of the students did not like the scheme originally but eventually came to like it. The majority of reasons given for this were that it felt like students got credit for trying, they studied harder than they would have in a traditional course, and that they really appreciated the chance to redo assignments and tests as many times as they needed.

The final survey given at the end of the semester had only a 37.8% response rate, with most responses being from students who either didn't like the grading structure from the very beginning or stopped liking it halfway through the semester. However, even though the majority of responders to the final survey didn't like the grading structure as a whole, the following statistics from this survey are of interest:

- 75.5% found the ability to resubmit assignments, quizzes, tasks, and challenge prompts beneficial.
- 66.0% found the flexible test dates beneficial.
- 49.1% found the flexible assignment dates beneficial.
- 54.7% found the ability to choose between assignments, quizzes, tasks, and challenge prompts beneficial.
- 66.0% found the ability to retake tests beneficial.
- 65.4% liked not needing to complete all assignments in order to fulfill course expectations.
- 78.9% thought that resubmitting/reviewing assignments was a good way to prepare for tests.

- 51.9% felt that the grading scheme helped them move past failures and try again.
- 46.2% thought the grading scheme helped students focus more on understanding the material than on getting good grades.

These are surprisingly positive responses from a group of students who primarily did not like the grading structure. At the end of the course, the primary reason given for not liking the grading structure was that test scores were the only thing that influenced their final grade.

When asked about certain possible changes to future gamified grading schemes, students who responded to the final survey answered as follows:

- More than half the students agreed that enforcing a two-week test window would have a positive impact on the grading structure.
- There was not a consensus as to whether enforcing a two-week assignment submission window would have a positive or negative impact on the grading structure.
- There was not a consensus as to whether students felt they would have done better or worse in the course if it had used a more traditional grading scheme.

About half the students felt that due date strictness was perfect, while the other half would have preferred more strictness in due dates and penalties.

2.4.3 Evaluation Results

While the data seem to suggest an overall positive response to the gamified grading structure, these feelings were not as wide-spread across the students as was expected, especially since course data suggest that the gamified grading structure accomplished much in the way of removing barriers to student success.

Most students who didn’t like the grading structure were bothered that tests were the only thing that was reflected in their final grade. While the researchers do not believe that formative assessments such as assignments and quizzes should compose too much of a final grade, which is meant to be a summative assessment, perhaps having assignments contribute to a very small portion of the final grade would alleviate this consistent concern from students.

It is difficult to tell how the overall spread of final grades was affected by this new grading structure, because this program was tested in the Fall 2020 semester, which was the first semester students were forced to do primarily online and distance learning due to the COVID-19 pandemic. The researchers are interested to see how this gamified grading program fares in a more traditional face-to-face environment where students have more ample opportunity to develop a personal relationship with their teachers and teaching assistants.

With the data gathered on students’ experience with this new grading structure, the following comments can be made on each of the evaluation questions posed earlier in this paper.

How difficult are students finding the test material? Though the materials presented on the tests were considered fairly difficult, typical scores on each test ranged from 80-90%. However, about 25% of students scored below a 75% on at least one unit test. Those who procrastinated their first attempt on a test were more likely to find the test more difficult, and less likely to have the time to attempt the test again. There may be some merit in enforcing a small submission window for the first submission of each unit test to keep students from procrastinating their first attempt, and to ensure all students have time for one or more additional attempts if they are needed.

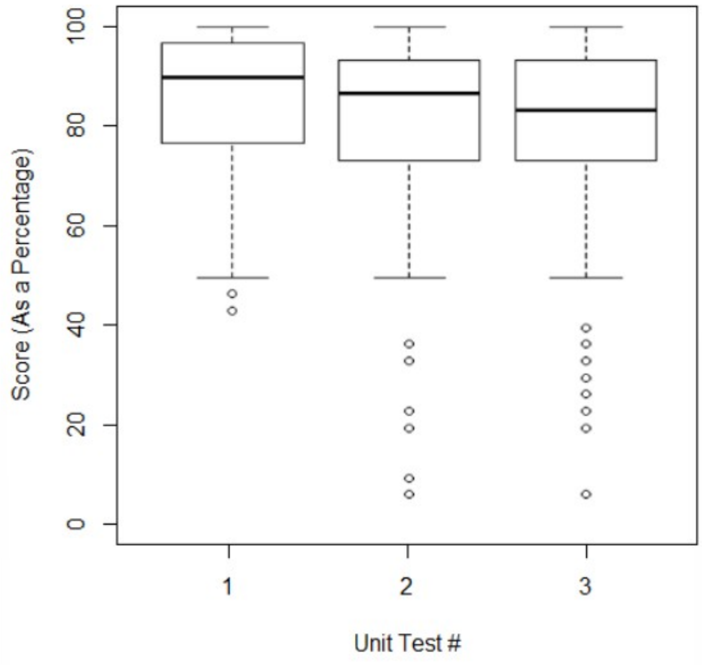


Figure 3: Student Unit Test Scores

How many students require a retake of the test? 40% of students in the class made a second attempt at one or more of the unit tests, even though it took earning more "stars" in order to unlock those retakes.

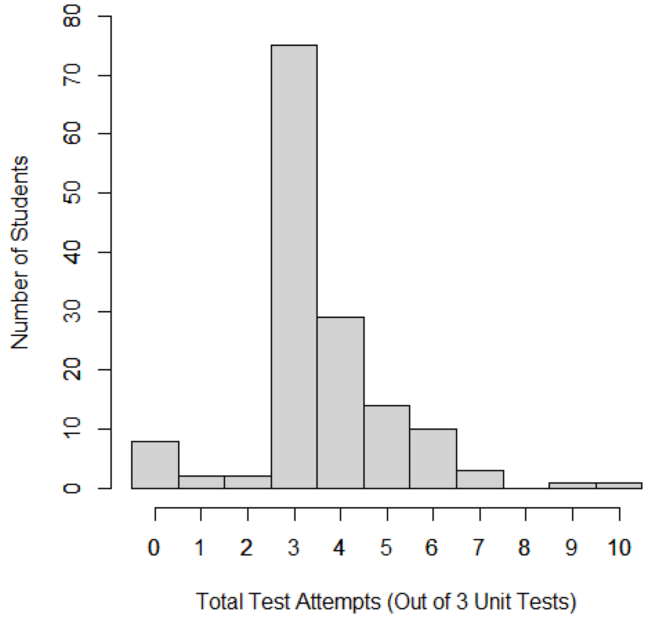


Figure 4: Student Total Test Attempts

Is the ease/difficulty students are experiencing a result of the test questions themselves, or a result of the program's preparation (or the students' lack of preparation) for the taking of those tests? Referring to Figures 1 and 2, the more that students used the program's preparation opportunities, the higher their minimum test scores were. Those who found the test material difficult appear to be the students who did not put in the necessary preparation. Instructors may be able to help these students by finding more effective ways to help them see the value of preparation for unit tests.

To what extent does this program increase students' willingness to try, and motivation to learn? Many students appear to have focused more on learning the material than on simply getting good grades on assignments, and that many students who normally may have had to settle for low grades on their initial attempts seized the opportunity to try again and again (even to making up to 10 attempts on unit tests) to achieve the scores they desired in the course.

2.4.4 Future Implementations

2.4.4.1 Analysis of possible adjustments

Using Gilbert's Performance Matrix (Gilbert, 2007), the researchers explored reasons why there were not as noticeable results of better grades and attitudes from the students had been expected. This analysis reveals some adjustments that can be made to the experience in a way that will help students catch the vision and implement the tools given to them well in order to succeed in the class. The portion of Gilbert's performance matrix that Gilbert most often focused on, which are stages III, IV, and V, as highlighted in Table 3 (MacDonald & Reardon), were used.

Table 3: Gilbert's Performance Matrix, with the portion used in this dissertation highlighted

Stages	Models	Measures	Methods
I. Philosophical Level	Ideals	Integrity	Commitment
II. Cultural Level	Goals	Conformity	Policy
III. Policy Level	Missions	Worth	Programs
IV. Strategic Level	Responsibilities	Value	Strategies
V. Tactical Level	Duties	Cost	Tools
VI. Logistical Level	Schedules	Material Needs	Supplies

At the “Policy” level, the “Missions” for the participants are to get an A, and to learn the material thoroughly. The “Worth” of these missions is measured by a juxtaposition of how important it is for them to pass statistics with good understanding and high marks, and how much of a sacrifice it is for them to do 6-9 hours of work every week. Some possible “Programs” to enhance students’ ability to accomplish these missions and perceive the worth of them are (1.1) to put a greater emphasis on the value of the material, (1.2) to make adjustments to the design of assignments, lectures, and due dates to give as much flexibility for when to work on the material as possible, (1.3) to build opportunities into assignments for students to accomplish other personal tasks while completing the assignment, and (1.4) brainstorm policy changes that make a passing grade accessible to all students, even at the end of the semester.

At the “Strategy” level, the “Responsibilities” of the participants in order to fulfill the missions are to stay ahead of their assigned work, actively engage with their instructors, and to redo unsatisfactory work in a timely manner. The “Value” gained when these responsibilities are fulfilled are a good grade, useful knowledge, a sense of accomplishment, and progression in their program. The value lost when the responsibilities go unfulfilled result in discouragement, not being able to apply the material at crucial future moments, and possibly having to retake the course. Some "Strategies" to enable students to gain the value that come from engaging in these responsibilities are (2.1) to make shorter assignments, (2.2) automate as much of the grading as possible to provide immediate feedback, (2.3) create tasks that have students apply material in other relevant personal settings, and (2.4) free up instructors' time so they can do more tutoring and reaching out to students one-on-one.

At the "Tactics" level, the "Duties" of the participants in order to fulfill their responsibilities are to attend lectures and recitations, start their assignments early, to proactively ask questions, and to focus on the material rather than the grade. The "Cost" of completing these duties include six to nine hours a week outside of class (≈160 hours total), conjuring up self-motivation rather than being motivated by deadlines, the energy and concentration to think deeply during class time rather than be a passive receiver of information, and to be willing to make fast and frequent "failures" during the learning process. Some "Tools" that can help students fulfill these duties while minimizing costs are (3.1) faster return of grades and feedback, (3.2) easy-to-follow, clear instructions for applying material to other current jobs or classes, (3.3) a clear and detailed map of what will take place throughout the course, and (3.4) brief, low-stakes opportunities for students to practice the material.

The programs, strategies, and tools above can be organized into each of a few broad categories, as shown in Table 4.

Table 4: Categories of possible useful interventions for the gamified grading structure, with programs, strategies, and tools labeled according to numbers in section on the "analysis of possible adjustments" of this article

Intervention Type	Programs, Strategies, or Tools			
Alterations to Course Structure and Policies	1.2	1.4	2.4	3.4
Adjustments to Assignments	1.3	2.1	2.3	3.2
Changes to Grading or Instructor Feedback	2.2	3.1		
Modifications to Class Introduction	1.1	3.3		

2.4.4.2 Structural design intervention

The analysis above using the performance matrix is meant to help behaviorists determine which type of human performance intervention to apply. The researchers determined that the best intervention to move forward with first is a “Structural Design” intervention, in which some alterations would be made to the structure of the course utilizing technology. In particular, this structural design intervention will be to convert all assignments into shorter online assignments that can be graded by an online grading system. Students will be able to complete their work in smaller chunks, they will be able to receive feedback on their work immediately, and instructors will have more time to work one-on-one with students since they will spend much less of their time grading work.

This intervention will allow students to remain more accountable and ahead of the work they need to do in the course, and will provide a larger amount of human interaction needed to help students understand the material deeply.

2.4.4.3 Awareness campaign

Finally, providing an awareness campaign to help students see the reasons for, and benefits of, the change to a gamified grading structure should have an impact on those students who remained less motivated to utilize the new grading structure from the beginning. With reference to Jeffrey M. Hiatt’s awareness campaign guidelines

(Hiatt, 2006), the following campaign was developed to help students become converted to the new system:

- Give reasons for the change.
    - Give a short (10-20 minute) presentation on the first day of class in large lecture, about research on video games and how certain elements of those games help players feel motivated to continue trying at something difficult even in the face of persistent failure.
    - Show students the TED talk "The Super Mario Effect" by Mark Rober (2018) during their first recitation before covering the course syllabus.
- Explain why the change is being made in this way.
    - On the first day of class, present the elements of traditional grading systems that have been shown to psychologically drain students of motivation as the course progresses. Present how certain gamification techniques can remove that motivation drain to a significant degree.
    - During the second recitation, recitation leaders facilitate a short (five to seven minute) activity that gets students comfortable with "failing" or being wrong in front of each other. Then, they briefly describe the research that shows that "failing" at something contributes more to accurate long-term retention on information than succeeding does.
- Communicate excitement for the change.
    - During the first several lectures, the instructor will find some way to bring up the new grading structure and how excited they are that

    students will have less obstacles between them and the grade they want in this difficult course.
    - In each recitation, recitation leaders consistently encourage students to try assignments early so they can "fail" quickly at material they do not yet fully understand, so they can get helpful feedback on how they can improve before upcoming due dates.
- Clarify the details of the change.
    - Map out the changes to the grading structure in great detail in the course syllabus.
    - During the first recitation, recitation leaders review the syllabus with students to facilitate any questions they may have about the new structure.
    - During the first week, assign an online quiz for a small amount of points where students can show that they grasp each of the major details of the new grading structure and how to take advantage of it.

2.5 Conclusion

The gamified grading structure implemented here did not show strong evidence of *creating* motivation in students to learn, but it did show evidence of *removing barriers* to motivation that many students have been experiencing, in part, due to traditional college grading schemes. (Research Question 1)

The majority of students who experienced the gamified grading structure had a positive experience, explaining that they weren't as afraid to experience failure in the classroom as they learned new material. (Research Question 2)

Differentiating between formative and summative assessments by how they contributed to the final grade appears to have helped students better utilize formative assessments as learning tools rather than graded activities. (Research Question 3)

The small group of students who did not like the grading structure primarily disagreed with having three summative assessments as the only contributors to the final grade. Future iterations of the gamified grading structure may include an awareness campaign for the change in the grading plan, shorter online assignments with immediate feedback, formative assessments contributing to a small portion of the final grade, and short submission windows for the first attempt at each of the unit tests. (Research Question 4)

CHAPTER III – PROJECT 2 – A QUALITATIVE ANALYSIS OF THE UTILITY OF "YOUR AVERAGE GAME" AS A TOOL FOR INTRODUCTORY STATISTICS STUDENTS TO CONSTRUCT AND EXPLORE THE STEPS OF A HYPOTHESIS TEST FOR A MEAN

3.1 Introduction

3.1.1 Background and Context

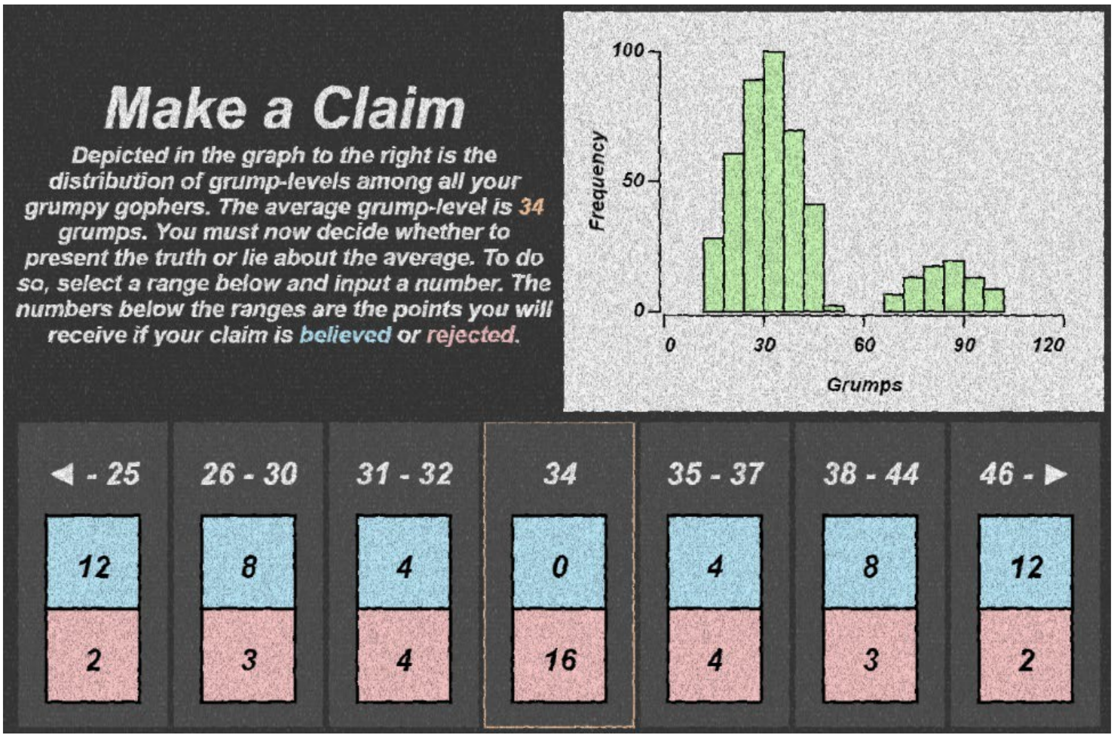
In this project, the researchers explored the results of having introductory statistics students play a game called "Your Average Game" which has been developed by Partridge. Several undergraduate participants played the game while the researchers examined what types of statistical methods and strategies the participants naturally came up with as they played the game, if any.

Figure 5: Two Game Options on Site Homepage

"Your Average Game" can be found at http://tinyurl.com/YourAverageGame at the time of writing. It is designed to be played between two or more players. There are two different versions of the game, called "Your Average Lying Game" and "Your Average Guessing Game."

The lying game puts players in a position to conduct informal hypothesis tests for a mean, and the guessing game puts players in a position to create informal confidence intervals for a mean.

This project focuses on “Your Average Lying Game,” and leaves the guessing game for future research. Throughout this paper, when “Your Average Game” is referred to, it is a reference to the lying version of the game.



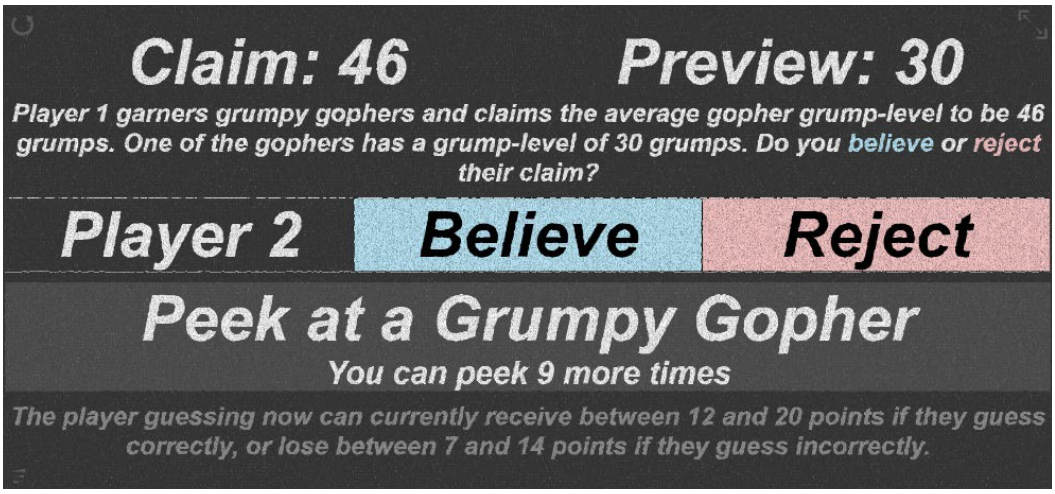
Make a Claim

Depicted in the graph to the right is the distribution of grump-levels among all your grumpy gophers. The average grump-level is 34 grumps. You must now decide whether to present the truth or lie about the average. To do so, select a range below and input a number. The numbers below the ranges are the points you will receive if your claim is believed or rejected.

◀ - 25	26 - 30	31 - 32	34	35 - 37	38 - 44	46 - ▶
12	8	4	0	4	8	12
2	3	4	16	4	3	2

Figure 6: Player 1's Screen When Playing as the Poser

Players take turns posing a claim about a dataset while the other players try to decide whether they think the claim is true or false. The player making the claim is given a fun scenario for a random variable with a given average value. The player then makes a claim to the other players about the average value of the random variable.



Player 1 garners grumpy gophers and claims the average gopher grump-level to be 46 grumps. One of the gophers has a grump-level of 30 grumps. Do you believe or reject their claim?

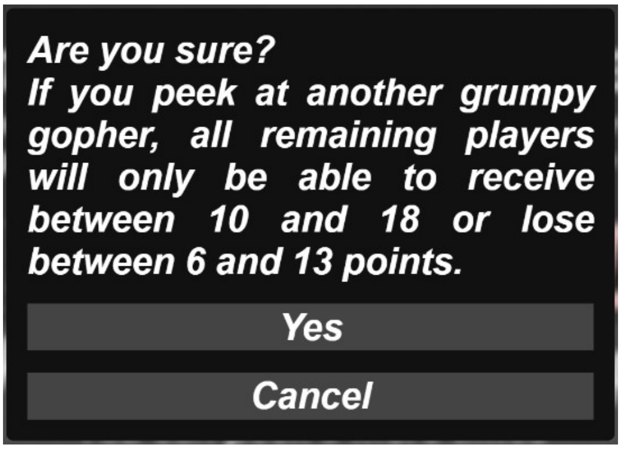
You can peek 9 more times

The player guessing now can currently receive between 12 and 20 points if they guess correctly, or lose between 7 and 14 points if they guess incorrectly.

Figure 7: Player 2's Screen When Playing as the Peeker

The other players may then randomly select observations from the population one by one, and then decide whether they believe the claim or not.

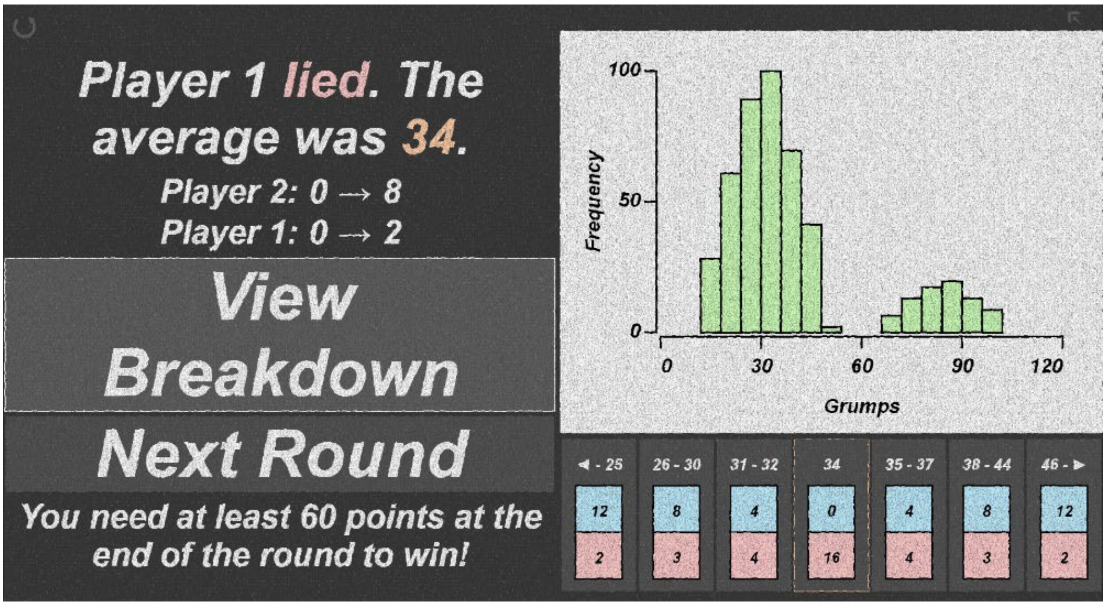
Points are awarded in such a way that the player making the claim must choose to tell the truth for a moderate amount of points and hope the others think it is a lie, or to tell a small lie for a small number of points or a big lie for a large number of points and hope the others think it is the truth. If the others are fooled, the player making the claim is awarded points.



Are you sure?
If you peek at another grumpy gopher, all remaining players will only be able to receive between 10 and 18 or lose between 6 and 13 points.

Figure 8: Warning of Decreasing Points When Choosing to Peek

The other players hope to determine correctly whether the claim is true or false, but the number of points they can receive decreases with every observation they select from the population, which encourages them to take risks and make decisions about the claim before they've obtained a lot of information.



Player 1 lied. The average was 34.

Player 2: 0 → 8

Player 1: 0 → 2

You need at least 60 points at the end of the round to win!

Figure 9: Results Screen After Peekers Have Made Their Decisions

The first player to reach a predetermined number of points after a full round wins. The full rules as detailed on the game's website, along with an example playthrough of a round of gameplay, are provided in the Appendix.

3.1.2 Problem Statement

Students who are learning how to conduct basic formal hypothesis tests in the classroom make a great number of common mistakes in the process. At the procedural level, many of these mistakes can be corrected through practice and memorization (Evangelista & Hemenway, 2002). At the conceptual level, students can find it difficult to connect the procedures together in a way that logically makes sense to them (Glaser, 2003). Many instructors and researchers have observed that even a great deal of emphasis on the logical reasoning behind the steps of a hypothesis test is sometimes not enough to help students overcome the impression that these steps are a complicated mathematical process that must be memorized for lack of simplicity or straightforwardness (Evangelista & Hemenway, 2002).

Smith (2008) and Sotos et al. (2009) showed several years ago that there was a lack of literature exploring what can be done about helping students learn and understand the logical procedure of a hypothesis test. Since that time, we are still searching for better ways to provide students an environment where they can more naturally make these connections.

"Your Average Game" has been designed as a casual, low-stakes way for players to informally construct the basic steps of a hypothesis test for a population mean in their own way and through their own logic. Evidence that this game helps players come up with the process on their own would suggest that using the game in the statistics classroom may make it easier to help students see in a statistics course that the formal steps they are learning are, in fact, the natural and straightforward way to go about answering the question at hand.

3.1.3 Research Questions

The questions addressed in this project are as follows:

Research Question 1: Concerning players of early college age who have not yet learned about hypothesis testing in a statistics course, what strategies and procedures do they come up with as they play against an opponent in "Your Average Game"?

Research Question 2: What similarities, if any, do these strategies and procedures have with the formal steps of a statistical hypothesis test of significance for a population mean?

3.1.4 Relevance and Importance of the Research

While there are a large number of researched games that can be found for teaching elements of algebra, geometry, or even calculus, there are very few researched games for teaching statistical concepts outside of probability. When it comes to games for learning hypothesis testing methods, the size of the pool of researched games reaches almost zero. Most published papers on games that teach "hypothesis testing" are actually historical, social, or logic-based puzzles where students will make guesses about the correct answer or what will happen in a science experiment, and then test to see how valid their guess is. Such games have been described by Adams et al. (2012), Rickard & Titley (1988), and Maloney & Masters (2010). These games are not actually grounded in the formal steps of a hypothesis test for a population parameter as taught in introductory statistics courses. There is one study that showed the effectiveness of a gamified approach to teaching each individual step of the statistical hypothesis testing process (Delgado-Gómez et al., 2020), but it was not an actual game to be played as much as it was an approach to comprehending each step of the process.

Thus, research on a game that focuses on helping students develop a concept of the logic and reasoning behind the steps of a statistical hypothesis test would be one of the first of its kind, and could serve as a precedent for future games and studies to be developed for assisting students in their understanding of these and other statistical methods.

3.2 Literature Review

3.2.1 Key Concepts, Theories and Studies

The vast majority of studies and papers that have been written on the subject of using games as part of a classroom curriculum show arguments and evidence that games engage students with the material in a way that a simple lecture, assignment, or video simply cannot. Such research and conclusions have been shown by Squire (2005) in world history courses, Sugar (1999) in history, math, biology, and business training, Chou (2019) in sociology and public policy, Gardner & Hatch (1989) in elementary education, Armstrong (2000) in psychology, Langran & Purcell (1994) in English language learning, and Afari et al. (2012) in mathematics.

Kurt Squire, a Professor of Informatics at UC Irvine, made the case using research done in world history courses that the completion rate for online courses barely reach 50%, but gamers will not only spend hundreds of hours mastering a game, but will write extensive papers about their strategies and even set up virtual “universities” in order to teach others how to play those games. “In short, while e-learning has a reputation for being dull and ineffective, games have developed a reputation for being fun, engaging, and immersive, requiring deep thinking and complex problem solving” (Squire, 2005).

Steve Sugar is the author of five performance measurement game systems that are actively used in the U.S. and over 20 other countries, which written games and techniques were developed through studies across many different education subjects such as history, math, and biology, before Sugar focused the research on using games for training employees of all business types. He claimed that games in the classroom are a natural product of how the focus of teaching has changed over the years. The general public used to imagine classroom learning as students sitting at desks listening to lectures and copying down information. Today, they expect learners to be much more active in the classroom. Introducing games into the curriculum is a natural way to encourage students to actively participate in the material. (Sugar, 1999)

While Yu-kai Chou (2019), whose research is focused on gamification on a larger scale for social change, posited that humans have eight core motives behind everything they do, and that gamification allows us to tap into many of those motives, Gardner & Hatch (1989) theorized in their research on the abilities and intelligences of four- and five-year-old children that humans have eight “intelligences,” and that each topic in the classroom should be approached in at least six different ways in order to tap into most of these intelligences throughout the learning process. One of these ways is “the personal way”, where one aims to see if something is possible through role play or personal interactions with others, which well-designed games can easily provide.

Thomas Armstrong, Executive Director of the American Institute for Learning and Human Development, made the case in his psychological research that spanned hundreds of classrooms and subjects across the world that using games as a teaching strategy allows for an excellent low-stakes setting for students to interact with each other while learning the material (Armstrong, 2000). Langran & Purcell (1994) agreed as a result of their research on teaching a second language to adults, stating that games are a great way to encourage the more introverted students, or those with low confidence, to speak in front of a smaller audience in an atmosphere that is not so serious or as risky as a group project or a presentation.

There are more articles connecting teaching theory to the use of games in the classroom than there are formal quantitative studies on the effects of using games in the classroom. However, there are several such studies that show quantitatively that games are likely a positive inclusion in a classroom curriculum. For example, Afari et al.

(2012) demonstrated in their research on using games for teaching tertiary-level mathematics in the classroom that student perceptions of a class or of the material being taught were statistically significantly more positive after playing relevant games for learning. In particular, students felt more strongly that the teacher supported them and was invested in their education, more involved in the class, that the material was more relevant to them, that the mathematics was more enjoyable to learn, and they felt greater confidence in their ability to learn it.

3.2.2 Key Debates and Controversies

While it is difficult to find many naysayers to the use of games in the classroom, there are a few. Lee Su Kim, associate professor at the School of Language Studies and Linguistics in Universiti Kebangsaan Malaysia, and president of the Peranakan Baba Nyonya Association of Kuala Lumpur, found that in many countries there is a common perception that learning should be rigorous and formal by nature, and that if one is having fun then it is not really learning (Lee, 1995).

Older arguments from professionals such as Caillois (1957) regarded games as unpredictable ways of learning, as the instructor will be uncertain about what students will actually take away from the experience. Additionally, arguments were made that games usually have arbitrary rules and fictitious settings that do not relay well to the "real world," and thus are unproductive activities for learners to engage in. However, so many elements of education have changed dramatically since the 1950s, these arguments are more a reflection of a different era than representation of opinions held by today's researchers in education.

Most published arguments are not so much aimed at removing games from the classroom, but in making sure they are being used appropriately. For instance, one researcher stated that while using games targeted at certain learning objectives may be a useful tool, instructors should not be revolutionizing their curriculum to ensure games are the main source of instruction (Mayer, 2016). Others warned that games "should not be regarded as a marginal activity filling in odd moments when the teacher and class have nothing better to do," (Stojković & Jerotijević, 2011) but rather should be used carefully and deliberately in the classroom if they are used at all.

3.2.3 Gaps in Existing Knowledge

Many researchers have found that well-designed games are a demonstrably effective way to help learners become more engaged in the material being taught. The largest "gap" in the current knowledge is a lack of unique learning games that have been shown as efficient methods of addressing specific learning objectives in mathematics and statistics classrooms.

Suat Khoh Lim-Teo, who has done more targeted research on the subject of using games specifically for teaching mathematics, has offered a possible classification for the six types of games that can be used in a mathematics classroom: (1) games for drill and practice, (2) games for concept reinforcement, (3) games which lead to concept formation, (4) games which lead to mathematical investigations, (5) games which apply mathematical knowledge, and (6) games for fun. She remarked that games for drill and practice tend to be the most-used because they require the least amount of attention from the instructor and they are the easiest to create. However, since these games do not foster the use of critical thinking skills or actively seeking out winning strategies, learners soon recognize them as mere disguises for drill and practice and eventually lose interest. However, types 2, 3, 4, and 5, while much harder to create effectively, are the most influential at encouraging student engagement and understanding. (Lim-Teo, 1991)

3.3 Research Design

3.3.1 Participants and Setting

The participants for this study were freshman and junior college students who had begun but not yet completed an introductory statistics course. The 150 students enrolled in various sections of a lower-level introductory statistics course and of a higher-level introductory statistics course at Utah State University in the summer semester of 2023 were informed of the opportunity to participate in the study, with a small amount of extra credit provided by their professor for filling out a brief survey and volunteering to be a part of the study. The lower-level course generally has students who are either not as comfortable with math or who have declared majors with minimal math requirements, and the course is designed to cover statistical methods in a very approachable way to the general public. The higher-level course generally has students who are pursuing degrees in fields such as engineering, mathematics, or biology, and the course is designed to introduce statistical methods along with the underlying mathematics and calculus behind them.

Of the fifty-six students (thirty-six lower-level, twenty higher-level) that volunteered, thirty-six (twenty-three lower-level, thirteen higher-level) had not had previous exposure to hypothesis testing in college or high school math or statistics courses. Eight students (four lower-level, four higher-level) were randomly selected from these thirty-six to play the game in a research interview setting with a researcher. Seven of the eight (three lower-level, four higher-level) followed through with the interview. Of those that were interviewed, two were women and five were men. The other forty-nine volunteers (thirty-three lower-level, sixteen higher-level) were invited to play the game at home, and then to answer a few survey questions about their experience afterward. Fourteen of the forty-nine filled out the survey (ten lower-level, four higher-level). Of the survey respondents, nine were women and five were men.

3.3.2 Research Interviews

The seven students who participated in the research interviews met in a neutral room on the Utah State University campus where they were briefly introduced to the researcher and then to the game. Participants were encouraged to bring their own opponent so they could play against someone they were comfortable with, but none of the participants chose to do so. Instead, one selected member of the research team played as the opponent for each participant. This worked in the project's favor, as this opponent was able to play with a consistent strategy and demeanor across every round and with every participant.

Each participant played the game with their opponent for about 45 minutes, while the researcher asked them questions about what they were thinking, why they were making each of their choices, and what kinds of strategies they thought would help them perform better during the game.

During the game, the participant regularly took on two different roles: that of the "Poser," and that of the "Peeker." While playing as the Poser, the participant was given the average value of a large numeric data set, as well as a histogram depicting the distribution of the data. They then decided whether to present the true average to their opponent and hope the opponent would not believe them, or to present a lie as the true average and hope the opponent would believe them. The Poser only stands to gain points during the round. They will gain a minimal number of points if the Peeker accurately determines whether they told the truth or a lie, or they will gain a larger number of points if they deceive their opponent. Most questions the researcher asked the participant during this portion of the game were patterned after the following:

- Do you think it's better to tell the truth or to lie this time? Why?
- Why did you choose that range for your claim?
- What on-screen elements are you considering as you make your decision?
- How do you hope your opponent will be deceived by your claim?
- How likely do you think it is that this claim will deceive your opponent?
- Are you approaching this differently than you have before? Why/How?
- What do you think is the best strategy to follow during this phase of the game? Why?
- Do you think you made the best decision even though you didn't deceive your opponent? Why?

While playing as the Peeker, the participant is given a claimed average of a population by their opponent, as well as the true value of a single randomly selected subject in that population. The participant may choose to believe or reject the claim as the true average of the population, or they may choose to "peek" at another randomly selected subject in the population. They may peek up to ten times during any given round. The Peeker stands to gain points if they correctly determine whether their opponent told the truth or a lie, or to lose points if they are incorrect. At the beginning of each round, the Peeker stands to gain or lose a large number of points, but with each peek the number of points to be gained or lost decreases. Most questions the researcher asked the participant during this portion of the game were patterned after the following:

- Do you think your opponent is lying or telling the truth this time? Why?
- Do you think it's worth it to get more information before making your decision? Why?
- Why are you choosing to make your decision now instead of after another peek?
- Why are you choosing to peek again before making your decision?

- How likely do you think it is that you're about to make the wrong decision?
- Are you approaching this differently than you have before? Why/How?
- What do you think is the best strategy to follow during this phase of the game? Why?
- Do you think you made the best decision even though your decision was incorrect? Why?

Audio recordings of each participant's session, as well as a screen recording of gameplay, were made as the researcher had the participant think aloud during the game. Zoom was used to record the interview audio and the gameplay on the screen, as well as to automatically create a rudimentary transcript of each interview. The researchers then reviewed the Zoom transcripts line by line and made edits to ensure transcription accuracy.

3.3.3 Asynchronous Surveys

Fourteen students participated in the asynchronous portion of this study. These students were given a link to "Your Average Game" and invited to play the game against another college-age opponent of their choice. After playing the game through to the end, they were invited to fill out a survey that asked some questions about their beliefs or strategies about the game. The questions posed in the survey were as follows:

- Describe the basic strategy you developed to have the best chance at gaining the most points while trying to determine whether the Poser was telling the truth.
- Explain how peeking is helpful for determining whether the Poser is telling the truth.
- When you were the Poser, did you use the histogram provided to inform your decision about what claim to make about the average? If so, how? If not, why not?

- What is the smallest number of peeks you think a Peeker could make and still feel fairly confident they will make the right decision? Why?
- Do you believe there may be a strategy you could develop where you could determine whether the Poser told the truth correctly every time? Why or why not?
- Were there any times, as a Peeker, where you believe you made the best decision based on the information you had, even though you got it wrong? Why or why not?

3.3.4 Practical Considerations

Due to the time constraints of the project, it was not possible to include more than 8 participants in the research interviews. However, for a qualitative study such as this one, this is a reasonably high number of participants.

The researchers did not foresee any ethical problems to consider, as participants would merely be playing a game with someone else for under an hour, with no real risk of distress or serious embarrassment. However, participants were informed that they could terminate the interview at any time if they felt uncomfortable, and had the right to request that any data gathered during the interview be deleted if needed. All participants who started the interview followed through until the end with no concerns about their experience or the data collected.

All participants read and signed an electronic informed consent form when volunteering for the study. It was not necessary to keep personal information of the participants in order to analyze the data. Subject anonymity and confidentiality were kept throughout the interview process, as they were not asked to state their name during the recording, and the video remained a simple screen recording of the game as it was played. All interview and asynchronous participants provided their names for extra credit on separate survey pages which were not kept after the instructors were informed about who participated.

One anticipated obstacle was that participants may not have been as forthcoming to explain their thought processes or strategies while they were sitting across from the opponent they were trying to beat in the game, as the opponent could then use anything the participant said to their own advantage, allowing the opponent to anticipate the participant's next moves and react accordingly. To overcome this, a large pair of noise-cancelling headphones was provided for the opponent to wear, connected to their choice of music, whenever the researchers paused and ask the participant brief questions so the participant could talk freely without feeling that they were giving their opponent an advantage over them.

Since professors gave a small amount of extra credit to students who participated in this study, there was the issue of fairness, as these extra credit points should be available to all students, and not just to those who end up participating in the study. For this reason, a pre-survey was provided where students gave some limited information about themselves and about their statistical understanding, as well as set up a time to participate in the face-to-face part of the study if they were selected. They could choose to volunteer to fully participate if chosen to be a part of the research interviews, or to volunteer only as an asynchronous survey participant, or to participate in a way that did not contribute to the study but that could still earn them extra credit in their class. Those who wanted extra credit but did not want to contribute to the study were given the same link to the game and survey questions as those in the asynchronous portion of the study, but were told to submit their survey answers to their instructor rather than to the researchers. Instructors gave the extra credit to any student who participated in an interview, or who filled out a survey and sent it either to the research team or to their instructor.

It is important to note that only a third of the students who were presented with the opportunity to participate in this study volunteered, and that less than half of those volunteers ended up choosing to follow through with participation. This presents the high likelihood of non-response bias in the study results. However, most participants who attended the interviews mentioned really needing the extra credit in their course, so the data likely reflect how "Your Average Game" is viewed and approached by those who are not confident in their understanding of basic statistical concepts. As such, any valuable insights into the process of hypothesis testing that participants showed or developed during the game are very encouraging as to the utility of "Your Average Game" as a tool for introductory statistics students to construct and explore the steps of a hypothesis test for a mean.

The researchers accept that, as with all interviews, there is risk that the researcher who interviewed participants could have introduced bias in the results yielded by the participants. As the research team is made up of statistics instructors with a strong drive to turn situations such as this into valuable learning opportunities as students make valuable observations or ask sincere and relevant questions, there is a chance, through inadvertent facial expressions or other body language, that the interviewer may have influenced participants to answer, or modify their answers, in certain ways. However, every effort was made to keep all questions and reactions as unbiased and unhelpful as possible, so as to allow the natural thought processes of the players to come out as much as possible.

The Institutional Review Board determined this study was exempt from review under Exemption 2a, for use of surveys, interviews, and/or educational tests involving adults for educational research purposes.

3.3.5 Qualitative Coding Methods

This was a qualitative study in which actions, strategies, questions, and thought processes players constructed as they attempted to beat an opponent at the game were recorded and coded. Data were collected through audio recordings and subsequent transcriptions of the participants playing the game and answering questions posed by a researcher.

The researchers implemented a hybrid approach to the coding process, using inductive coding to address research question 1, and deductive coding to address research question 2.

The researchers have designed "Your Average Game" to fall somewhere within type 3 (games which lead to concept formation) and type 4 (games which lead to mathematical investigations), referring to the types of games outlined by Lim-Teo in Section 3.2.3. It is the aim of this project to provide a new game to the short list of games that target a specific learning objective effectively. Should the conclusions to the first research question show that players are investigating the mathematics of the game and how to utilize that knowledge in order to win, it will suggest that the game contributes to those of type 4. Should the conclusions to the second research question show that the strategies and procedures players come up with are similar to that of a formal hypothesis test, it will suggest that the game contributes to those of type 3.

3.3.5.1 Inductive Coding

The inductive coding process involved a first coding, organization, and grouping, and then a second coding.

The first coding was a modified process coding, in which the researchers marked and classified each action being taken, as well as each strategy or belief explicitly stated by each participant. These took the form of basic action words in the game, such as "believe," "reject," "peek," "lie," "tell the truth," etc., other action words applicable to the participant, such as "hesitate," "regret," "calculate," "analyze," "gamble," etc., and any strategies or beliefs expressed when the researcher asked why they were doing something, such as "trying to go for the most points," "playing the mind game," "the truth is hard to believe in this distribution," "peeks will center around the average", etc.

After organizing and grouping the coded actions, strategies, and beliefs, the second coding implemented was an axial coding, in which the researchers investigated relationships and links between the observed actions, strategies, and beliefs. Here, it was noted when two or more codes were consistently found next to each other, or in a particular reoccurring sequence, or if there is a pattern to the codes that immediately precede or follow certain codes of interest (such as "regret").

This coding process was deemed most appropriate by the researchers to address research question 1, pertaining to the strategies and procedures the players perform as they play "Your Average Game," since the question primarily focuses on what participants are doing and why they are doing it.

3.3.5.2 Deductive Coding

The deductive coding process requires that the researchers determine beforehand which behaviors and attitudes they are looking for before coding the recorded data. For this study, a coding structure was implemented that identified elements of the formal steps of a hypothesis test, as taught in introductory statistics courses.

These elements were as follows, with examples of phrases in the interviews that were coded accordingly:

1. Identify a hypothesis: *"okay, so you're saying that there's 61 average," "so, the preview is 69 and the claim is 87," "118 square centimeters…I feel like that's a big mole rat."*
2. Gather data relevant to the hypothesis through sampling: *"I'm gonna do a peek to begin with," "I want to peek another time just to see," "I'll just give it like three peeks."*
3. Summarize sample data: *"so I've seen 126, 134, and 125," "if the result doesn't convince me, I will take the average of all the samples," "but most of my numbers are higher and close together, with that one low one that might be an outlier."*
4. Compare the sample statistic to the hypothesized population parameter: *"well, 133 is pretty close to 127, but probably she is lying slightly," "because the peeks are between the average, so there's a possibility that the claim is true," "okay, the ones I'm looking at are both below the claim."*
5. Determine whether the difference observed between the statistic and hypothesized parameter is due to chance: *"the 123 could be an outlier, or there could be a wide distribution," "well, 234 is nowhere near 165, so something's not right," "I think the numbers are all so close to the average that her claim is believable."*
6. Draw in-context conclusions based on the observed difference: *"so, the average should be less than 133," "I'm gonna be inclined to believe that the claim is too high," "If I get something close to 139 I'll believe. And 139!! I think I have to believe."*

7. Recognize that conclusions have a nonzero probability of being incorrect: *"I will say 75% of the time that I will get it right," "I still would have rejected it, I think, no matter what. So yeah, I think I made the best decision I could with what I had," "if the true average is 133, it's too easy to put 134…like, I feel like I'm just less likely to get it right."*

This was the most appropriate coding process to address research question 2, pertaining to the similarities that may exist between players' strategies during the game and the steps of a statistical hypothesis test, since these steps were already established, and the researchers needed only to note any places during gameplay where participants were formally or informally following one or more of the steps.

3.4 Discussion

3.4.1 Interview Results

The think-aloud research interviews are the primary source of data to be used for interpretation throughout this project.

All seven interviewees participated while about halfway through their first college statistics course. None of them had yet learned about hypothesis testing in their current course or in previous courses.

Table 5 shows the strategies or beliefs that were emergent across the interviews as the participants took on the role of the Poser (seeing the true population distribution and deciding what to tell the opponent is the true average).

Table 5: Emergent strategies and beliefs of interviewees as Poser

Strategy or belief	# of instances across all 7 interviews	# of participants who mentioned
Analyze the histogram to inform decision	23	7
Play the mind game with opponent	18	7
Do whatever stands to get you the most points	15	5
Opponent will likely peek near a certain false claim	13	6
Opponent's peeks will likely center around the mean	12	6
With large variation, truth is harder to detect/believe	9	6
Take a gamble without much thought	7	4
To deceive, lies should stay close to the mean	7	5
Regrets choice that didn't deceive opponent	7	5
Still happy with choice that didn't deceive opponent	3	3

Table 6 shows the strategies or beliefs that were emergent across the interviews as the participants took on the role of the Peeker (given a value for the population mean by their opponent, and must determine through a small number of peeks whether to believe or reject the claim).

Table 6: Emergent strategies and beliefs of interviewees as Peeker

Strategy or Belief	# of instances across all 7 interviews	# of participants who mentioned
More data is necessary for making a good decision	16	7
Play the mind game with opponent	16	6
More peeks help to establish population data range	13	6
Detecting truth from a sample is difficult and random	10	5
Postulate what the histogram might look like	8	6
Regrets incorrect decision	8	6
Do whatever stands to get you the most points	6	6
Try to determine whether a peek was an outlier	4	3
Still happy with incorrect decision	3	2
Take a gamble without much thought	2	2

Table 7 shows how often evidence that a participant was formally or informally thinking through one of the seven previously established steps of a hypothesis test for a mean during the exercise.

Table 7: Evidences of hypothesis steps being formally or informally taken by interviewees

Hypothesis Test Step	# of instances across all 7 interviews	# of participants who mentioned
1 Identify a hypothesis	29	7
2 Gather data relevant to hypothesis through sampling	61	7
3 Summarize sample data	18	6
4 Compare sample statistic to hypothesized parameter	27	7
5 Determine whether difference is due to chance	23	6
6 Draw conclusions based on observed difference	25	7
7 Recognize nonzero probability of being incorrect	13	7

There was no discernible difference between the students in the lower-level course and those in the higher-level course with regard to any of the data outlined in Tables 5 through 7.

3.4.2 Survey Results

The survey responses from those who played the game at home are a supplementary source of data to be used for interpretation throughout this project.

All 14 survey respondents participated while about halfway through their first college statistics course. Some of these respondents had learned about hypothesis testing in a previous high school statistics course.

Table 8 summarizes the main themes that arose from respondents' answers to each of the six questions in the survey.

Table 8: Emergent themes from asynchronous survey questions

Survey Question	
Emergent Theme	**# of participants who mentioned**
Describe the basic strategy you developed to have the best chance at gaining the most points while trying to determine whether the Poser was telling the truth.	
Peeking a lot of times increases ability to make educated guess about mean	6
Use mind games and tricks to determine if opponent is lying	6
Evaluate initial difference between the first preview and the claim	5
Trusting gut reaction	3
More skeptical of big numbers than small numbers	2
Look at difference between the claim and the average of several peeks	2
Peeking a lot of times minimizes point loss	1
Fewer peeks maximizes point gain	1

Explain how peeking is helpful for determining whether the Poser is telling the truth.	
More peeks help calculate a better estimate of the true average	6
It gives more context to the first number that was given	5
Gives a better idea of the range of the population data	2
Didn't initially understand why peeking was useful	2
With every peek comes another chance to read opponent's body language	1
When you were the Poser, did you use the histogram provided to inform your decision about what claim to make about the average? If so, how? If not, why not?	
When you want to lie, you can find a number that is believable	4
If the histogram is skewed, then lie on the side with the bulk of the values	3
Great way to picture where the average is	3
Did not use the histogram, the average was enough information	2
Did not use the histogram, but randomly selected to lie or tell the truth	1
Don't bother trying to interpret weird histograms, look when they are simple	1
Puts range into perspective to determine how big a lie can be	1
What is the smallest number of peeks you think a Peeker could make and still feel fairly confident they will make the right decision?	
1, it's a gamble, but that makes it fun	1
1, more might just confuse you	1
1 or 2	1
2 or 3, 1 tells you almost nothing	1
3 to 5	4
3, less is not enough information, more loses too many points	3
4 gives you a good spread to work with	2
5, unless there are outliers, then more	1
Do you believe there may be a strategy you could develop where you could determine whether the Poser told the truth correctly every time? Why or why not?	
No, the randomness of peeks makes it impossible to nail down exact average	3
No, different opponents will have different strategies	2
No, but maybe you could see a pattern over many games	1
No, there are not enough peeks to know the average for certain every time	1
Yes, with a large number of peeks it is obvious if the opponent lied	3
Yes, if you developed a mathematical method for interpreting peeks	2
Yes, if you play with the same person enough, you can read them	2
Yes, there is always a strategy for everything	1
Were there any times, as a Peeker, where you believe you made the best decision based on the information you had, even though you got it wrong? Why or why not?	
No, I wish I had known mathematically how to reliably use the peeks	3
No, I should have used the data instead of going with my gut	2
Yes, the lie and true average were so close it would be impossible to tell	4
Yes, the lie was another mode in the histogram so peeks made it believable	2
Yes, the random peeks were all so far away from the truth I believed the lie	2
Yes, I misinterpreted the question	1
Yes, the peek was so close to the claim it made sense to believe	1

There was no discernible difference between the students in the lower-level course and those in the higher-level course with regard to the answers provided in the table above.

3.4.3 Emerging Themes

Some themes were consistent across participants regarding how long they had been playing the game. The following is a description of behaviors and beliefs that were regularly observed during each round. The time allotted for the interview allowed almost all participants to play four rounds of the game. Participant 3 only played three rounds, as they spent a great deal of time thinking and analyzing and did not have time for a fourth round. They went through the following process, with their Round 3 following the typical third round as the Poser, and the typical fourth round as the Peeker. Participant 5 actually managed to play six rounds in the time allotted due to much faster decision-making than most participants. They went through the following process, with their Round 1 following the typical first round, Rounds 2 and 3 following the typical second round, Round 4 following the typical third round, and Rounds 5 and 6 following the typical fourth round.

3.4.3.1 Round 1 – Gambling and Insecurity

Participants start the game as the Poser, without feeling a firm grasp on what their opponent will be doing as the Peeker. When asked how they were deciding whether to tell the truth or to lie about the population average, most participants said something to the effect of "I'm just gonna take a gamble that she'll reject it, so I can get the most points" (Participant 1), or "I want to tell the truth, but I also want to get familiar with what I'm going through too" (Participant 6). Most understood that their opponent would be shown random subjects from the population, but few of them felt like they knew how helpful that would be in revealing whether their claim was true or false.

When it was time to take on the role of the Peeker, participants still felt uncertain about how to decide whether their opponent had lied or told the truth. All participants peeked at least once, with most of them peeking only twice. It was clear the participants were not really peeking with the understanding of what information they could obtain, but that they were peeking because that was the big button on the screen and the researcher probably wanted them to press it. These peeks were accompanied by statements like "I don't know…I don't know the possible distributions of this stuff. I'm just gonna go reject" (Participant 5), or "so there's nothing that I, there's no strategy that I can have on this part that predicts how she plays" (Participant 4).

3.4.3.2 Round 2 – The Histogram is the Key

During their second opportunity as the Poser, they really latched on to the idea that the provided histogram of the population distribution was their main means of making a clever decision. Most participants at this point said something like "if you get the distribution and you decide to like say the average is 33, but there's not a lot of data that says that it's 33…then she would know that you're lying" (Participant 2), or "the graph is my biggest thing. I think the graph is the most important thing" (Participant 7). Most came up with the same strategy here as the survey respondents described in question 3 of the survey. If it was a skewed distribution, they lied in the direction of the bulk of the data. If it was a multimodal distribution, they lied by claiming the average was one of the modes. If was more uniform or normal, they either told the truth or gave a lie that was very close to the true average.

As the Peeker, participants remained focused on the histogram. Though they had not been shown the graph for the population, most of them attempted to use a few peeks to imagine what the histogram might look like. They made statements such as "I have another question. The range goes from ten, or…I mean, like, in the graph" (Participant 3), or "I mean you can't be shown the distribution, but like, if I had more of a hint of what it was…" (Participant 5), or "so thinking about that, I'm trying to picture what she was seeing" (Participant 6). After constructing a possible image of the histogram with three or four peeks, they determined whether or not they believed the claim could really be the average of that constructed possible histogram. During this stage, nearly every participant correctly identified whether the opponent was telling the truth.

3.4.3.3 Round 3 – The Mind Game

At this point, most participants drifted from the mathematics of the game and really focused on getting into the mind of their opponent. This was expected, because though the game fosters opportunities to think through hypothesis testing, it is also at its core a lying game. As the Poser, most participants still took a small amount of time looking at the histogram, but spent most of their decision-making time taking stock of the deception that had happened thus far, making comments like "because of how many times I've told the truth so far, I think she's caught on to that" (Participant 2), or "I'm thinking, like, if she thinks that I am like a liar person, she will be confident that I'm more likely to lie" (Participant 3).

This mentality continued to prevail as participants took their turn as the Peeker. Most participants only peeked once or twice during this round, but made most of their judgments based off the history and body language of their opponent, saying "I think that the better way to play is to like try and figure out whether or not you like to lie, and kind of run with that" (Participant 4), or "so it seems like a lot of it is a reading thing and like, I don't know, you've been lying a lot" (Participant 5). Some participants still let their peeks give some sway to their decision, but the primary influencer was their belief about the opponent. During this stage, most participants incorrectly identified whether the opponent was telling the truth.

In preparation for this study, the researchers were interested to find out just how much participants would focus on the deception and mind games aspect of the activity. The results here were quite satisfactory, as there seems to have been just enough attention on the mind game to make the experience fun for players, without overshadowing the true purpose of the exercise.

3.4.3.4 Round 4 – A Larger Sample Size is Necessary to Secure the Win

At this point in the game, participants had varying methods of deciding whether to lie to their opponent or not. However, as they thought through their different strategies during this round, most made statements about recognizing that the peeks their opponent would see would most likely center around the true average, and that they needed to take that into account. They observed "because the 'reject' and 'believe' points are the same, and with the frequency…I might lie, but close to the original average" (Participant 2), and "so I know, like the average, it will be more inclined to the left side…so I think it will be a stupid decision if I choose, well, something over 200" (Participant 3).

When participants were told that this was the final round of the game for the interview, nearly all of them forgot about the mind game as the Peeker and tried to use the peeks to their advantage as best they could. Nearly every participant peeked three or more times during this last round, giving reasons such as "I just want to maximize my chances of getting some points rather than maybe getting a lot of points" (Participant 1), or "I only got four points last time, you know, so maybe peeking ten times might be like…like I don't think ten times, but enough to where you're confident" (Participant 7). It seems when the final points were on the line and they just needed to make the right decision, participants intuitively recognized that a larger sample size would be their greatest asset toward accomplishing that goal.

3.4.4 Evidences of Conducting Hypothesis Tests

It was expected that many elements of a hypothesis test would be observed throughout the interviews, as some of the steps are inherent in the gameplay of the Peeker role. For example, when the Peeker reads the claim of their opponent, they have identified the hypothesis (step 1). If a Peeker chooses to peek, they are gathering data (step 2). To some degree, as long as the Peeker has taken the data they obtained into consideration when they click 'Believe' or 'Reject,' they are drawing conclusions based on the difference between the statistic and the parameter (step 6).

Steps 3, 4, and 5 of hypothesis testing (summarizing sample data, comparing the statistic to the hypothesized parameter, and determining whether the difference is likely due to chance error) were of the most interest to the researchers in this study, as these steps are not built into the mechanics of the game. These steps would only be followed during gameplay if players chose to follow that line of thinking. While it was not expected that players would actually calculate sample means, standard deviations, or p-values, it was expected that participants would make statements and observations that showed they were considering related concepts such as measures of center, distances between observed and expected values, and observed ranges of values as they made their decisions.

There is not much depth to the instances where participants followed step 1, and the notable thought processes behind step 2 have been mentioned in the section on emerging themes. Following are the observations of interest regarding steps 3 through 7.

3.4.4.1 Step 3 – Summarize Sample Data

To some degree, participants informally summarized their sample data any time they peeked one or more times in a round. Most participants focused on how many of their data points were lower than the claimed average, and how many were higher, making statements such as “the two numbers…the higher numbers are very close together, but then there’s the lower one that’s closer to the claim” (Participant 6). Some postulated early on that taking the average of their sample points might be helpful, and some even estimated sample averages in their head, but only one participant asked to pull out a calculator and actually compute the sample average to compare to the claimed population average. Most of the time, in this casual setting, participants would simply look at the list of sample data points they had obtained, and then tried to decide if the claimed average could be the center-point of those sample points.

3.4.4.2 Step 4 – Compare Sample Statistic to Hypothesized Parameter

Since participants rarely calculated a concrete sample statistic, they instead compared groups of data points to the hypothesized parameter, making statements like “then the two peeks that you got were clear out on the side” (Participant 4), or “the average is between the peeks, so there’s a possibility the claim is true” (Participant 2). Because they focused on individual data points rather than a sample average, they were easily swayed by data points that were within a distance of three or four from the claimed average. If even one peek was right next to the claimed average, they were much more likely to believe the claim, regardless of the other values in the sample.

3.4.4.3 Step 5 – Determine Whether Difference is Due to Chance

In an introductory statistics course, the instructor will often spend a great deal of time demonstrating that the variance of a population or sample is a crucial element of determining whether a result may simply be due to chance error. While no participant tried to calculate the variance of their sample data, they certainly thought about the concept as they were making their decision to believe or reject the opponent’s claim. Phrases like “234 is nowhere near 165, and I just want to peek again…maybe I got an outlier" (Participant 7), or "142 seems a little bit on the higher side, but I would need more information to be able to determine how much higher is that really, comparatively" (Participant 6), or "the 123 could be an outlier, or there could be a wide distribution" (Participant 2), were very encouraging evidence that participants understood that the range or spread of the unknown distribution was an important factor to take into account while going through this process.

3.4.4.4 Step 6 – Draw Conclusions Based on Observed Difference

When participants were not too caught up in the deception aspect of the game and concentrated on using data to inform their decision, the difference between what they saw and what their opponent claimed was the predominant influence on their final choice. Participants consistently gave reasons for their decisions such as "I believed her because I took a few peeks, and they were all decently close" (Participant 5), or "if most of the ones I'm looking at are all below the claim, I'm gonna be inclined to believe that the claim is too high" (Participant 1). Whenever a participant followed a strategy like this and still made an incorrect decision, if they expressed regret for their decision, it was usually centered in the wish that they knew a reliable mathematical way to use the information they were given to make a better decision. This was an especially promising result, as students learn best when they believe what they are being taught will answer a question or solve a problem they have come up against. This suggests that students who have played this game may be more likely to be attentive and commit things to memory when an instructor lectures about the mathematics of a formal hypothesis test.

3.4.4.5 Step 7 – Recognize Nonzero Probability of Being Incorrect

In statistics, the ability to appropriately interpret the p-value is incredibly important. This ability begins with a recognition that there is always some chance that one may observe very deviant data even though the null hypothesis is true. In a very rudimentary way, participants recognized this by stating that they "will be wrong 25% of the time" (Participant 3) or have a "one out of seven" (Participant 4) chance of being incorrect even though their sample observations were far away from the opponent's claim. With the random nature of their sample, all participants intuitively recognized that there was a chance that an outlier was secretly skewing their sample data.

Most introductory statistics courses present either 'rejecting' or 'failing to reject' the null hypothesis as the two appropriate conclusions. Most students will, at some point, question why they must fail to reject the null hypothesis rather than simply believing or accepting the null hypothesis. Several participants answered this very question as they thought aloud their decision-making processes, stating that "it can give you like, a false sense of security for believing something that's still likely a lie" (Participant 5), or "if the true average is 133, it's too easy to put 134…like, I feel like I'm just less likely to get it exactly right" (Participant 1). These participants began to recognize that if the true population mean and an untrue claim are very close to each other, it is nearly impossible to detect.

3.4.5 Misconceptions

There were a few misconceptions about the hypothesis testing process that were brought to light as participants played "Your Average Game" which were notable.

Participant 4 felt adamant for most of the experience that a sample was not very useful unless it was a certain percentage of the population. Since they had seen that the populations in the game had over 1,000 data points in them, they believed that any sample of size smaller than 200 or so was effectively meaningless.

Participant 2 was certain that if their decision about the Poser's claim was incorrect, then they must have made a foolish mistake at some point. It was not feasible for this participant that chance error could have a hand in influencing them to make a well-informed decision that was ultimately incorrect.

For the first half of the game, four different participants believed that they invariably had a 50/50 chance of correctly identifying whether the Poser had told the truth. When pressed about why, even after peeking at several data points, they would characterize their chances of being correct in this way, the participants felt that since there are two options (believe or reject), and the Poser effectively had one of two options (tell a lie or the truth), they "have a 50/50 chance any time" (Participant 1).

Three participants, when acting as the Poser and presented with a multimodal distribution, expressed that, rather than choosing a lie that was close to the true average, that the next most believable claim to the truth would be one of the modes that was not the true average. The fact that this other mode was more likely than most other values to be obtained in the opponent's sample gave them a false sense of hope that the opponent's peeks might also center around the mode they claimed as the true average.

Two participants had the sense that the random nature of the peeks made them largely irrelevant. While statistics instructors will be sure to instill in their students' minds that a random sample is the most unbiased and therefore most useful type of sample, these participants felt that randomly selecting the values from the population created the least useful type of sample.

3.4.6 Future Classroom Use

The researchers believe that having students play "Your Average Game" either in class or as a homework assignment can be a valuable use of time. In several preliminary studies as the game went through different iterations of development, the research team used the game as a tool to prompt deeper thinking about hypothesis testing in various introductory statistics courses at both college and high school levels, and were met with modest success.

The game can be played before students learn about hypothesis testing, followed by an activity or worksheet where students think through the questions posed on the asynchronous survey in this study. These questions encourage students to formally consider the usefulness of a larger sample size, how different population distributions might affect sample values, and how well-informed statistical conclusions still have a probability of being incorrect conclusions. These thought exercises, accompanied by an experiential desire to know the mathematics behind the game, have the potential to prepare students well to understand the logic and formulas behind the steps of a hypothesis test when they are taught in class.

The game can also be played after students have learned the steps of a hypothesis test, while they are still relatively unfamiliar with the mathematics of those steps. This gives students an opportunity to construct their own ideas about how one could mathematically follow the steps they have learned in a fun and relaxed setting, where they won't necessarily assume that they are bad at math when they make incorrect conclusions. In this case, the gameplay could be followed up with not only the asynchronous survey questions, but also with a worksheet or discussion about where certain parts of a hypothesis test could be found in the game, such as the null hypothesis, the alternative hypothesis, the population, the sample, the parameter, and the statistic (the statistic is the only element here that is not explicit in the game, though students could have easily calculated it at any time). Future research would be merited regarding students' approach to the game after having learned about hypothesis testing, as the interviews conducted for this study were only conducted with students who had not previously been exposed to statistical hypothesis testing in a classroom setting.

This game may also have some utility outside of the classroom context, as most introductory statistics students have not tried to answer a question or solve a problem that required the statistical thinking behind hypothesis testing before they enroll in the course. “Your Average Game” may help players better develop that statistical thinking even without guidance from a worksheet or instructor, but that is outside the scope of this project.

3.5 Conclusion

It is evident that college-age players of “Your Average Game” who have not yet learned formal processes of hypothesis testing develop a wide range of beliefs and strategies as they address the question, ‘is this the true population mean?’ While experiences obviously varied, most participants who were interviewed started playing the game feeling uncertain and approaching the game like a basic gambling game, but then started recognizing the value of being shown a histogram of the population, or of obtaining random sample values to inform their decisions. When the effort of using this valuable data felt too time-consuming or taxing, participants often began employing a dominant strategy of mind games and deception in an attempt to come out on top, but eventually realized that this was not as reliable for detecting the truth, and subsequently redoubled their efforts to effectively interpret larger sample sizes than they had gathered before. Players had to weigh the utility of a larger sample size with the cost of obtaining it, and to accept that randomly selected values are more useful than using their gut to determine what is true. Interviewees and survey respondents alike developed a desire to know a reliable mathematical way to use random sample data to determine whether a stated value is the true mean of a population distribution. This suggests that “Your Average Game” may be an excellent primer to help students develop a deeper desire to learn the material related to hypothesis testing in a statistics course. (Research Question 1)

Participants consistently questioned, explored, and enacted relaxed forms of each of the steps of a statistical hypothesis test for a population mean, without the researchers inviting or encouraging them to follow or consider those steps. While three of the steps leading up to making a conclusion are inherent in the gameplay itself, the other three are merely opportunities the player can pursue without suggestion or guidance by the game. Nonetheless, participants still showed evidence of performing strategies corresponding to each of these implicit steps, despite not yet having learned them in a classroom setting. It is especially notable that participants recognized that larger samples are easier to use and more reliable than smaller samples, that calculating the sample average is a useful step during the process, that the distance between what was observed and what was expected should be considered in context of the spread of the observed data, and that if a sample is randomly selected it necessarily introduces a small probability that a true claim may appear unbelievable due to chance error. This suggests that not only might "Your Average Game" be an effective instrument for allowing students to construct the steps of a hypothesis test themselves and to recognize that the process taught in class is the most natural and efficient way to answer the question at hand, but it may also be beneficial as a tool for providing students a concrete way to grapple with the difficulties of hypothesis testing and interpretation early in their exposure to these big ideas throughout their coursework. (Research Question 2)

CHAPTER IV – DISSERTATION IMPLICATIONS AND CONCLUSION

4.1 Implications and Contributions to Knowledge

4.1.1 Practical Implications

The first project is a concrete example of how the traditional grading system in a classroom can be altered to allow students more autonomy in the way they learn the material and to remove elements of the traditional grading system that can feel discouraging as students make mistakes throughout the learning process. The efficacy and consistency of the traditional grading system has been questioned for decades, but most instructors continue to use these traditional methods because there are not many concrete examples of how the system can be changed for the better without drastically changing how the class must be structured or taught (Cowan, 2020). This project can serve as one such concrete example to help those instructors who would like to change the method by which they assign grades to their students' work.

The second project shows that players are constructing concepts and strategies while playing "Your Average Game" that are helpful steppingstones to understanding general methods of hypothesis testing. A clear practical implication of this study is that this game can immediately become a resource for instructors to use as a part of their statistics curriculum. This means that a small but immediate enhancement is available to help instructors break up the monotony of lecture and repetition with something memorable and productive.

4.1.2 Theoretical Implications

The first project strengthens the theory laid out by Yu-kai Chou about how effective gamification must go beyond the points, badges, and leader boards that most often come to mind when people think of gamification. The study focuses on several elements of Chou’s Octalysis Framework and shows the results produced from using the key drives stated in the framework in the process of gamifying a part of the classroom structure. (Chou, 2019)

The second project demonstrates that the use of games in the classroom can go beyond mere memorization and recollection experiences. “Your Average Game” is an example illustrating that a well-designed game can naturally guide a student through a process or algorithm in a salient way that fosters connection and understanding rather than the gathering of seemingly unrelated facts. When using games for mathematics learning, instructors and curriculum designers often utilize activities that employ unproductive or even counter-productive practices such as speed pressure, timed testing, and blind memorization, which stand in the way of good number sense (Boaler et al., 2015). “Your Average Game” challenges the assumption that games are merely for practicing an algorithmic skill and aims to show that games can be used to help students recognize connections between concepts and the reasoning behind processes in a setting that feels more natural and freeform.

The project also strengthens theory that shows that games are a very useful tool for helping students with mathematical modelling. Powell, Cangelosi, and Harris (1998) said “Game playing is a good technique for creating data sets which students firmly understand and feel responsible for. Intuitive understanding encourages students to apply mathematics and supplies an idea of when things work and when they don't. The sense of responsibility keeps them going when the going gets tough . . . students become responsible for the mathematics because they are creating it, as opposed to being victimized by it.”

4.1.3 Research Limitations

Both projects outlined in this dissertation were conducted with undergraduate students at Utah State University. Consequently, the majority of participants were white, native English speakers, ages 20 to 26. This limits the scope to which these data can be generalized. The gamified grading structure may be more or less effective for differing learning age groups or for students of other various ethnicities. Similarly, the lack of participation by non-native English speakers in the research on "Your Average Game" is an issue since the game is primarily text-based.

The grading gamification in the first project, while designed with more types of courses in mind that exclusively statistics courses, has not been researched in other settings. Without further investigation, it is unclear whether such a gamification approach would be effective or appropriate in other non-mathematical courses.

"Your Average Game" in the second project was only formally and informally explored with players who were currently taking an introductory statistics course, and who were either about to reach the hypothesis testing unit of the course, or had recently begun learning about hypothesis testing in the course. Therefore, these data cannot necessarily be generalized to a wider audience of players who have not yet had any formal introductory statistics instruction.

4.1.4 Future Work

Further research would be merited from both projects outlined in this dissertation.

Further research into the first project would allow more specific exploration into which barriers to motivation are inherent in the traditional grading system, and which gamification techniques are most effective at removing those barriers.

Further research into the second project would allow a more quantitative understanding of how "Your Average Game" may affect students' long-term recollection and comprehension of formal hypothesis testing methods when it is included as part of the statistics course curriculum. Research could also be done on "Your Average Guessing Game" as a tool for exposing players to the concepts involved with confidence intervals. Across the game site, for both versions of the game, it would be very useful develop methods for collecting and analyzing the game data. By recording mouse clicks, text entries, wait times, correct/incorrect decisions, rounds, and scores, more insight could be obtained on play patterns and prompt difficulty.

It would be pertinent to conduct similar studies to both projects outlined here in other schools and settings across the country, in order to explore effectiveness and reception across more diverse ethnic groups, as well as for non-native English speakers.

References

Adams, D. M., Mayer, R. E., MacNamara, A., Koenig, A., & Wainess, R. (2012). Narrative games for learning: Testing the discovery and narrative hypotheses. *Journal of Educational Psychology*, *104*(1), 235. https://doi.org/10.1037/A0025595

Afari, E., Aldridge, J. M., & Fraser, B. J. (2012). Effectiveness of using games in tertiary-level mathematics classrooms. *International Journal of Science and Mathematics Education*, *10*, 1369-1392. https://doi.org/10.1007/s10763-012-9340-5

Armstrong, T. (2000). *Multiple Intelligences in the Classroom* (2. edition). Alexandria, VA: Association for Supervision and Curriculum Development.

Boaler, J., Williams, C., & Confer, A. (2015). Fluency without fear: Research evidence on the best ways to learn math facts. *Reflections*, *40*(2), 7-12.

Bowman, N. A., & Seifert, T. A. (2011). Can college students accurately assess what affects their learning and development? *Journal of College Student Development, 52*(3), 270-290. https://doi.org/10.1353/csd.2011.0042

Caillois, R.. (1957). *Les jeux et les hommes*. Paris: Gallimard.

Chou, Y. (2019). *Actionable Gamification: Beyond Points, Badges, and Leaderboards*. Birmingham, UK: PACKT Publishing Limited. https://www.packtpub.com/product/actionable-gamification/9781839211706

Cowan, M. (2020). A legacy of grading contracts for composition. *Journal of Writing Assessment*, *13*(2). Retrieved from https://escholarship.org/uc/item/0j28w67h

Currie, C. T. (2014). *Reciprocal Effects of Student Engagement and Disaffection on Changes in Teacher Support Over the School Year*. MS Thesis. Portland State University, Department of Psychology. https://doi.org/10.15760/etd.1645

Deif, A. (2017). Insights on lean gamification for higher education. *International Journal of Lean Six Sigma*, *8*(3), 359-376. https://doi.org/10.1108/IJLSS-04-2016-0017

Delgado-Gómez, D., González-Landero, F., Montes-Botella, C., Sujar, A., Bayona, S., & Martino, L. (2020). Improving the teaching of hypothesis testing using a divide-and-conquer strategy and content exposure control in a gamified environment. *Mathematics*, *8*(12), 2244. https://doi.org/10.3390/math8122244

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011, September). From game design elements to gamefulness: defining "gamification". *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments* (pp. 9-15). *Association for Computing Machinery*. https://doi.org/10.1145/2181037.2181040

Evangelista, F. & Hemenway, C. (2002). The use of the Jigsaw in hypothesis testing. *2nd International Conference on the Teaching of Mathematics at the Graduate Level.* Hersonissos, Crete, Greece.

Gardner, H., & Hatch, T. (1989). Educational implications of the theory of multiple intelligences. *Educational Researcher, 18*(8), 4–10. https://doi.org/10.3102/0013189X018008004

Gilbert, T. F. (2007). *Human Competence: Engineering Worthy Performance*. San Francisco, CA: Pfeiffer.

Glaser, R. (2003). Assessing expert knowledge representations of introductory statistics. *CSE Tech. Rep. No. 600.* Los Angeles. University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gressick, J., & Langston, J. B. (2017). The guilded classroom: Using gamification to engage and motivate undergraduates. *Journal of the Scholarship of Teaching and Learning*, *17*(3), 109-123. https://doi.org/10.14434/v17i3.22119

Hebert, S. (2018, March). *The Power of Gamification in Education* [Video]. TEDxUAlberta Conference. https://www.ted.com/talks/scott_hebert_the_power_of_gamification_in_education

Hiatt, J. (2006). *ADKAR: A Model for Change in Business, Government, and Our Community*. Loveland, CO: Prosci Learning Center Publications.

Huang, R., Ritzhaupt, A. D., Sommer, M., Zhu, J., Stephen, A., Valle, N., ... & Li, J. (2020). The impact of gamification in educational settings on student learning outcomes: A meta-analysis. *Educational Technology Research and Development*, *68*, 1875-1901. https://doi.org/10.1007/s11423-020-09807-z

Jones, M. G., Jones, B. D., & Hargrove, T. Y. (2003). *The Unintended Consequences of High-Stakes Testing*. Lanham, MD: Rowman & Littlefield.

Kohn, A. (2013). The case against grades. *Counterpoints, 451*, 143-153. http://www.jstor.org/stable/42982088

Landers, R. N., Auer, E. M., Collmus, A. B., & Armstrong, M. B. (2018). Gamification science, its history and future: Definitions and a research agenda. *Simulation & Gaming*, *49*(3), 315-337. https://doi.org/10.1177/1046878118774385

Langran, J., & Purcell, S. (1994). *Language Games and Activities. Netword 2: Teaching Languages To Adults.* London, England: Centre for Information on Language Teaching and Research.

Lee, S. K. (1995). Creative games for the language class. *English Teaching Forum*, *33*(1), 35-36. https://www.scribd.com/document/99650330/Vol-33-No-1-LEE-SU-KIM

Lew, M. D., Alwis, W., & Schmidt, H. G. (2010). Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education, 35*(2), 135-156. https://doi.org/10.1080/02602930802687737

Lim-Teo, S. K. (1991). Games in the mathematics classroom. *Teaching and Learning, 11*(2), 47-56.

MacDonald, M., & Reardon, T. (n.d.). Gilbert's Performance Matrix. Retrieved March January 18, 2021, from https://hptmanualspring16.weebly.com/gilberts-performance-matrix.html

Majuri, J., Koivisto, J., & Hamari, J. (2018). Gamification of education and learning: A review of empirical literature. In *Proceedings of the 2nd international GamiFIN conference, GamiFIN 2018*. CEUR-WS. Pori, Finland.

Maloney, D. P., & Masters, M. F. (2010). Learning the game of formulating and testing hypotheses and theories. *The Physics Teacher, 48,* 22-24. https://doi.org/10.1119/1.3274353

Mayer, R. E. (2016). What should be the role of computer games in education? *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 20–26. https://doi.org/10.1177/2372732215621311

Mertens, D. M., & Wilson, A. T. (2019). Program evaluation theory and practice: A comprehensive guide. In *Program Evaluation Theory and Practice: A Comprehensive Guide* (p. 230).

Ortiz, M., Chiluiza, K., & Valcke, M. (2016). Gamification in higher education and STEM: A systematic review of literature. *EDULEARN Proceedings*. https://doi.org/10.21125/edulearn.2016.0422

Partridge, T., & Schneiter, K. (2023). Evaluation of effects of gamifying grading strategies on student motivation and resilience. *College Teaching*, 1–11. https://doi.org/10.1080/87567555.2023.2227985

Powell, J. A., Cangelosi, J. S., & Harris, A. M. (1998). Games to teach mathematical modelling. *SIAM Review*, *40*(1), 87–95. https://doi.org/10.1137/s0036144596310021

Richter, F. (2020, March 11). Infographic: Super Mario: The Timeless Bestseller. Retrieved August 19, 2020, from https://www.statista.com/chart/5764/best-selling-super-mario-games/

Rickard, K. M., & Titley, R. W. (1988). The hypothesis-testing game: A training tool for the graduate interviewing skills course. *Teaching of Psychology, 15*(3), 139-141. https://doi.org/10.1207/s15328023top1503_8

Rober, M. (2018, April 7). "The Super Mario Effect – Tricking Your Brain Into Learning More [Video]." TEDxPenn Conference. https://www.ted.com/talks/mark_rober_the_super_mario_effect_tricking_your_brain_into_learning_more

Sanchez, D. R., Langer, M., & Kaur, R. (2020). Gamification in the classroom: Examining the impact of gamified quizzes on student learning. *Computers & Education*, *144*, 103666. https://doi.org/10.1016/j.compedu.2019.103666

Sheldon, L. (2011). *The Multiplayer Classroom: Designing Coursework as a Game.* Boston, MA: Cengage Learning PTR.

Smith, T.M. (2008). *An Investigation into Student Understanding of Statistical Hypothesis Testing.* Doctoral Dissertation. Retrieved from Digital Repository at the University of Maryland. (umi-umd-5658.pdf)

Sotos, C., Vanhoof, S., Noortgate, W. & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? *Journal of Statistics Education 17*(2). https://doi.org/10.1080/10691898.2009.11889514

Squire, K. (2005). Changing the game: What happens when video games enter the classroom?. *Innovate: Journal of Online Education, 1*(6). Retrieved February 21, 2023 from https://www.learntechlib.org/p/107270/

Stojković, M. K., & Jerotijević, D. M. (2011, May). Reasons for using or avoiding games in an EFL classroom. In *1st International Conference on Foreign Language Teaching and Applied Linguistics* (pp. 5-7). Sarajevo, Bosnia.

Subhash, S., & Cudney, E. A. (2018). Gamified learning in higher education: A systematic review of the literature. *Computers in Human Behavior*, *87*, 192–206. https://doi.org/10.1016/j.chb.2018.05.028

Sugar, S. (1999). *Games that Teach: Experiential Activities for Reinforcing Learning*. Hoboken, NJ: Jossey-Bass.

Von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, *51*(8), 58–67. https://doi.org/10.1145/1378704.1378719

Appendix: “Your Average Game” Rules and Example Playthrough

RULES OUTLINED ON "YOUR AVERAGE GAME" WEBSITE

As mentioned in Section 3.3.1, the following is the set of rules as described on the website for "Your Average Game" (found at https://tinyurl.com/YourAverageGame) when, under "Your Average Lying Game," the "How to Play 'Your Average Lying Game'" button is selected.

"Your Average Lying Game"

(2 to 5 players)

A TURN OF PLAY

One player takes on the role of the "Poser" and the others of "Peeker". Whoever's turn it is will be the Poser, who looks privately at the computer/phone screen and clicks "Only I can see the screen!". The Poser is then shown a chart depicting the distribution of some set of numbers, such as the amount of time (in seconds) that each of Ms. Chapman's 5th grade students could hang from the monkey bars before falling. The Poser will also be told the average of the distribution (in this case, let's say the average was 102 seconds). They must then decide whether they will tell the truth or lie about the true average to the other players. If they choose to tell the truth, they simply click the center button at the bottom of the screen with the average on it and then type the average and submit. If they choose to lie, they must choose one of the other buttons at the bottom of the screen containing different intervals in which the Poser's lie could be. For instance, the Poser could select the button that says "94-99" and then type in "98", claiming to the other player that the true average is 98, even though it really is 102. The Poser can use the shape of the distribution's graph, as well as the points outlined on the buttons at the bottom of the screen, to aid their decision in whether (and/or how) to lie. For each button at the bottom of the screen, there is a number in a blue box and a number in a red box. The number in the blue box is how many points the Poser will win if a Peeker believes their claim. The number in the red box is how many points the Poser will win if a Peeker rejects their claim.

Once the Poser has submitted a claim, it is now the Peekers' job to determine whether they believe the Poser has told them the true average of the distribution or not. To help accomplish this, the Peekers are given information on one random subject in the population (for instance, they are told that the claimed average amount of time Ms. Chapman's 5th Graders can hold onto monkey bars before falling is 98 seconds, and one randomly selected 5th Grader was able to hold on for 116 seconds). Peekers can choose to peek at another random subject in the population in order to get more information In this case, a Peeker could click the button "Peek at a 5th Grader", after which a random 5th Grader will be selected, and the number of seconds they held onto the monkey bars before falling would be revealed. A Peeker may decide after the first subject is revealed that they believe or do not believe the Poser's claim, or they may decide after any subsequent peek. Peekers may ask for up to 10 peeks, always with the choice to stop after any peek and decide whether or not they believe the Poser's claim.

SCORING A TURN

Once all the Peekers have decided whether or not they believe the Poser, it is revealed whether or not the Poser's claim was the true average. The Poswer wins more points for each Peeker they managed to deceive (each Peeker that believed a lie or rejected the truth). Each Peeker wins points if they make the correct decision, and they lose points if they make the incorrect decision. The amount of points won or lost goes down with each Peek they made before making their decision. So, a Peeker who peeked 8 times before deciding doesn't stand to gain very many points if they were correct, but they also won't lose very many points if they were incorrect.

WINNING THE GAME

One round of play consists of every player taking a turn being the Poser. At the end of a round, if at least one player has the minimum number of points needed to win (as stated on the scores page at the end of any player's turn), the game ends. Then, whoever has the most points wins. If no player has the minimum number of points needed to win, another round is played.

EXAMPLE PLAYTHROUGH

The following is an example playthrough between Anandi (Player 1) and Blair (Player 2). Figures 5 through 9 in Section 3.3.1 show screenshots of some of this gameplay.

At the start of play, the screen says "Player 1, It's your turn!" with a button below that says "Only I can see the screen!" Anandi moves the screen so Blair cannot see it, and then clicks the button. She is shown a histogram of a bimodal data set, with the first very tall peak around 30, and another much smaller peak at around 85. The screen says "Make a Claim. Depicted in the graph to the right is the distribution of grump-levels among all your grumpy gophers. The average grump-level is 34 grumps. You must now decide whether to present the truth or lie about the average. To do so, select a range below and input a number. The numbers below the ranges are the points you will receive if your claim is believed (in blue) or rejected (in red)."

Anandi sees that if she tells the truth (34), she will not receive any points if Blair believes her. However, if Blair doesn't believe the 34, Anandi could get 16 points. But, Anandi thinks it would be more fun to lie about the average, and thinks she could get away with a large lie because of the second mode around 85. So, Anandi chooses the largest interval (46 and above), and then types in 46. With this choice she will gain twelve points if Blair believes her, and only two points if Blair recognizes that she's lying.

They turn the screen so Anandi and Blair can both see. Blair reads the following: "Player 1 garners grumpy gophers and claims the average gopher grump-level to be 46 grumps. One of the gophers has a grump-level of 30 grumps. Do you believe or reject their claim?" Right now, Blair could receive between 12 and 20 points if he makes the correct decision, or lose between 7 and 14 points if he makes the incorrect decision. He decides that one gopher is not enough information, so he clicks "Peek at a Grumpy Gopher." A pop-up then warns him that if he peeks at a gopher, then he can only gain between 10 and 18 points for a correct decision, and lose between 6 and 13 points for an incorrect decision. Blair decides that's work it and moves forward with the peek.

It shows that the new randomly chosen gopher has a grump-level of 31 grumps. Now his two gophers are a 30 and a 31, and Anandi claimed the true average is 46. Anandi is keeping a straight face through all of this, and Blair really wants to be right, so he peeks one more time, even though he'll only be able to gain between 8 and 16 points or lose between 5 and 12 points. The third gopher has a grump-level of 37. After seeing 30, 31, and 37, Blair feels confident that Anandi is lying, so he rejects her claim.

The screen reveals that Anandi was indeed lying. Blair receives eight points for being right, and Anandi only receives two points since she didn't manage to fool Blair. Anandi points out to Blair the second mode near 85, now that they can both see the histogram, to justify why she dared make such an outrageous claim.

It shows here that a player will need at least 60 points at the end of a round in order to win. They click "Next Round," after which the screen shows "Player 2, It's your turn!" Blair turns the screen so only he can see, and clicks "Only I can see the screen!"

Blair receives a prompt about the girths of his fat cats. This time, the histogram is bell-shaped, with very low variability. He's not sure what to do, because he thinks that if he tells the truth, any of Anandi's peeks will be close to the truth and she'll believe him, so he won't get many points. But, if he tells a large lie, he thinks it will be obvious that the peeks are not centered around the lie, and Anandi won't believe him, so he won't get many points. He decides that his best course of action is a really small lie. He makes a claim that is just two away from the true average, which provides five points of Anandi believes him, or five points if Anandi doesn't believe him. Now, it doesn't matter to him what Anandi decides!

They turn the screen so they both can see. Anandi reads the prompt, and receives a value that is very close to Blair's claim. She decides to peek even though she stands to gain less points, and it's another number close to Blair's claim. They're so close, in fact, that Anandi just decides to believe Blair after gathering just two data points.

It is revealed that Blair lied. Blair receives five points, which was already guaranteed to him. Anandi loses 10 points for making an incorrect decision. Now Blair has 13 points and Anandi has -8 points. They click "Next Round" and it's Anandi's turn to be the Poser. Since she has negative points now, she's going to have to decide if she wants a guaranteed small number of points with a small lie, or if she wants to risk trying to get a large number of points by either telling the truth or telling a large lie.

CURRICULUM VITAE
Todd Partridge

CAREER OBJECTIVE:

To obtain a position at a prestigious university as an instructor and researcher of statistics education. Special areas of interest: games and gamification, assessment strategies with growth mindset, and discovery-based learning.

EDUCATION:

BS in Mathematics & Statistics with Minors in Computer Science, Psychology, Economics, and Family & Human Development, Utah State University, Logan, Utah. (Dec. 2017) GPA: 3.54, Graduate Cum Laude. Ph.D. in Statistics and Instructional Technology & Learning Sciences, Utah State University, Logan, Utah. (expected Aug. 2023) Grad GPA: 3.81. Dissertation research conducted at Utah State University in Logan, Utah 2020-2023.

EXPERIENCE:

STATISTICS EDUCATOR, Utah State University, Logan, Utah

(September 2015 – May 2023).

Courses Taught: STAT 1040, MATH 2020.

Recitations Taught: STAT 1040, STAT 1045, STAT 2000,

STAT 2300, STAT 3000, MATH 1050.

Teaching Assistant for: STAT 3000, STAT 4010, MATH 5010, MATH 5720.

WORK/ACCOMPLISHMENTS:

AWARDS:

- Graduate Student Teaching in Excellence Award (2023, 2022, 2020),
- Excellence in Teaching Award (2019),
- Outstanding Undergraduate Recitation Leader Award (2017, 2016).

PRESENTATIONS:

- “Celebrating Mitsake-Making” at USU Graduate Student Teaching Showcase. October 25, 2023, Logan, UT.
- “You Think This Is A Game? Gamification Strategies for Student Resilience” at USCOTS 2021. Virtual event on June 28, 2021.
- “Stat Starters: Rules and Routines I Set In My Classroom That Students (Eventually) Thanked Me For” at USU TWeT Conference. August 20, 2019, Logan, UT.

PUBLICATIONS:

- Partridge, T., & Schneiter, K. (2023). Evaluation of effects of gamifying grading strategies on student motivation and resilience. *College Teaching*, 1–11. Milton Park, Oxfordshire: Taylor & Francis. https://doi.org/10.1080/87567555.2023.2227985