Improving Educational Game Design Methods: A Rubric to Assess the Engagement and Educational Value of Educational Games

Ken Hoff
University of Colorado at Boulder
kendall.hoff@colorado.edu

December 5, 2013

Abstract

Existing educational games lack the combination of education and engagement. The objective of this thesis was to research, synthesize, and test a more effective educational design method. First, an educational game design rubric was synthesized from current literature and existing educational games. Then, players applied this rubric to known educational games. In addition, players completed a quiz on game material that accompanied some of the games, to test if the game had improved their content knowledge or skills.

Amazon's Mechanical Turk was used to find players to complete this survey. It produced a significantly large number of responses at a very small cost. Afterwards, the quizzes were scored and analyzed for statistical significance using a two-tailed t-distribution method. The game design rubric responses were analyzed for consistency using an inter-rater reliability metric.

Only one of the games (The Oregon Trail) had a statistically significant improvement in the quiz scores (5%), and only a few of the rubric items placed in the "Slight Agreement" category as measured by inter-rater reliability. There are no statistically significant conclusions that can be drawn from this research, but it provides an effective first step for future research using similar content or procedures.

Contents

List of Figures

Chapter 1

Introduction

Educational games lack the combination of education and engagement. In ineffective games, the content being taught is either uninteresting or unrelated to the gameplay. A good example is Math Blaster; players answer simple arithmetic questions to fill up their energy, which they then use to shoot space debris. These games are usually just homework problems, either included as an unrelated component inside the game, or interspersed between gameplay segments in a quiz format. It seems that game designers do not know how to produce effective and fun educational games. They are unaware or unfamiliar with the attributes their educational game should or shouldn't have in order to be as effective and fun as possible.

1.1 Thesis Statement/Solution

My goal in this research is to synthesize an educational game design rubric by creating a list of attributes that educational games should have a high degree of in order to be effective and engaging. The list is not a set of standards, but a set of guidelines; it is possible (though unlikely) for a game to contain all of the attributes and still be ineffective, or for a game to have none of the attributes and be very effective. It is simply meant to infer that game designs that contain these elements are more likely to be effective and engaging educational games.

The question to be answered in this research is the following: Are there attributes that an educational game should have in order to be effective and engaging? In this work, I will develop a rubric of game attributes that educational games should have, and investigate whether some selected educational games incorporate these attributes. In addition, I will test the effectiveness of the selected games' ability to educate the players.

1.2 Approach

I will begin by examining current educational game design rubricss, through research of published literature. I will then synthesize my own educational game design rubric; a list of attributes found in effective and engaging educational games. Afterwards, I'll attempt to apply my rubric to existing educational and semi-educational games, and see if the most

effective educational games correlate to my design rubric. In contrast, ineffective educational games won't correlate at all to my design rubric.

1.3 Organization

The rest of this thesis is organized as follows:

Chapter 2 contains background research relevant to this field of research. It explores existing educational frameworks and research that influenced the selection of design elements for the rubric.

Chapter 3 contains the design rubric that games are scored on. It contains 13 game design elements, each of which have been synthesized from the resources in the Research chapter. Each game design element may be found in existing games, to varying degrees described herein.

Chapter 4 contains background information on how our example games were selected, as well as a listing for each game in our study. The game's overview and educational content is described, and screenshots, sources, and notes on implementation are included for each.

Chapter 5 contains all information related to the procedure of our study. It includes how Amazon's Mechanical Turk works, as well as design decisions for the content and process of administering the survey. It also includes the survey that was administered, as well as all permutations and additions.

Chapter 6 contains the aggregated results of the administered survey, as well as non-statistical observations on the data by myself. It includes results from the game opinion, quiz, and rubric sections of the survey.

Chapter 7 contains the statistical analysis of the prominent parts of the survey, the quiz and rubric. We use the t-distribution to determine the statistical significance of the quiz score differences, and use an inter-rater reliability metric to determine the cohesiveness of our design rubric.

Chapter 8 contains conclusions that we have explored in the Results and Analysis sections, as well as a retrospective and future options for additional research.

Chapter 2

Research

2.1 Edutainment vs Educational Games

It is important for us to differentiate between edutainment and educational games.

For this paper, we're going to define edutainment as the simple gamification of a task. The purpose of edutainment is to increase the player's skill at something by getting them to play a game; the player mimics or actually does the skills within the game, and the game rewards them and keeps them engaged in order to continue increasing their skill. These kinds of games are effective at reducing a skill (for example, solving arithmetic problems, or identifying countries on a map) to a formula, where players can 'rote-and-drill' until they are able to complete a problem quickly and without error. However, eductainment games are ineffective at teaching complex concepts and systems, where players have to reason about their environment without having much prior instruction.

For the purpose of teaching complex concepts and systems, we use educational games; players can learn more from educational games because they cannot utilize the "skill and drill" method of education. Educational games typically contain large, complex systems (for example: levels in The Logical Journey of the Zoombinis, Kerbal Space Program) that the player must fundamentally understand in order to make progress within the game. Players who fundamentally understand the system instead of just knowing simple ways to interact with it (e.g. math problems, geography identification) will likely have a much higher retention content retention rate.

2.2 Bloom's Taxonomy

Bloom's Taxonomy [1] is one of the most widely cited classifications of learning. It provides milestones for three different domains of learning: Psychomotor, Affective, and Cognitive.

The first domain, Psychomotor, deals with the ability to perform intricate physical tasks, like hammering a nail or operating a complex machine or instrument. Games are currently very good at teaching concepts within the Psychomotor domain. The best examples are ones where the game provides a near-simulation to the task being performed in real life. Music games like *Guitar Hero* and *Rock Band* improve coordination and motor control while playing an instrument. Simulation driving games like *Forza* and *Gran Turismo* provide players with

continuous feedback to help them learn the correct response to stimulus given the car they are driving, to help them improve their driving skills. Other examples include any game that improves motor control; Kinect's *Dance Central* improves player's bodily movements, and even first person shooter games like Halo or Call Of Duty improve player's hand-eye coordination.

The Affective domain is one that is far harder for games to address. It deals with learning empathy, emotion, and includes things that are commonly learned through interaction with people and not necessarily through formal education. Games commonly expose players to this in two ways. The first is through game narrative, where players are exposed to characters, situations, and choices where they observe and learn about the human condition. Games with extreme character development (*Mass Effect, The Walking Dead*) typically explore human choices in great detail. There are also extremely minimalistic games that communicate powerful empathic ideas without a large amount of gameplay or graphics, or conversation (*Thomas was alone, Journey*), as well as completely narrative-based, choose-your-own-adventure Twine games that address extremely empathic concepts (*Howling dogs*). The second is through putting players in situations with other players where they will need empathy and skills from the Affective domain in order to solve their problems. All multiplayer games facilitate this to some degree, but puzzle games (*Portal, LittleBigPlanet*), and massively-multiplayer online games (*World of Warcraft, EVE Online*) generally encourage players to communicate, negotiate, and work together to achieve a common goal.

When we think of educational games, we commonly think of the Cognitive domain. The Cognitive domain deals with knowledge of facts, associations, and mechanics, and typically consists of the material that educators attempt to teach in schools. This is also the domain that games attempt to address the most; being able to communicate these concepts effectively would reduce the amount of material educators would need to teach, allowing them to focus on more complex concepts. Because games in the Psychomotor domain are well traversed, and games in the Affective domain are far more nebulous as to measuring their educational value, we will be focusing exclusively on the Cognitive domain.

2.3 Tangential Learning

In their paper called Serious Games and Learning, Breuer and Bente [2] outline a method called Tangential Learning. Originally presented as part of the Extra Credits video series online [7], they give us a method for which we can encourage significant learning using games. The theory is that, given the proper incentive, players will engage in a form of self-education separate from the game environment.

They base this on the principle that we're able to recall and understand material that we're more interested or passionate about, rather than material that is uninteresting or boring to us. The same principle applies to video games; if we are interested in the context of an educational game, we're far more likely to remember that instead of the content being taught as part of it. The example Extra Credits uses is geography; it is easy for some players to draw the entire map of Azeroth from memory, but they are unable to properly identify US states.

The most prominent field that this could be applied to is history; as most games are set

in the past, such as Ancient Rome or World War II, it would be easy for players to learn more about history on their own simply because they were playing the game in the appropriate context. There are other contexts that players can be placed in, in order to educate them about other topics. For example, players in music games can self-educate on music theory and different musical genres, and players in space simulations might self-educate on spaceflight and astronomy.

However, there are numerous other ways players can be educated apart from being placed in the context of the educational goal. An easy-to-implement way would be to include facts as part of the loading screens in the game; instead of having game tips or strategies, designers could include interesting bits of external information to help players retain information. Also, designers can include subtle references to real-life objects within their games. Extra Credits gives the example of Sephiroth from *Final Fantasy*, or the Excalibur and the Masamune. By naming characters or objects after real-life characters or objects, designers can encourage players to go out and seek the origin of the character or object; it also helps to have one object in a group be something that's easily recognizable, so that players immediately know that other objects in the group are referential to the real world.

Another excellent way to integrate tangential learning into games is by including an in-game encyclopedia, where players do not even have to leave the game to explore more about the topic they are interested in. Usually, these encyclopedias include information related to playing the game; for example, the stats of an object, or the rules of the game. However, some include information that isn't related to playing the game, like the history or origin of an object or character. A notable example of this is the game *Civilization*, which includes the Civilopedia, containing background information on every unit, building, and wonder in the game, as well as other information related to it. An easy way for games to integrate this would be just including Wikipedia links to every object in the game that has real world roots.

2.4 Flow

Dondlinger [6] explains why it is extremely important to give players the perception of free will. In educational games, we do not want to give the players no direction as to where to go within the game; we risk players missing the educational goals, and worse, becoming frustrated with the lack of progress. Conversely, we do not want to overly guide the players; too much hand-holding results in the game stifling creativity, as well as not allowing players to really learn by doing. it is important for us to give them the illusion of having an open world, where almost any actions are possible, in order to allow players to experiment and learn by doing.

In parallel to the illusion of free will is the concept of 'flow.' In games, we want to the player to experience just the right difficulty; games that are too hard will be frustrating, and games that are too easy will be boring. The same is true of education and educational games; material that is far beyond the comprehension of the learner won't be retained, but material that the learner has already covered sufficiently will be boring. We need to maintain a level of difficulty to our game that Csikszentmihalyi (1997) calls 'flow.' When players are in a state of 'flow,' the game provides to them a clear set of inputs and outputs; material

that they are somewhat familiar with, but also obscure enough to encourage to players to try several options in order to solve the problem.

Following the concept of 'flow,' we can extrapolate the concept of 'adaptive difficulty.' it is logical to assume that games that continually adapt their difficulty level to the players will be more effective in retaining the player's interest and educating the player than games that have a fixed difficulty curve. However, not every game can implement adaptive difficulty effectively; games need the players to continuously attempt problems and give feedback on how well they are doing to ascertain how well the difficulty matches up to the player.

2.5 Play

Similar to the illusion that players have a seemingly limitless game world to interact with is the concept of 'play.' Paras [13] defines a world where play can happen simply as a world with a series of constraining rules. We do not want the rules to be too constraining, because then we inhibit creativity and discourage alternative solutions to problems, but at the same time, we want the rules to be constraining enough to guide the player towards the solution and mimic the rules of the real world.

The way the game guides players around these rules is extremely important. Players learn the rules of complex systems by exploring them personally, instead of reading tutorials on how the system works. If the game gives the player copious amounts of tutorials and guidance as to what they can or can't do in the game, the player won't learn them as well. This isn't to say that the player should be encouraged to break the game rules, but instead to encourage the player to explore the world's rules; they won't understand what will and won't work unless they encounter it on their own and are enabled to learn from it.

The game also shouldn't penalize the player for exploring the boundaries of the game world, as it would discourage them from trying new things and discovering additional boundaries. However, we still need to encourage players to solve problems within the game world without breaking any boundaries, just to ensure that they understand the world that they are playing in; that's why rewarding the players for not breaking any rules works much better than penalizing the player for breaking rules.

Chapter 3

Rubric

In this chapter I propose a rubric for educational game design, created by researching and synthesizing properties of existing educational games. Games are not required to have all or any of these properties, but the theory is that games with more of these properties will be considered more effective and engaging educational games.

3.1 Location of Game Encyclopedia

Some educational games include a game encyclopedia as part of the game, either internally or externally as a game manual or wiki. Having such an encyclopedia available within the game greatly increases the chance of players using it for self-directed and self-motivated learning, as part of tangential learning.

The rubric to measure the quality of how a game satisfies this attribute is as follows.

1 Game contains no encyclopedia of game content

2 Game information is located online, in a non-central location

3 Game has an outside manual (or wiki) of game content

4 Game doesn't have an in-game encyclopedia, but points to a central location elsewhere

5 Game has an in-game encyclopedia of game content

Examples Examples of games with an encyclopedia with a good location include the Civilization series (for their Civilopedia) and the Total War series. Both of these games contain an in-game encyclopedia that players can access with one click.

3.2 Content of Game Encyclopedia

Having a game encyclopedia as part of the game encourages self-motivated players to seek out additional help. If the encyclopedia includes more than just game mechanics as part of its content, players may be more encouraged to self-educate themselves about those topics.

The rubric to measure the quality of how a game satisfies this attribute is as follows.

1 Game contains no encyclopedia or Encyclopedia contains no content

2 Encyclopedia contains content only related to game mechanics

3 Encyclopedia contains information about game mechanics, and also limited historical/-factual information

4 Encyclopedia contains content related to game mechanics and historical/factual information

5 Encyclopedia contains content related to game mechanics and historical/factual information and outside links or references

Examples Examples of games with an encyclopedia that contains both mechanics and historical/factual information include Civilization and the Total War series, as well as most historical simulation games. All of these games' encyclopedias contain copious amounts of game mechanics information, as well as historical and factual information about each item in the game.

3.3 Amount of Referential Material

By having many objects or events in the game that are references to real-life objects or events, players may recognize certain objects or events and self-motivate themselves to learn more about those objects or events. More objects like that in the game means more chances for a player to recognize a real-life object or event.

The rubric to measure the quality of how a game satisfies this attribute is as follows.

1 No game objects or events are references to real-life objects or events

2 At least one event or object is a reference to a real-world event or object

3 Several non-connected events or objects are references to a real-world event or object

4 At least one group of objects or events are references to real-life objects or events

5 Numerous groups of objects or events are references to real-life objects or events

Examples The Oregon Trail is a game with huge amounts of referential material; just about every item and event in the game is historically accurate.

3.4 Popularity of Referential Material

Having real-life objects or events in the game means that some players will recognize them and attempt to self-educate themselves about those objects or events. In addition, having a large amount of those references be popular means that more players will recognize the more popular objects/events, and by association, education about the lesser-known objects/events will be the result of self-motivated education about the popular items.

The rubric to measure the quality of how a game satisfies this attribute is as follows.

1 Existing references are extremely obscure

2 Some of the references are popular

3 About half of the references are popular

4 Many of the references are popular

5 Most or all of the references are popular

Examples Where in the World is Carmen Sandiego is a game with many popular references. In contrast to The Oregon Trail, where all of the content is specific to that time period and may be totally unfamiliar to players, Carmen Sandiego's wide range of concepts (geography, landmarks, cultures, religions, etc) is likely to be recognized by many different players.

3.5 Rewards for knowing Referential Material

If games allow players to use their knowledge of the referential material in a positive manner within the game, it reinforces the desire for the player to tangentially learn about all the referential material in the game. This is in contrast to traditional methods, where the content that the player is trying to be educated about serves as a barrier of entry to the later levels of the game, or the game penalizes players that don't know the content; with this, players can still play the entirety of the game, but are incentivized to do better by learning the referential material.

The rubric to measure the quality of how a game satisfies this attribute is as follows.

1 Knowing the referential material is purely irrelevant; doesn't affect the gameplay at all

2 Knowing the referential material is a little useful; only affects the gameplay a small amount

3 Knowing the referential material is somewhat useful; moderately affects the player's choices during gameplay

4 Knowing the referential material is very useful; usually affects the player's choices

5 Knowing the referential material always significantly affects gameplay

Examples In Kerbal Space Program, players aren't taught about aerodynamics, power-to-weight ratios, or how to transfer orbits. However, if they do know and understand that material, building, launching, and maneuvering an orbital craft becomes much easier.

3.6 Adaptive Difficulty

In order to incorporate flow as much as possible, the game must use a form of Adaptive Difficulty; the game tries to match the difficulty of the obstacles to the skill level of the player. This can be done in a rudimentary way by allowing the player to select the difficulty themselves, but is most effective when it recognizes how well a player is doing on a certain puzzle and adjusting the difficulty automatically. This ensures that the player isn't trivially challenged, but at the same time isn't pushed outside the bounds of their ability.

The rubric to measure the quality of how a game satisfies this attribute is as follows.

1 Game only has one difficulty

2 Game has several difficulties, but players can only select difficulty at beginning of game

3 Game has several difficulties, and players can change difficulty mid-game

4 Game has several difficulties, and prompts the player to increase or decrease the difficulty as needed

5 Game has several difficulties, and automatically adjusts the difficulty of the game as needed

Examples In The Logical Journey of the Zoombinis, players must complete challenges with 16 zoombinis. If the player passes the level, they continue with all their Zoombinis. If the player does poorly on the level, they can still pass the level; they just lose Zoombinis depending on how poorly they did. A lost zoombini returns to the previous checkpoint (spaced every 3 levels), and players can only embark from a checkpoint with 16 Zoombinis. If the player has a 100

Chapter 4

Games

4.1 Game Selection

In order to test our rubric, we chose to have a group of people play some educational games, and have the players apply the rubric to the games. Ideally, the players should score the games similarly. The players were solicited using Amazon's Mechanical Turk, as described in detail in the next chapter. The games were chosen according to the following criteria.

We looked for games marketed as educational games, as well as games that could have been considered educational by including tangential learning. We selected the games from existing knowledge, as well as from references in existing literature.

4.1.1 Web games

Due to the End User License Agreement (EULA) of Mechanical Turk, we could not ask Workers to install any software on their machines as part of our task. This meant that we couldn't select some popular educational games, like Typing of the Dead, because users would have to install the game on their machine.

An option that was considered was having Workers find a YouTube or other video of the game being played, and try to evaluate the game's properties based on the gameplay footage. However, we decided that Workers wouldn't be able to evaluate some of the game's properties (e.g. penalties for failure, or contextual tutorials) based on a gameplay video.

Thus, our primary selection criteria for educational games was that they had to be web-based. The players could not be required to install any software onto their machine in order to play the game. This left us with many interesting and high-quality options, like *Light Bot* and *Darfur is Dying*, but it eliminated some other high-quality options like *Logical Journey of the Zoombinis*.

It is important to note that we were allowed (or at least, did not run into any problems) to ask users to install web browser plugins in order to play games. Some of these plugins were commonplace (like the Java runtime), but others were obscure (like the Apple II emulator). This allowed us to find some popular educational games usually only available as a desktop applications, like *The Oregon Trail* or *The Incredible Machine*, even if they were very old versions of the games.

4.1.2 Math Blaster

One of the most prominent games that we were unable to include in our survey was the *Math Blaster* series. *Math Blaster's The Search For Spot*, released in 1994, was the earliest version of the game, and the franchise continues strong to this day. More editions and expansions were released, all as installable desktop applications.

Recently, the *Math Blaster* franchise has transitioned into a web-based format, where players can play various Math Blaster games online. This would have been perfect for our study, but the games exist behind a 'login wall,' where players must create accounts and verify their email addresses before they are allowed to play the games. We assume this violates the EULA of Mechanical Turk, but even if it does not, the login wall presents such an accessibility obstacle for Workers that we decided to omit Math Blaster altogether.

4.1.3 Where in the World is Carmen Sandiego?

Another extremely popular and highly-acclaimed educational game series, *Where in the World is Carmen Sandiego*, started releasing desktop games in 1985, but has since ceased producing sequels. Online, we found numerous emulators for the original game, but the performance and framerate for each of them was very poor. We decided that the average Worker would find it unplayable, and decided to omit it from the study.

4.1.4 Old Games

It is important to note that when we think of good, educational games, we commonly think of classic games; games that were created early in the lifecycle of educational games, and were highly acclaimed for their educational content. Recently, there haven't been many games that have been popularized for their educational content.

Why is that? there is a couple ways that we could explain the lack of popular educational games. It is possible that there exist as many or a greater number of excellent educational games presently, but due to saturation of the gaming industry (both as entertainment and educational games), educational games do not get as much visibility as they do before. For example, the most popular recent game included in our survey, *Darfur is Dying*, had over 800,000 plays in 6 months and was covered in the media, but received no 'educational game' awards.

Another option is that the nature of education in gaming is changing. Instead of players seeking deliberately edcuational games, where education is the primary goal of the game, players may be more interested in games where education is the secondary goal. Games with education as the secondary goal most likely use tangential learning in order to teach players without deliberately lecturing or quizzing them. *Light Bot*, one of the games in our survey, is a great example of this.

4.2 Game Listings

4.2.1 Darfur is Dying

URL `http://www.darfurisdying.com/`

Description Darfur is Dying, developed by mtvU, is a activism game released in April of 2006. The game consists of two main sections. In the first section, players must select a member of a Darfuri refugee; the family consists of a male, female, and several children. Once the player has chosen, the player must guide the refugee to a well using only compass and distance directions, while attempting to avoid Janjaweed militia patrols in trucks. If the player is caught, the game describes the fate of the refugee, and the player is prompted to select another refugee. Once the player has made it to the well, the player must return through the same section to their encampment, but lose water during their journey. Once they have returned to the encampment, the player enters a top-down strategy-like game simulation; they can use the water they have retrieved to grow crops and keep the encampment in good condition. However, if they run out of water, they will need to return to the first section and make the well run again. The goal of the game is to keep the encampment alive for 7 days. In addition, the community is constantly under threat from attacks from the militia; if the militia attacks, the encampment is lost, and the player must start again. The player can prevent attacks from the militia by participating in various viral and advocacy campaign tactics, such as inviting their friends to play, posting on social media, or writing to government officials.

Educational Content The educational content of Darfur is Dying is centered almost entirely around awareness. Through playing the game, players learn about the nation of Darfur, including its history, wars, environment, climate, and people. The players are forced to be aware of the troubles that plight Darfuri residents and families, such as militia attacks. Players are encouraged to read the backstories of every man, woman, and child, as well as stories associated with locations within the village that expose various events that have happened, such as being unable to fend off sickness without medical aid, or when a militia recently stormed the village and murdered numerous people.

Notes on Implementation Darfur is Dying was built in Flash by interactive media agency interFUEL.

4.2.2 Light Bot

URL `http://armorgames.com/play/2205/light-bot`

Description Light Bot, made by Danny Yaroslavski, is a programming and robotics puzzle game. It was originally a flash game, but has since been ported to iOS and Android. Players assume control of a robot on a grid of varying sizes and orientations. Each grid square can also have a height. The robot has the ability to move forward, turn left or right,

Figure 4.1: Darfur is Dying's title screen

Figure 4.2: A screenshot from Darfur Is Dying's village management section

jump up one level or down one level, and turn a square "on." The robot also has the ability to call "functions," where the robot can execute sequences of events and repeat functions several times or indefinitely. The goal of the robot is to navigate to all of the blue squares and turn them "on" to a yellow state. The robot can do this in any order and using any sequence they like, so long as it fits within the provided instruction spaces. There are 40 levels to the game, ranging from the simple to extremely difficult.

Educational Content Light Bot's educational purpose focuses on teaching programming at a very simple level. Players learn that the bot will follow sets of instructions. Initially, these instructions will be very simple (e.g. forward, turn, blink), but the player will realize quickly that the bot will follow the instructions explicitly, even if they do not solve the puzzle. This teaches players that computers are very powerful but very simple machines, and will do exactly what they are directed to do, even if it's not what the programmer intends to do. The game also teaches the concepts of functions and loops; players can "call" predefined functions numerous times, as well as have a function call itself to loop the function indefinitely. These programming concepts, while simple, are a wonderful introduction to programming for students.

Notes on Implementation Light Bot was built in Flash by Danny Yaroslavski, and the iOS and Android versions were built using Haxe 3 and OpenFL.

4.2.3 Pandemic 2

URL `http://www.crazymonkeygames.com/Pandemic-2.html#game`

Description Pandemic 2 is a flash-based strategy game involving infectious diseases, viruses, and bacteria. The player is in charge of designing and mutating an infectious organism, which infects the world population. The objective is to have the infection spread to and kill every human being on the planet, rendering the human race extinct. The virus starts out as being only mildly visible, lethal, and infectious, and can be mutated to more effective versions through "upgrades," received as more humans are infected and die. To combat the spread of the infection, world nations begin to close their borders, set up quarantines, and close off trade routes, cutting off the transmission of the infection to their nation and making it more difficult or nearly impossible for the disease to spread. The player typically alters between the disease upgrade screen and the world monitoring screen, which includes notable headlines and the statuses of the nations. Global high scores are given to players that successfully eliminate the human race in the shortest amount of time.

Educational Content Pandemic has two limited educational aspects to it. The first is the notion of learning about infectious diseases and organisms. While there isn't much science within the game behind mutating an organism to be more deadly, there are plenty of terms and game mechanics that the player can familiarize themselves with, such as organism's resistance to humidity, or how airborne diseases differ from waterborne. There's also an element of strategizing, risk-taking, and planning ahead associated with playing the game;

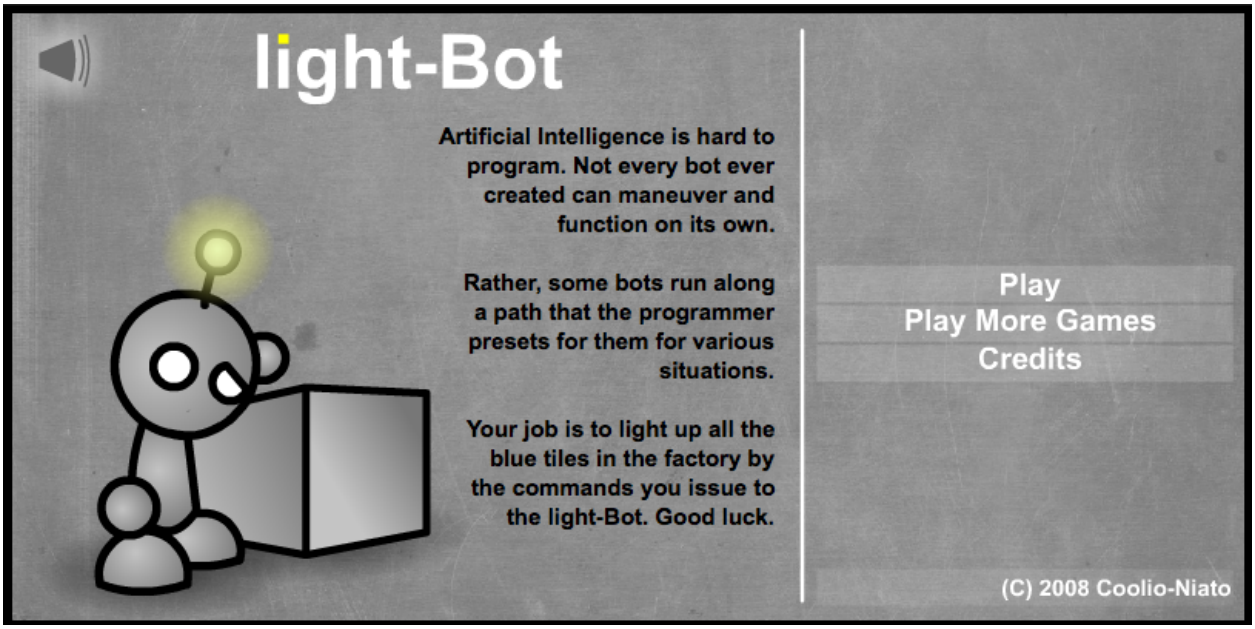

Figure 4.3: Light Bot's main menu

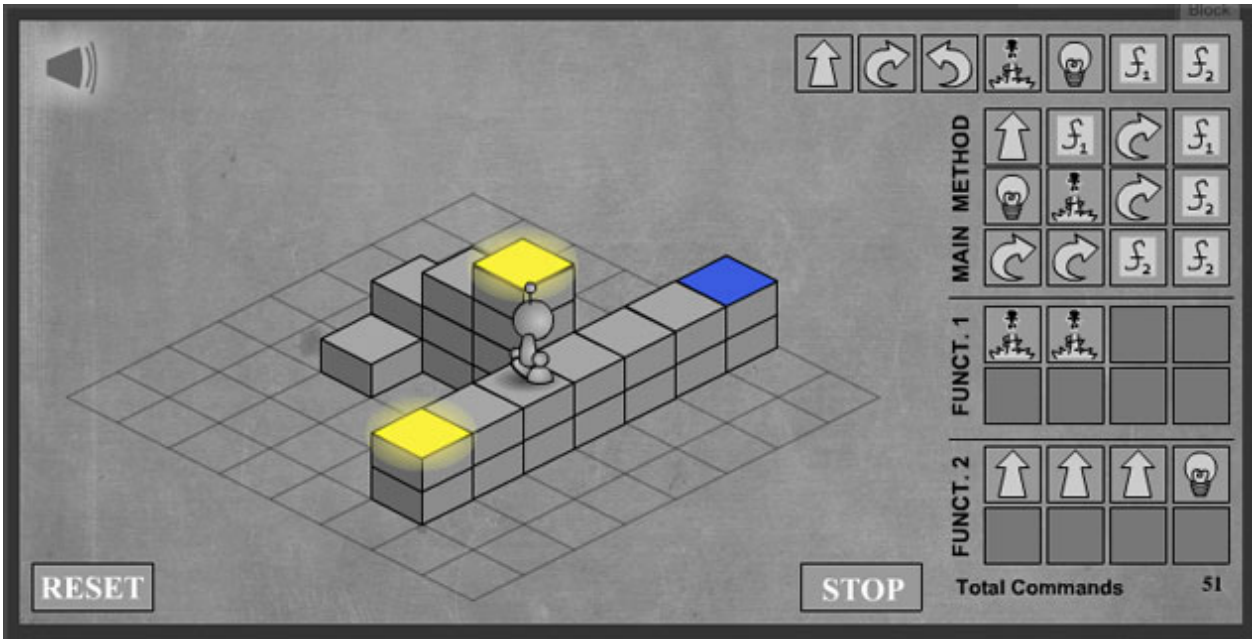

Figure 4.4: A screenshot from one of Light Bot's levels

if players run into a problem with a certain method, they may be able to solve the problem a different way, or circumvent the problem on another playthrough.

Notes on Implementation Pandemic 2 was built in Flash by Dark Realm Studios.

4.2.4 Oregon Trail

URL `http://www.virtualapple.org/oregontraildisk.html`

Description The Oregon Trail is a series of games detailing the experience of a pioneers traveling the Oregon Trail, a wagon route from the Missouri River to Oregon in 1848. The original game, developed for the HP 2100 in December of 1971, turned out to be extremely well received by middle and high school students. In the game, the player first selects an identity which determines how much starting money they have. They then select supplies to buy for their journey; wagons, spare parts, oxen, food, and other supplies. Once they embark on their journey, time passes and food is automatically consumed. Occasionally, players have the chance to hunt, where they can play a simple top-down 2D game to shoot animals that provide food to their party. In addition, numerous events will happen that the player needs to make descisions for, such as a party member falling ill, a wagon part breaking or oxen becoming injured, or needing to cross a river. The player's objective is to travel the entire Oregon Trail with using the minimum amount of resources; the player receives more points at the end for having living party members, items in inventory, and number of dollars, as well as receiving a multiplier if they started the game with less cash.

Educational Content Oregon Trail's educational value comes in two forms. The first is the most apparent one; though not explicitly sitting players down and teaching them, the game educates students on life in the 19th century, as well as the hardships and trials endured by explorers of the early Oregon Trail. It teaches them about the kinds of materials that were used in everyday life, like wagons and oxen, as well as the diseases that commonly plagued explorers (diarrhea, dysentery), and what day-to-day activities were like on the trails, such as maintaining the wagons, crossing rivers, and hunting for food. The other educational aspect that Oregon Trail focuses on is planning and risk management; though not explicitly teaching players how to assess the risk of various actions, players who properly evaluate their initial inventory options as well as the options they have during events on the trail will end up doing better than players who don't.

Notes on Implementation This is the very first edition of The Oregon Trail, for the Apple II. It's played in a browser-based Apple II emulator.

4.2.5 Number Munchers

URL `http://wallofgame.com/free-online-games/arcade/988/Number_Munchers.html`

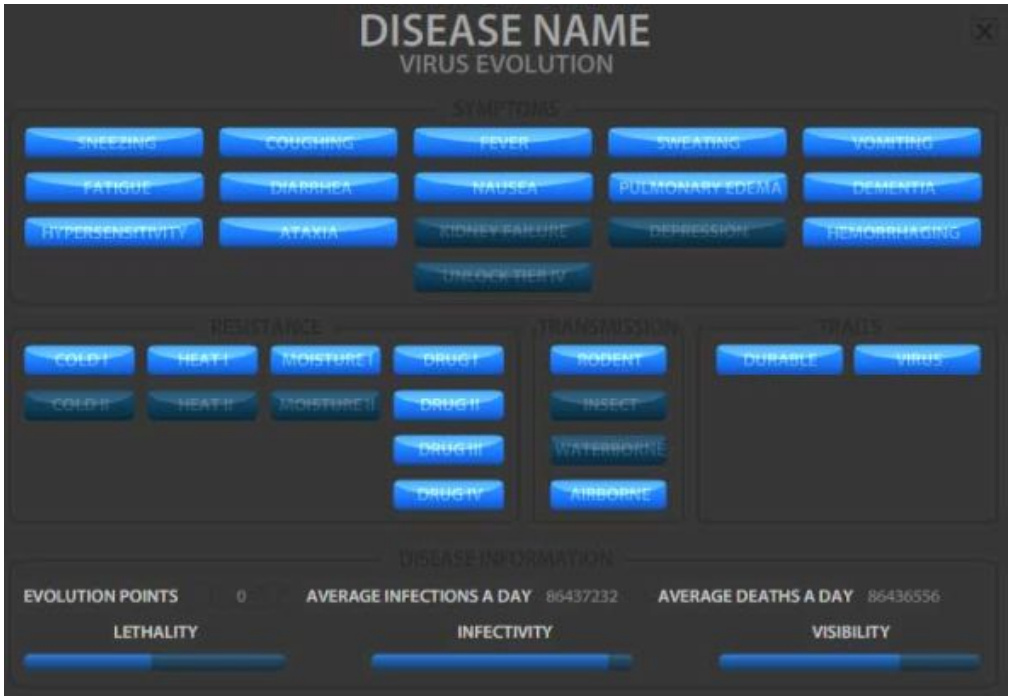

Figure 4.5: A screenshot from Pandemic 2's evolution view

Figure 4.6: A screenshot from the opening sequence of Oregon Trail

Description Number Munchers is part of a "Munchers" series that began as a strictly educational game around teaching children basic arithmetic and other numerical concepts. The series then expanded to teach words and other knowledge. The player first selects the educational content that they'd like to play, including grade level. Administrative controls are available for parents and teachers to restrict what kind of content is available; for example, restricting higher-grade students to their level, instead of allowing them to play at trivial levels. Once in the game, the player assumes control of a green 'Muncher,' which can move up, down, left or right. The game board is a grid, with each space containing a number that is a solution to the level they're playing (e.g. equal to 12, multiples of 3, divisible by 5). The player moves onto the appropriate space, and presses the space key to 'eat' the number. If the number matches the level criteria, the player gets points; if it does not, the player loses one of three lives. The player also has to deal with 'troggles,' which cross the grid slowly at random intervals. If the player occupies the same space as a troggle, the player loses a life and is reset to a different grid space. The level ends once all the appropriate numbers are eaten, and the player recieves a bonus for time.

Educational Content The educational content of Number Munchers is extremely straightforward. Players will learn arithmetic, primes, and other numerical concepts while playing this game. Players who learn the concepts will be able to identify the correct solutions faster, and consequently achieve a higher score while playing this game.

Notes on Implementation This is not an official version of Number Munchers. It's a Flash-based, Number Munchers-inspired game created by user Authorblues. It's functionally equivalent to the original edition of Number Munchers.

4.2.6 BotLogic

URL `http://botlogic.us/`

Description BotLogic is an educational programming game. In the game, the players assume control of a robot whose task is to return home. The game takes place on a grid, with the robot located on one space, the home on another space some distance away, and a number of obstacles in between. The player can direct the robot up, down, left, or right, and does so by queueing up commands before running the program. The player can wait for the program to run, then add more commands before running the rest of the program, which allows players to incrementally build their program. However, the robot has a limited amount of energy, which limits the number of moves the player can take. Later in the game, more obstacles and powerups are introduced, such as electric fences, buttons, and recharging stations. The game contains 20 levels.

Educational Content Botlogic teaches players about simple programming concepts, namely passing a sequence of instructions to a robot and watching them run. However, the game doesn't include any functional or object-oriented programming concepts, and is only slightly abstracted away from the player having direct directional control of the robot.

Figure 4.7: A screenshot from the gameplay of Oregon Trail

Figure 4.8: A screenshot from the gameplay of Number Munchers

Notes on Implementation BotLogic is built in HTML5/Javascript and developed by Dolphin Micro, a web development consulting firm located in Colorado.

4.2.7 Math Baseball

URL `http://www.funbrain.com/math/index.html`

Description Math Baseball is an educational math game. Initially, the player selects what type of arithmetic they'd like to do, as well as the grade level. Then, they assume the role of a baseball player at bat. For each throw, the players are given unlimited time to answer one arithmetic question. If they get the question right, the player earns a randomly selected single, double, or triple and gets a runner on base, as well as advancing any other runners. The player's score is determined by how many runs they get in. If they don't get the question right, then they recieve a strike. After three strikes, the player is "out," and the game ends.

Educational Content The education in "Math Baseball" is very straightforward. Players can learn addition, subtraction, multiplication, and division, with the player's knowledge of the topic reflected in their higher score of the game.

Notes on Implementation Math Baseball is written in HTML and an unspecified server framework.

4.2.8 Notpron

URL `http://notpron.org/notpron/`

Description Notpron is an ARG (alternate-reality game) for the browser. The player begins Level 1 with an image of a house with a partially open door in front, as well as some slightly opaque text that says "Enter the door." The player needs to click the door (not the image, but the door itself) to advance to Level 2. In Level 2, a finger points to the address bar, where the player can replace "level2.htm" with "level3.htm" to advance to the next level. In Level 3, the player must change "false" in the URL to "true" to advance to the next level. The game continues like this, adding in new elements each level. There are a total of 140 levels, and only 31 people have completed all 140 levels, out of about 16 million players.

Educational Content From the Notpron site:

> "[Players who finish the game] have persisted with a broad range of complex ways of thinking, while maintaining focus and dedication over a long period. [Their] detective skills have been tested to the limits, yet the smallest hint proved sufficient to solve the most complicated tasks. Furthermore, competence in the following areas have been displayed: Sound editing, Graphic editing, Musical understanding, Insight into HTML programming, Rapid learning of new programs,

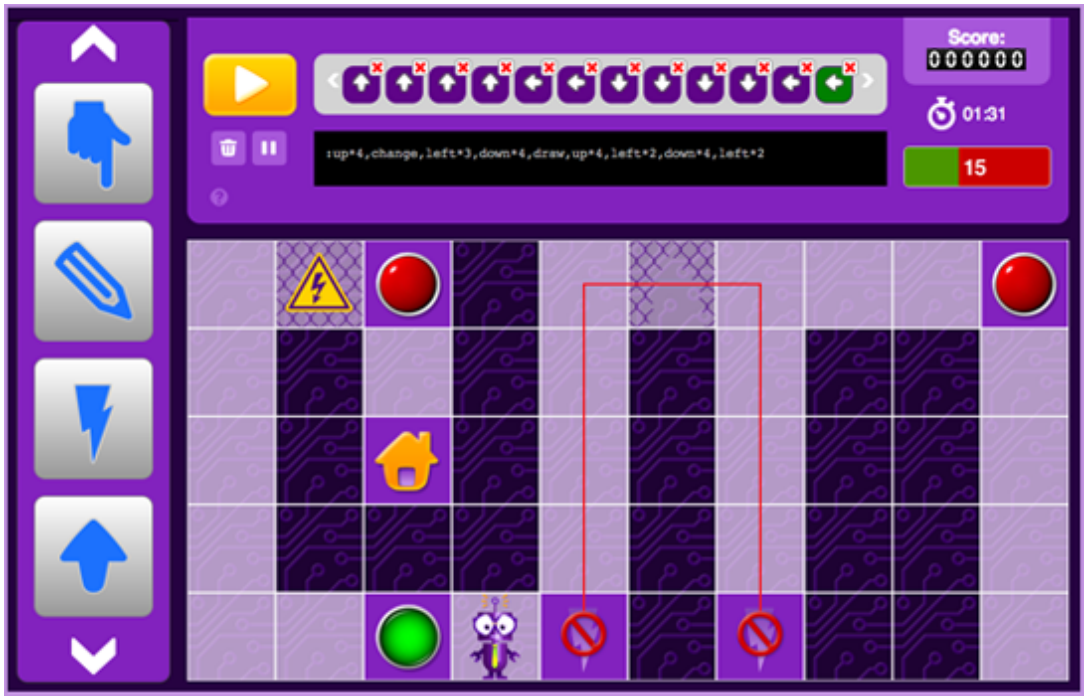

Figure 4.9: One of the levels from BotLogic

Figure 4.10: A screenshot from the gameplay of Math Baseball

Efficient online research techniques, [and] Insight into the complex workings of a computer."

Notes on Implementation Notpron is written in HTML and an unspecified server framework.

4.2.9 Lemmings

URL `http://www.elizium.nu/scripts/lemmings/`

Description Lemmings, a PC game, was originally developed in 1991. The game plays from a 2D side-scrolling perspective. The player directs "lemmings," small, humanoid creatures that are reminiscient of the mammal in their behavior. They operate almost entirely on their own, walking in one direction until they run into something, then reversing direction. They begin by dropping out of the entrance door, and successfully exit the level through another door in the level. However, the lemmings are very susceptible to dying; falling too far without a parachute will kill them, and numerous obstacles litter the courses, such as spike pits and smashers. The game contains four difficulty levels, with roughly 20 levels per difficulty level.

Educational Content Lemmings teaches players problem-solving, multitasking, and resource management. Because a fixed number of lemmings are required to finish the level, players must learn the proper methods for getting around obstacles by using the minimum number of lemmings possible.

Notes on Implementation This implementation of Lemmings is a Javascript port of the original version of the game.

4.2.10 The Incredible Machine

URL `http://www.classicdosgames.com/online/timdemo.html`

Description The Incredible Machine, originally developed in 1993, is a side-view 2D construction game. During a level, the player will have an objective (e.g. get the ball into the basket). The play area will already have some parts set up, so the player can use the parts that they have in reserve to construct the rest of the Rube Goldberg-style machine to accomplish the objective. There were around 80 levels in the game. The game was extremely successful, and spawned numerous sequels and ports.

Figure 4.11: The first level of Notpron

Figure 4.12: One of the levels from Lemmings

Educational Content The Incredible Machine teaches players about physics and problem solving. Players learn about the physical properties of various objects (for example, the tennis ball might not knock down the board, but the bowling ball might), and learn how to use limited combinations of those objects together to solve the puzzles.

Notes on Implementation This is the original edition of The Incredible Machine, run inside a Java container in the browser.

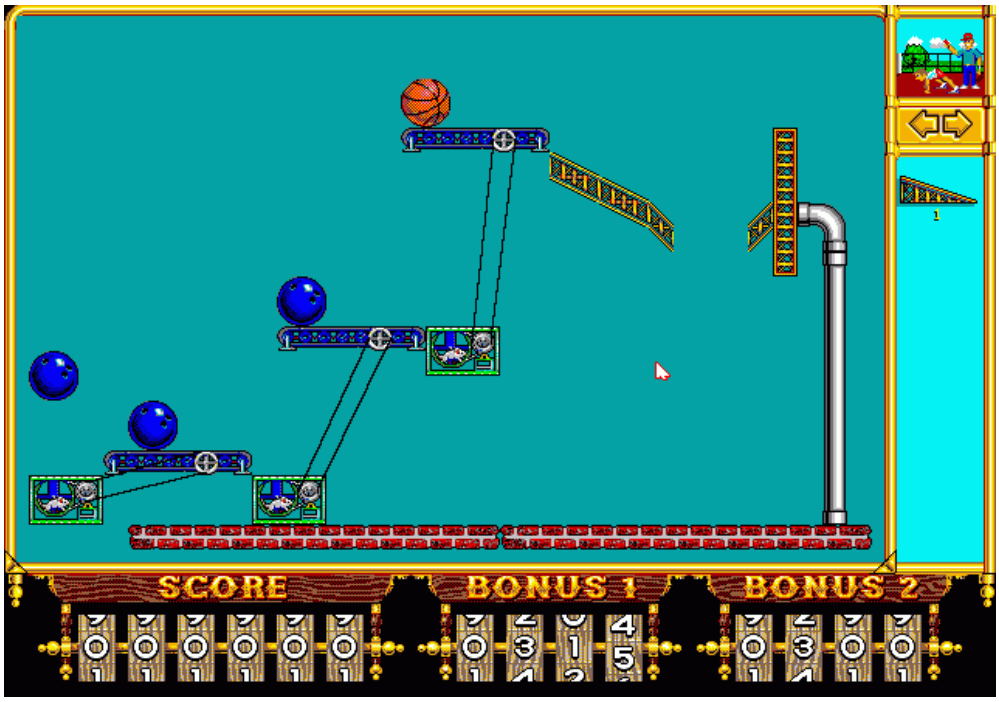

Figure 4.13: One of the levels from The Incredible Machine

Chapter 5

Procedure

5.1 Survey Design

In order to test the effectiveness of the rubric on determining good educational game design elements, it must be used by real players on real educational games. To get the best results possible, it would need to be evaluated by a large number of players across a variety of games.

Our procedure, then, would be very simple. A player would need to play a game for a short period of time, then fill out the survey, indicating what elements the game had. There's no need for additional feedback or contextual observation of the player as they play the game. This means that we can administer our survey remotely, through online quiz or survey-taking systems.

5.2 Mechanical Turk

I decided on Amazon's Mechanical Turk to adminster my surveys. Mechanical Turk is an online marketplace where Requesters (the people who want results from surveys or tasks) can submit HITs (Human Intelligence Tasks) into the marketplace, where Workers (people who fill out surveys or perform other tasks) can complete them, usually receiving some kind of small monetary reward for them. It's commonly used as a survey-taking platform, but Requesters also use it for tasks that computers are not yet good at doing, like determining the content of an image, writing a summary of an article, or finding the website of a lesser-known business.

In order to only allow a worker to take a survey once, I had to set up 10 different HITs, one for each game that we were testing. I also allocated a number of responses that I wanted to get for each HIT; this ensured that a worker would only be allowed to fill out a survey once, but was allowed to complete survey for all of the games if they wanted. I allocated 20 responses for each game, for a total of 200 survey responses. The initial expiration date of the HITs was 3 days later, but I extended the deadline a week.

For 6 HITs, the HIT was composed of the "Basic" survey included below. It

contained some simple demographic questions, some Likert-scale questions about their opinions on fun and educational games, and some freeform text response areas where they could leave feedback about the games and survey. It also included a link to one of 6 games (The Incredible Machine, Lemmings, Notpron, Math Baseball, BotLogic, or Pandemic 2), as well as a section with all of the rubric items, where players were asked to rate each game on all of the rubric items.

The other 4 HITs included a pre- and post-quiz in addition to the basic survey. The 4 games (Darfur is Dying, The Oregon Trail, Number Munchers, and LightBot) each had their own quiz, designed to test the knowledge that the games were intended to teach. The quizzes were 10 questions each, with either 4 or 5 multiple choice answers, or True/False answers. It's important to note that the pre- and post-quizzes were the exact same; the exact same questions and answers appeared before and after the players played and rated the game. The players were not given feedback on how well they did on either of the quizzes. Each quiz is included below.

5.3 Implementation

First, the information for the quizzes, surveys, rubric, and games were created in a Javascript Object Notation (JSON) format. From this format, the final surveys and quizzes specific to each game can be created, while also generating some LaTeXdocuments with the exact same format.

Then, the surveys are deployed to the Mechanical Turk platform. This involves authentication with my personal account, and setting up tasks with parameter specific to my research.

After the surveys are completed, they are retrieved and placed in a data directory. Then, the analysis script is run.

The analysis script first reads all of the files present in the data directory into a SQLite database stored in memory. It then grades all the quizzes, and uses PyPlot to generate the visualization graphs. It performs a similar task to generate the rubric and game score visualizations. Then, it calculates the t-distribution significance and inter-rater reliability scores, and graphs them. Finally, it writes the graphs to files.

Then, the LaTeXcompilation script is run, and all of the generated tex files, graphs, and other information is processed and placed into the final thesis document.

All of these instructions are placed in the Makefile, so that if more data is received, we only need to run make in order to generate new graphs and include them in the thesis document.

A total of about 2000 lines of code was written. All of the source code for the thesis can be found at `https://github.com/kenhoff/thesis`.

5.4 Survey

This is the survey that was given to the Mechanical Turk workers. It includes some demographic questions, followed by some questions on the worker's opinions on fun and educational games. Then, the worker is given the URL to one of the ten games. Then, the worker is asked to rate the game on the 13 rubric items, selecting one of the 5 options from each. After that, the worker provides more opinions on the game they just played.

5.4.1 Basic Survey

What is your gender?

- Male
- Female

What is your age?

(An empty text box where workers can only enter numbers)

How many years have you been playing video games?

(An empty text box where workers can only enter numbers)

Fun games can be educational.

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

Most fun games are educational.

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

Games are more fun when they are cooperative (instead of competitive).

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

Games are more fun when you play by yourself (instead of with others).

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

Games are more fun when you play online (instead of with others in the same room).

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

Educational games can be fun.

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

Most educational games are fun.

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

Games are more educational when they are cooperative (instead of competitive).

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

Games are more educational when you play by yourself (instead of with others).

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

Games are more educational when you play online (instead of with others in the same room).

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

List some fun games.

(An empty text box where workers can write anything)

List some educational games.

(An empty text box where workers can write anything)

5.4.2 Rubric

Does this game have an encyclopedia of game content? If so, where?

1. Game contains no encyclopedia of game content
2. Game information is located online, in a non-central location
3. Game has an outside manual (or wiki) of game content
4. Game doesn't have an in-game encyclopedia, but points to a central location elsewhere
5. Game has an in-game encyclopedia of game content

If this game contains an encyclopedia of game content, what kind of content does it contain?

1. Game contains no encyclopedia or Encyclopedia contains no content
2. Encyclopedia contains content only related to game mechanics
3. Encyclopedia contains information about game mechanics, and also limited historical/factual information

4. Encyclopedia contains content related to game mechanics and historical/-factual information
5. Encyclopedia contains content related to game mechanics and historical/-factual information and outside links or references

How many game objects or events reference real-life objects or events?

1. No game objects or events are references to real-life objects or events
2. At least one event or object is a reference to a real-world event or object
3. Several non-connected events or objects are references to a real-world event or object
4. At least one group of objects or events are references to real-life objects or events
5. Numerous groups of objects or events are references to real-life objects or events

Of the objects/events that are in the game, how recognizable/popular are they?

1. Existing references are extremely obscure
2. Some of the references are popular
3. About half of the references are popular
4. Many of the references are popular
5. Most or all of the references are popular

How much does knowing the real-life objects/events affect the gameplay?

1. Knowing the referential material is purely irrelevant; doesn't affect the gameplay at all
2. Knowing the referential material is a little useful; only affects the gameplay a small amount
3. Knowing the referential material is somewhat useful; moderately affects the player's choices during gameplay
4. Knowing the referential material is very useful; usually affects the player's choices
5. Knowing the referential material always significantly affects gameplay

How many difficulty levels are there in the game, and how is the difficulty changed?

1. Game only has one difficulty
2. Game has several difficulties, but players can only select difficulty at beginning of game
3. Game has several difficulties, and players can change difficulty mid-game
4. Game has several difficulties, and prompts the player to increase or decrease the difficulty as needed
5. Game has several difficulties, and automatically adjusts the difficulty of the game as needed

How often are tutorials offered in the game?

1. Tutorials aren't given in the game
2. Tutorials are given at the beginning of the game
3. Tutorials are given every few levels/sections of the game
4. Tutorials are given at the beginning of every level/section
5. Tutorials are offered continuously

If the player fails, how many game resources do they lose?

1. Game resource penalty is extreme (e.g. restart the game)
2. Game resource penalty is large (e.g. 1/3 lives, 10 health points / 100)
3. Game resource penalty is moderate
4. Game resrouce penalty is small
5. Game resource penalty is nonexistent (e.g. unlimited lives)

If the player fails, how long is the wait until they restart play?

1. Reset time is very long (e.g. reloading level takes a long time)
2. Reset time is long (e.g. greater than 10 seconds)
3. Reset time is moderate (e.g. up to 10 seconds)
4. Reset time is short (e.g. a few seconds)
5. Reset time is very short (e.g. near instant)

How often are there in-game checkpoints?

1. Zero checkpoints
2. Checkpoints are few and far between (e.g. levels are the only places to restart)

3. Checkpoints are moderate
4. Checkpoints are numerous (e.g. players can restart at the beginning of each puzzle)
5. Checkpoints are frequent (e.g. players can restart part of the way through puzzles)

How much freedom of choice is there in the game, including both game world and choice of lessons?

1. Players are placed in a strictly linear world or lesson progression
2. Players are allowed just a few large-scale choices in their game world
3. Players have the option to make choices about the direction of their progression in the world, but it is largely linear
4. Players can choose from many choices within the game world to explore, including lessons
5. Players are free to choose the direction they want, both educationally and within the game world; allowed to jump between parallel lessons

How much and how often is there feedback regarding progress on the game?

1. Game gives no feedback other than high-level progression through the game
2. Game gives feedback after each level
3. Game gives feedback at various points through a level, after a series of puzzles
4. Game gives feedback after each puzzle
5. Game constantly gives feedback (e.g. during a puzzle)

Are problems able to be solved multiple ways, or circumvented entirely?

1. There is only one way to solve any given problem, with one given progression that is valid as a solution
2. Some problems within the game may be solved more than one way
3. Multiple solutions are available for each problem, but players are limited to using one of those solutions
4. Players can solve each problem via any solution, so long as they do not circumvent it
5. Players can solve a problem any way they like, or even circumvent the problem, and be given full (or bonus) points

This game was fun.

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

I had fun playing this game.

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

This game was educational.

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

I learned something from playing this game.

- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

What did you learn from playing this game?

(An empty text box where workers can write anything)

Provide some comments and feedback on the game.

(An empty text box where workers can write anything)

Provide some comments and feedback on this survey.

(An empty text box where workers can write anything)

Estimate the number of minutes that you've played this game.

(An empty text box where workers can only enter numbers)

Estimate your highest score/level from your play session.

(An empty text box where workers can only enter numbers)

5.5 Quiz

These are the quizzes that were given to the Mechanical Turk workers. There are 4 quizzes, one for each game that was selected as a quiz game. The workers are given the corresponding quiz both before and after the 'Survey' section; both the before and after quizzes are the exact same.

The quizzes are each composed of 10 multiple choice questions. The questions were selected by first examining the reported educational goals of the games, then synthesizing the quizzes using existing online surveys.

5.5.1 Darfur is Dying Quiz

When did the violence in Darfur first start?

1. 2001
2. 2002
3. 2003
4. 2004
5. 2005

Who is the fighting between?

1. two different tribes
2. arabs and non-arabs
3. Hutu and Tutsi
4. Blacks and Whites
5. rich and poor

A genocide is defined as killing, causing serious bodily or mental harm and preventing births of ...

1. a specific national group
2. a specific ethnical group
3. a specific racial group
4. a specific religious group
5. all of the above

How did the conflict start?

1. bombing in a main city
2. government's drafting of soldiers
3. accidental killing of 5 civilians by police
4. rebellion attacks of government targets
5. public protest against new law

How many people have died as a result of the violence, hunger, and disease?

1. up to 3,000
2. up to 30,000
3. up to 300,000
4. up to 3,000,000
5. up to 3.5 milion

Where is Darfur?

1. northern Sudan
2. southern Sudan
3. eastern Sudan
4. western Sudan
5. central Sudan

When was the height of the violence in Darfur?

1. 2003-2004
2. 2003-2005
3. 2004-2005
4. 2004-2006
5. 2005-2006

What group is referred to as Devils on horseback?

1. Sunnis
2. Shiite
3. Tumoils
4. Janjaweed

Darfur is a region in the African country of Sudan. Which countries border Sudan?

1. Rwanda, Somalia, Tanzania
2. Algeria, Morocco, Tunisia
3. Central African Republic, Chad, Libya
4. Kenya, Ethiopia, Malawi

Which two major powers united with the Sudanese government to keep the United Nations out of Darfur?

1. United States and Canada
2. Iraq and Iran
3. Egypt and Libya
4. Russia and China

5.5.2 The Oregon Trail Quiz

How many miles long was the Oregon Trail?

1. 500 Miles
2. 1000 Miles
3. 2000 Miles
4. 3000 Miles
5. 20000 Miles

What is the name of a disease like malaria that the pioneers might catch?

1. ague
2. pneumonia
3. epilepsy
4. measles

What year was the first Oregon Trail wagon train organized?

1. 1847
2. 1843
3. 1899
4. 1876
5. 1836

What disease killed more people on the trail than any other?

1. smallpox
2. plague
3. cholera
4. scarlet fever

When was the first transcontinental railroad finished that eventually ended the Oregon Trail?

1. 1867
2. 1870
3. 1899
4. 1869

How many modern states did the travelers travel through when crossing the trail?

1. 6
2. 8
3. 3
4. 10

How many people died on the Oregon Trail?

1. 50,000-60,000
2. 20,000-30,000
3. 90,000-100,000
4. 10,000-20,000

Where did the Oregon Trail begin?

1. Independence, Mississippi
2. Independence, Missouri
3. Independence, Michigan
4. Independence, Montana
5. Independence, Massachusetts

Where did the Oregon Trail end?

1. Vancouver, Washington
2. The Dalles, Oregon
3. Portland, Oregon
4. Stevenson, Washington
5. Oregon City, Oregon

How many people came west on the Oregon Trail?

1. around 500
2. around 5000
3. around 50000
4. around 500000
5. around 5000000

5.5.3 Number Munchers Quiz

39 + 33 =

1. 67
2. 87
3. 73
4. 81
5. 72

15 + 82 =

1. 112
2. 106
3. 97
4. 81
5. 113

-77 - -94 =

1. 14
2. 36
3. 35
4. 17
5. 10

68 - 92 =

1. -8
2. -15
3. -24
4. -29
5. -38

18 * 9 =

1. 172
2. 162
3. 150
4. 173
5. 157

15 * 8 =

1. 121
2. 131
3. 122
4. 120
5. 119

14 * 6 =

1. 88
2. 71
3. 84
4. 68
5. 72

192 / 16 =

1. 24
2. 18
3. 27
4. 25
5. 12

143 / 11 =

1. 23
2. 13
3. 31
4. 30
5. 14

276 / 12 =

1. 38
2. 42
3. 41
4. 23
5. 20

5.5.4 Lightbot Quiz

What does this conditional evaluate to? (not ((true) and (false))) and (not ((false) or (not false)))

1. true
2. false

What does this conditional evaluate to? ((not false) and (not true)) and ((false) and (not true))

1. true
2. false

What does this conditional evaluate to? not (((not true) and (false)) and ((true) and (false)))

1. true
2. false

What does this conditional evaluate to? ((true) and (true)) or ((false) and (not true))

1. true
2. false

Examine the following code:
int count = 0; while (count <= 6) { System.out.print(count + " "); count = count + 2; } System.out.println( );
What does this code print on the monitor?

1. 1 2 3 4 5 6
2. 0 2 4 6 8
3. 0 2 4
4. 0 2 4 6

Examine the following code:
int count = 7; while (count >= 4) { System.out.print(count + " "); count = count - 1; } System.out.println( );
What does this code print on the monitor?

1. 1 2 3 4 5 6 7
2. 7 6 5 4
3. 6 5 4 3
4. 7 6 5 4 3

Examine the following code:
int count = -2; while (count < 3) { System.out.print(count + " "); count = count + 1; } System.out.println( );
What does this code print on the monitor?

1. -2 -1 1 2 3 4
2. -2 -1 1 2 3
3. -3 -4 -5 -6 -7
4. -2 -1 0 1 2

Examine the following code:

int count = 1; while (count < 5) { System.out.print(count + " "); } System.out.println( );

What does this code print on the monitor?

1. 1 2 3 4
2. 1 2 3 4 5
3. 2 3 4
4. 1 1 1 1 1 1 1 1 1 1 . . . .

Examine the following code:

function foo() { bar(); print('foo'); } function bar() { print('bar'); }

What does the code "foo(); bar()" print on the monitor?

1. foobar
2. barfoobar
3. foobarfoo
4. barbarfoo

Examine the following code:

function foo() { bar(); print('foo'); } function bar() { print('bar'); }

What does the code "bar(); foo()" print on the monitor?

1. foobar
2. barfoobar
3. foobarfoo
4. barbarfoo

Chapter 6

Results

6.1 Opinions on educational and fun games

These are the aggregated responses to the opinion Likert scale questions introduced in the very first part of the survey. While they don't have any influence on our results or analysis, it's interesting to note what kind of opinions or mindset workers when they begin taking the survey.

These questions include opinions about whether educational games can be fun (or vice versa), and if most of them are. There were also some questions about fun/educational games being competitive/cooperative, and more fun/educational if playing online/offline, solo or with friends.

The most interesting sections of this chart are the "Fun/educational games can be educational/fun" sections; almost unanimous agreement there. However, "most fun games are educational" met with a substantial amount of disagreement. This is expected; games that are commonly considered fun include shooters, adventure and action games. In addition, the "most educational games are fun" section was far more evenly distributed, implying that a substantial amount of people think educational games aren't fun.

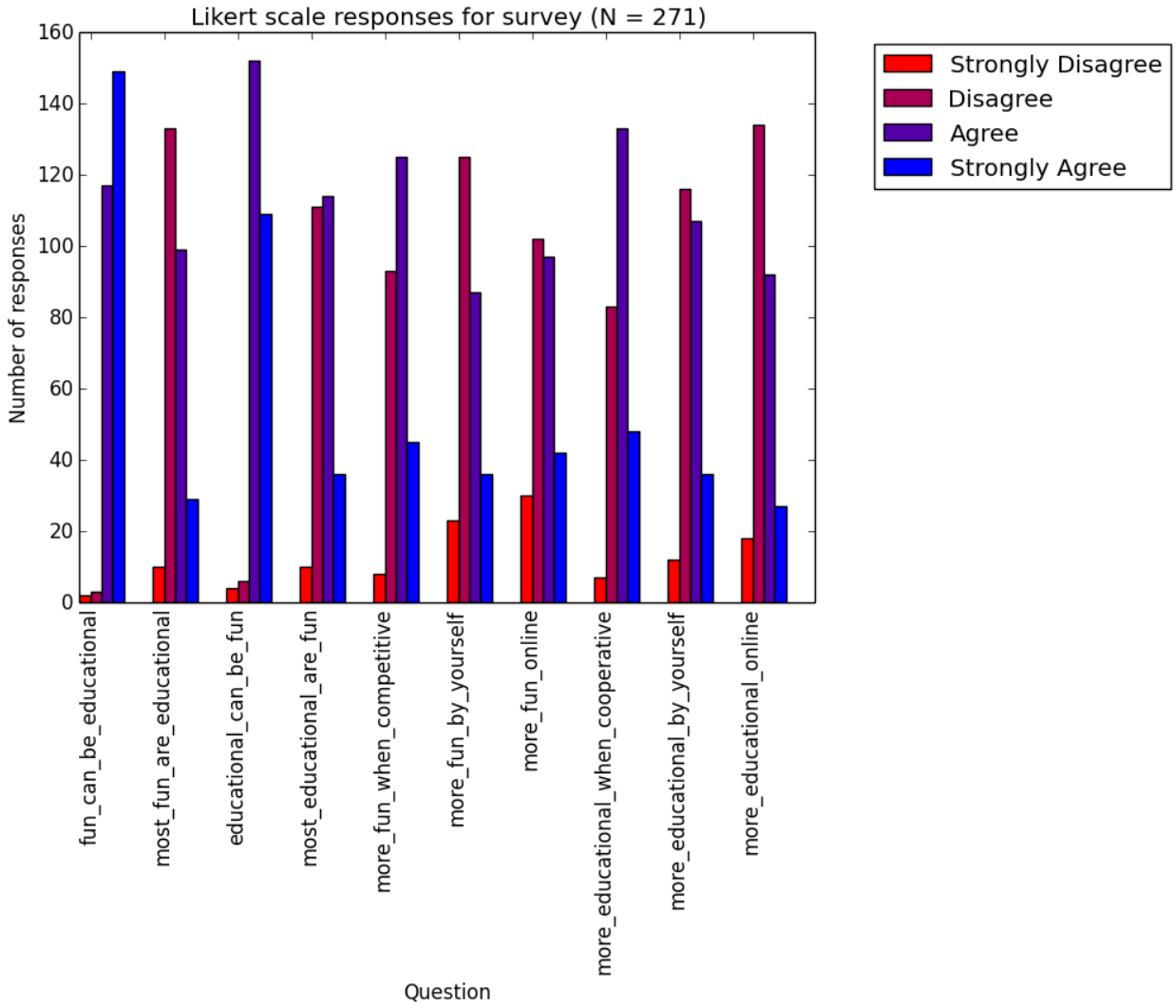

Figure 6.1: Likert scale responses for several questions on fun and educational games.

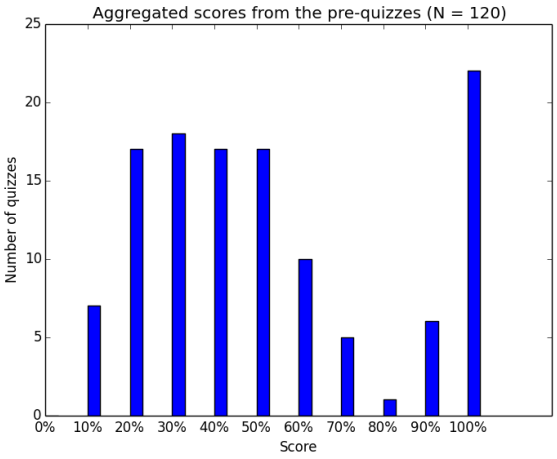

Figure 6.2: The aggregate pre-quiz scores across all games.

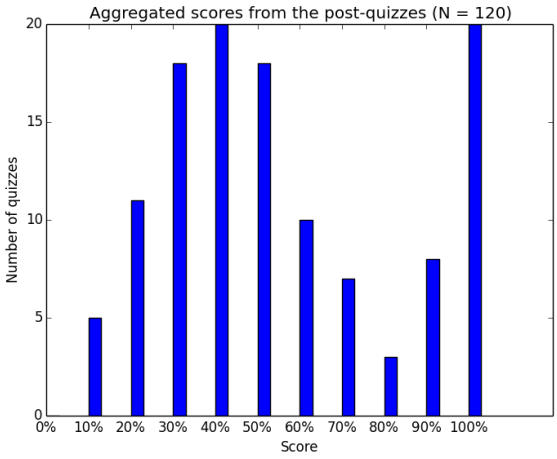

Figure 6.3: The aggregate post-quiz scores across all games.

6.2 Quiz Results

This section contains results of all of the quiz responses, displayed in 3 graphs. The pre-quiz and post-quiz graphs indicate how many responses were received with a specific grade, from 0% to 100%. The score differences graphs indicate how many responses improved from the pre-quiz to the post-quiz, and by how much.

For each graph, I'll include some non-statistical observations and insights. For detailed statistical analysis, see the Analysis chapter.

6.2.1 Aggregate Quiz Scores

The complete lack of a bell curve on our pre-quiz (Figure 6.2) and post-quiz (Figure 6.3) aggregates means that our quizzes may not be as evenly distributed as we'd like. As explained in the next chapter, aggregating the responses across all of our quizzes does not yield score increases that are statistically significantly different from zero.

The massive peak in the score differences chart (Figure 6.4) shows how little of a change there was in the majority of the responses. This means that most quizzes had zero improvement or change in score. This was unexpected, because the pre-quiz and post-quiz for each worker was almost identical; with very little variation in the answers.

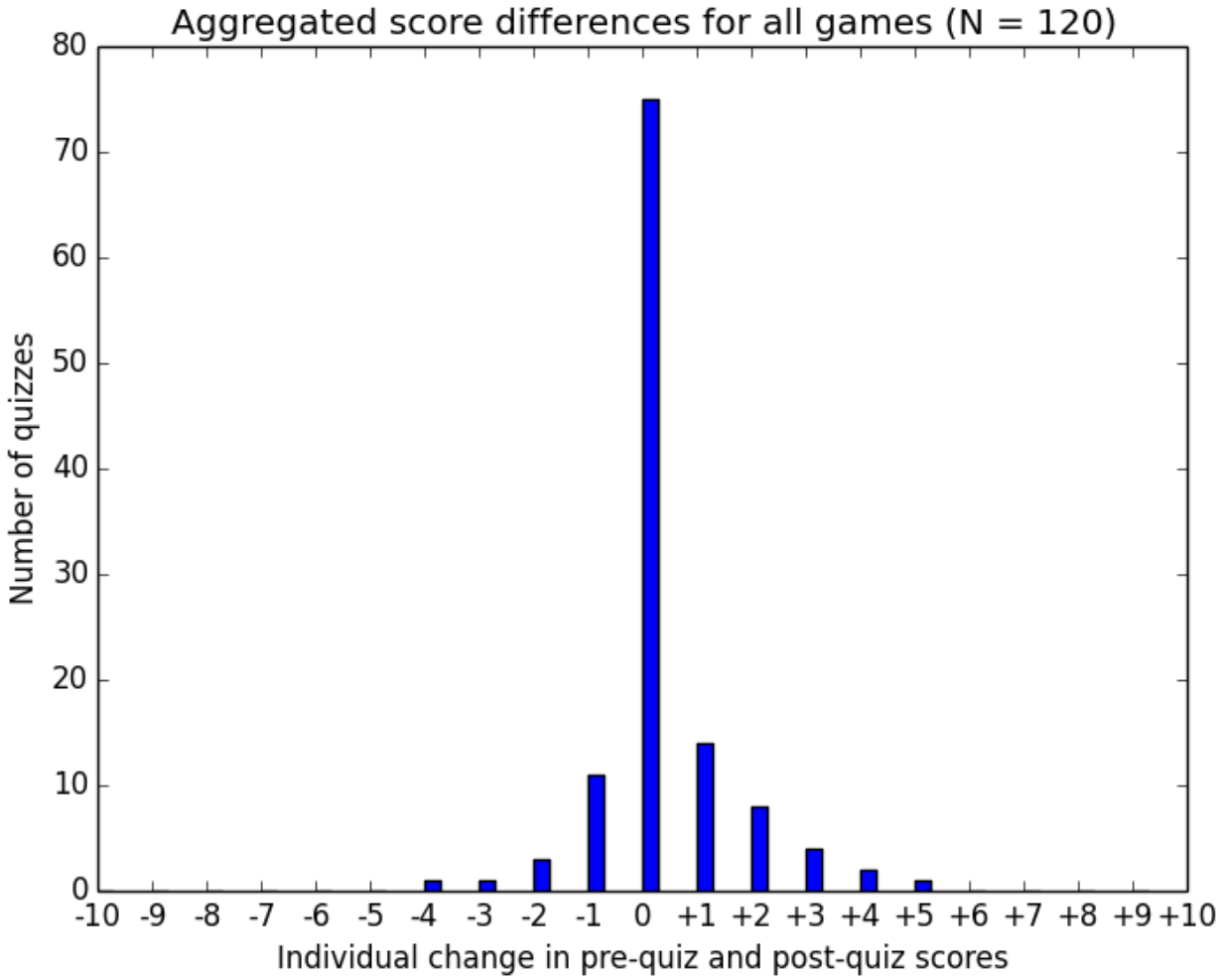

Figure 6.4: The aggregate score differences across all games.

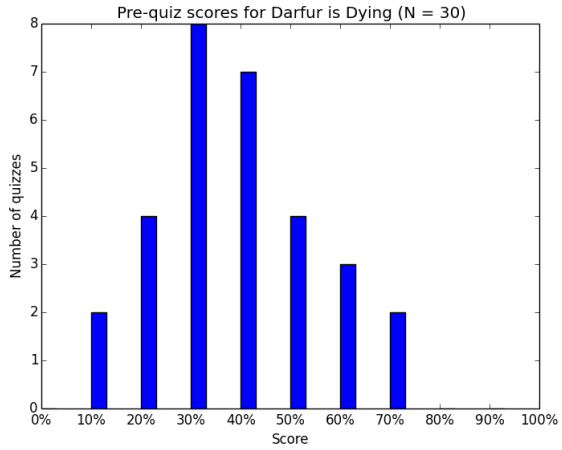

Figure 6.5: The pre-quiz scores for Darfur is Dying.

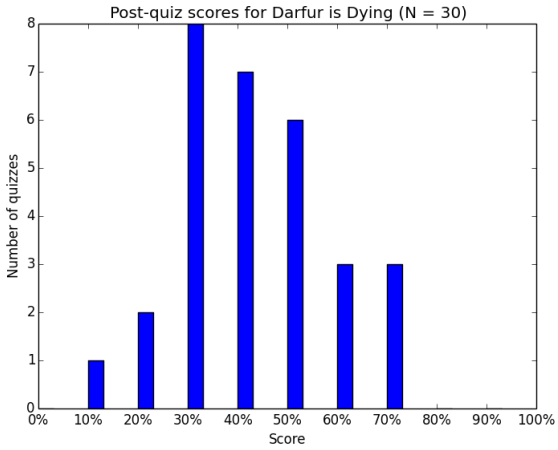

Figure 6.6: The post-quiz scores for Darfur is Dying.

6.2.2 Darfur Quiz Scores

The pre-quiz (Figure 6.5) and post-quiz (Figure 6.6) graphs show a roughly standard distribution. The score differences graph (Figure 6.7) shows the massive peak around zero, with minor improvements.

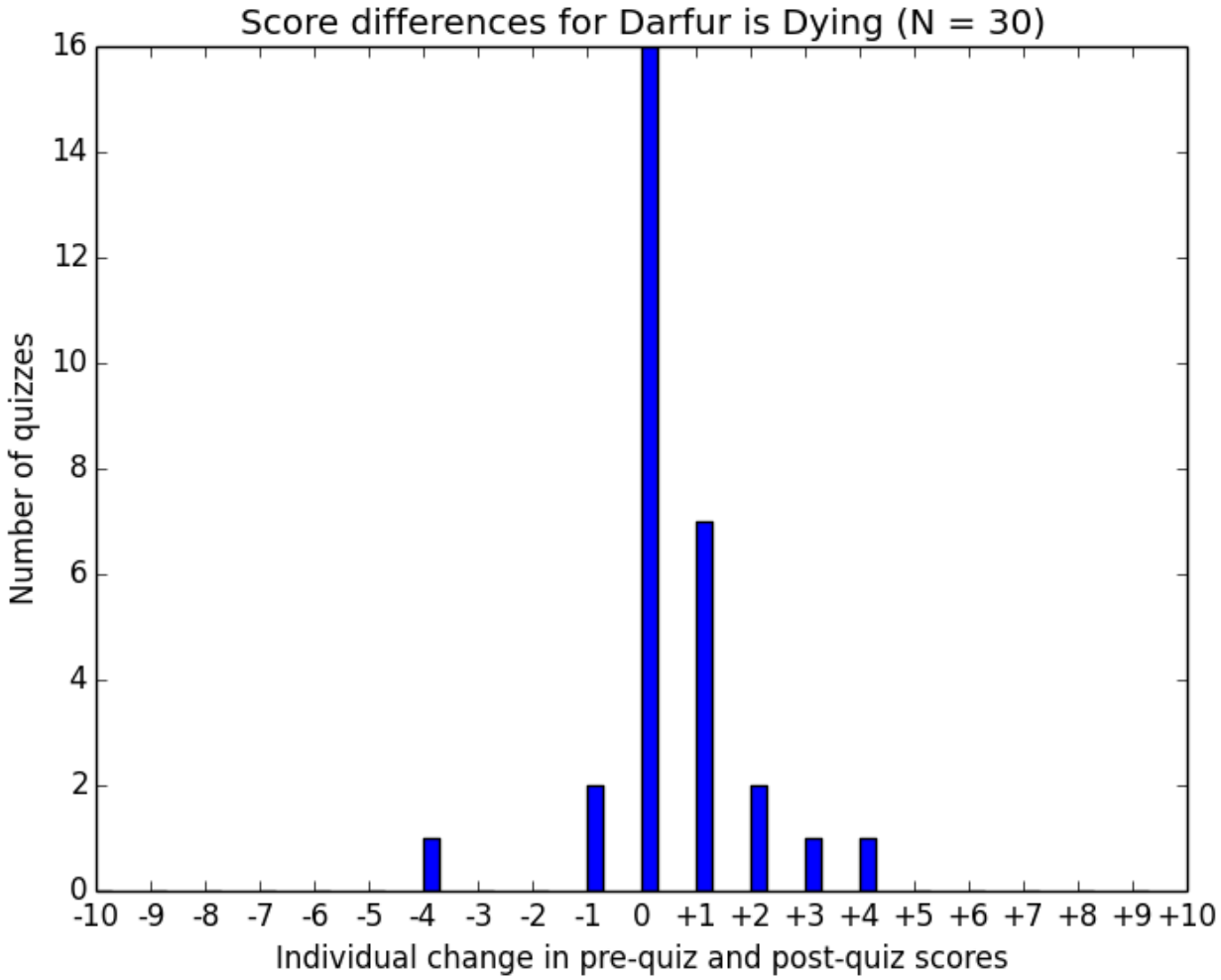

Figure 6.7: The score differences for Darfur is Dying.

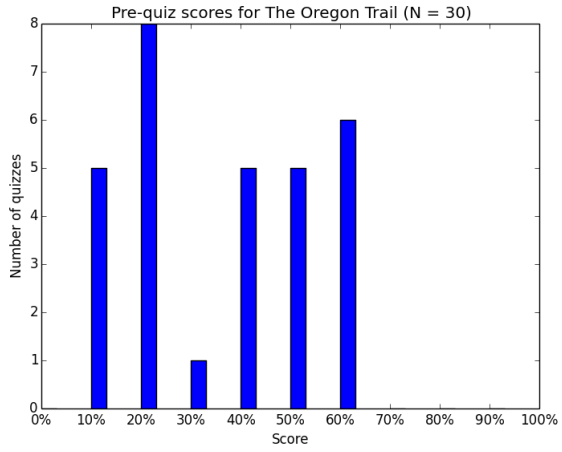

Figure 6.8: The pre-quiz scores for The Oregon Trail.

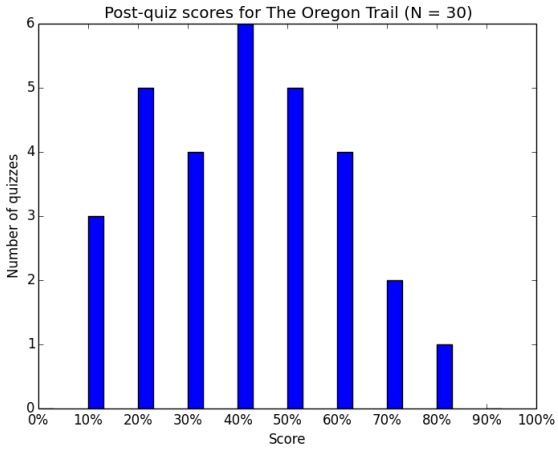

Figure 6.9: The post-quiz scores for The Oregon Trail.

6.2.3 The Oregon Trail Quiz Scores

The pre-quiz (Figure 6.8) and post-quiz (Figure 6.9) graphs show an increase in some scores, although there was almost no change for the majority of the quizzes (Figure 6.10)

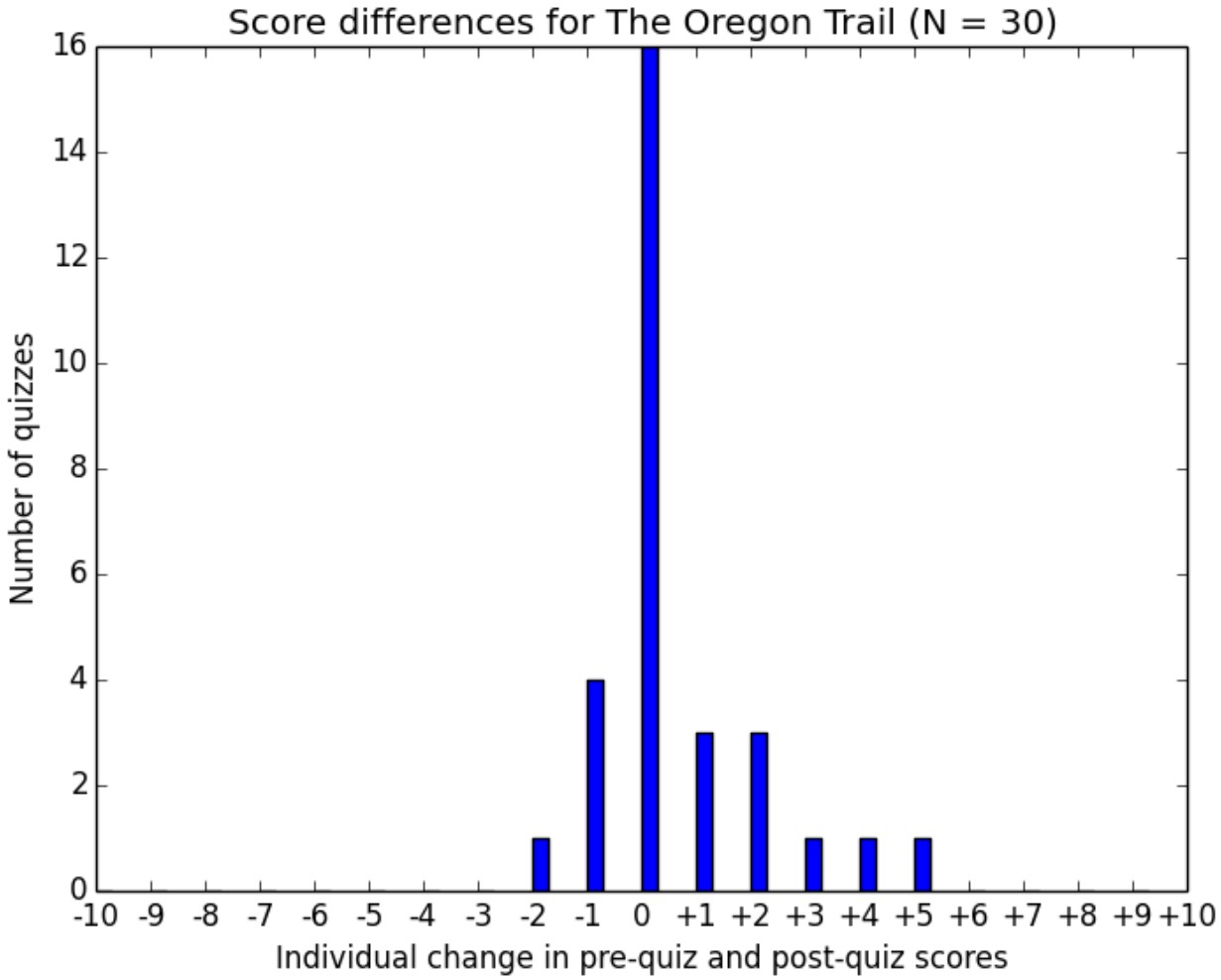

Figure 6.10: The score differences for The Oregon Trail.

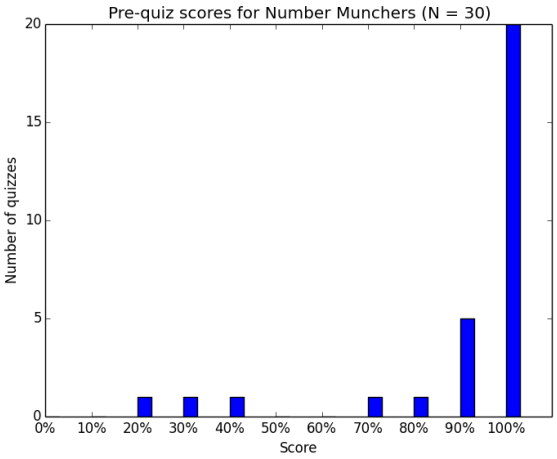

Figure 6.11: The pre-quiz scores for Number Munchers.

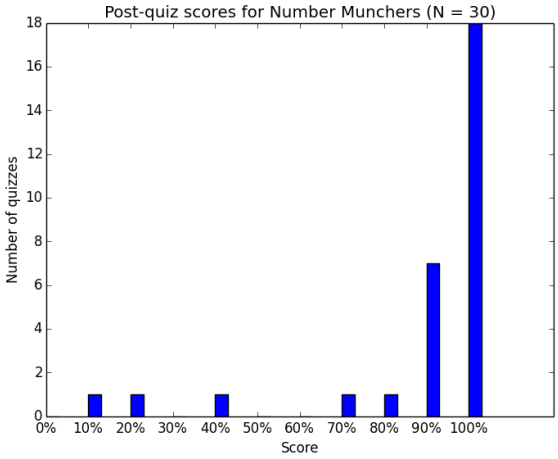

Figure 6.12: The post-quiz scores for Number Munchers.

6.2.4 Number Munchers Quiz Scores

The pre-quiz (Figure 6.11) and post-quiz (Figure 6.12) obviously ended up skewed towards 100%. This is likely caused by the simplicity of the quiz; simple arithmetic is easy for Mechanical Turk workers aged 18+ to do, and some are likely assisted by a calculator (even though we expressly asked them not to use one).

The score differences graph (Figure 6.13) actually shows zero positive improvements to quiz scores, with only 3 or 4 being negative.

This is a clear example of the ceiling effect. Because the workers scored so highly on the pre-quiz and post-quiz, we can assume that the quiz was not difficult enough. If the quiz were more difficult, we would see more workers getting lower scores. This would result in a better distribution of scores, and more useful information in score differences.

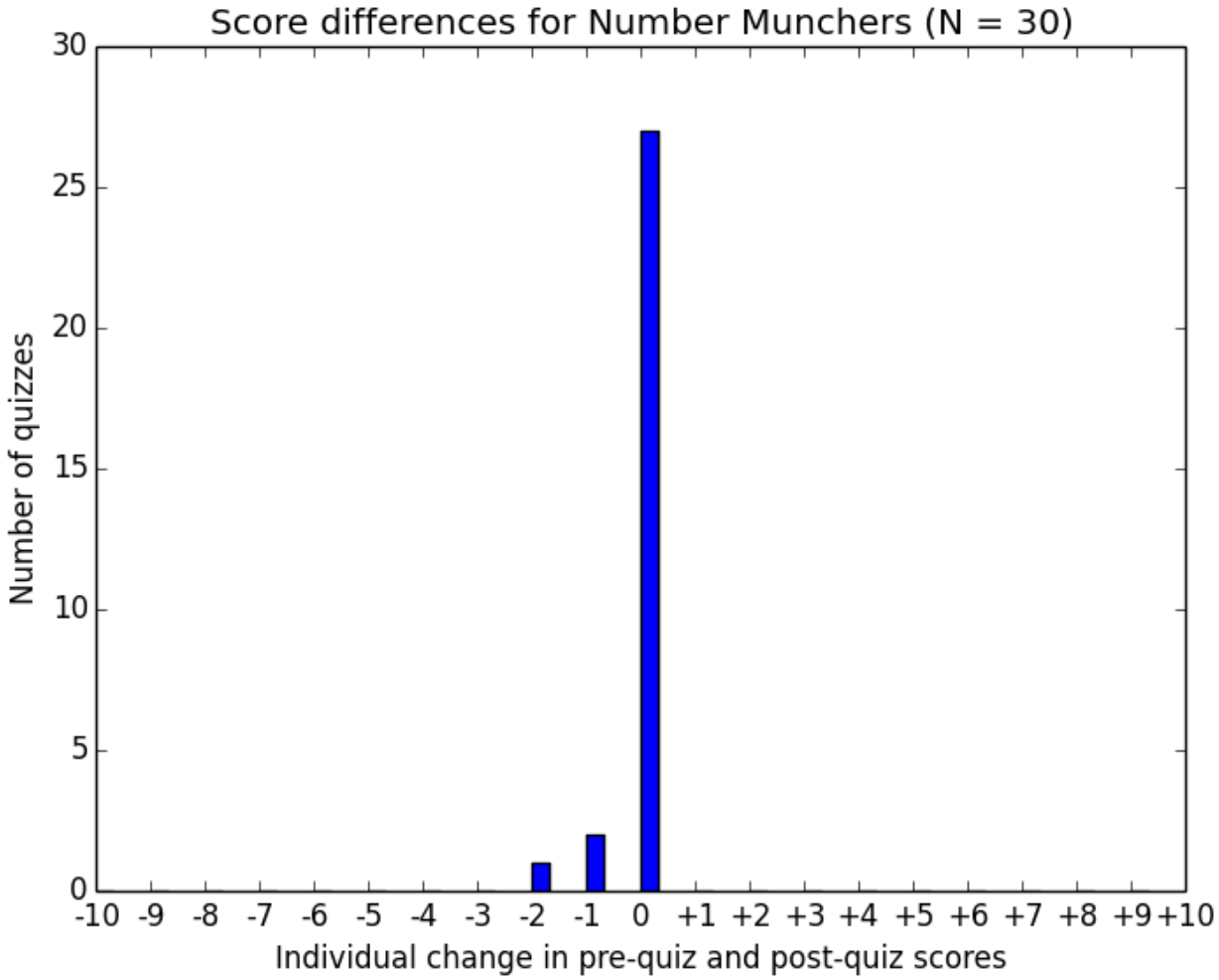

Figure 6.13: The score differences for Number Munchers.

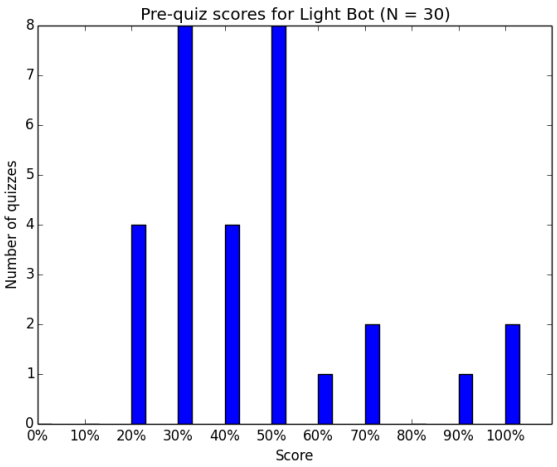

Figure 6.14: The pre-quiz scores for Light Bot.

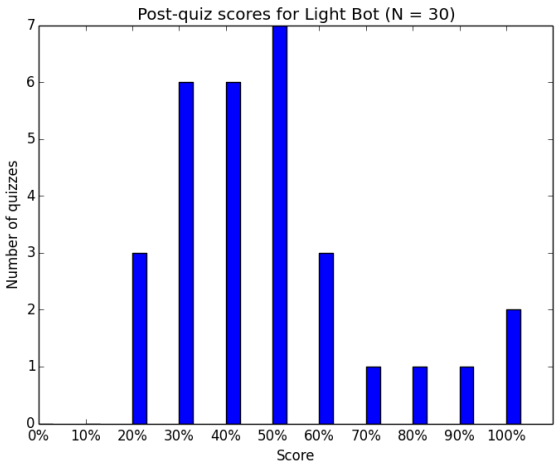

Figure 6.15: The post-quiz scores for Light Bot.

6.2.5 Light Bot Quiz Scores

Light Bot's pre-quiz (Figure 6.14) and post-quiz (Figure 6.15) scores show a roughly standard distribution skewed toward the high end. This could mean that some reviewers were already familiar with this material, or that they were so engaged with the questions that they spent extra time on them. In either case, the ceiling effect will slightly affect our score differences results (Figure 6.16). The score differences show a roughly standard disribution with a large peak at 0.

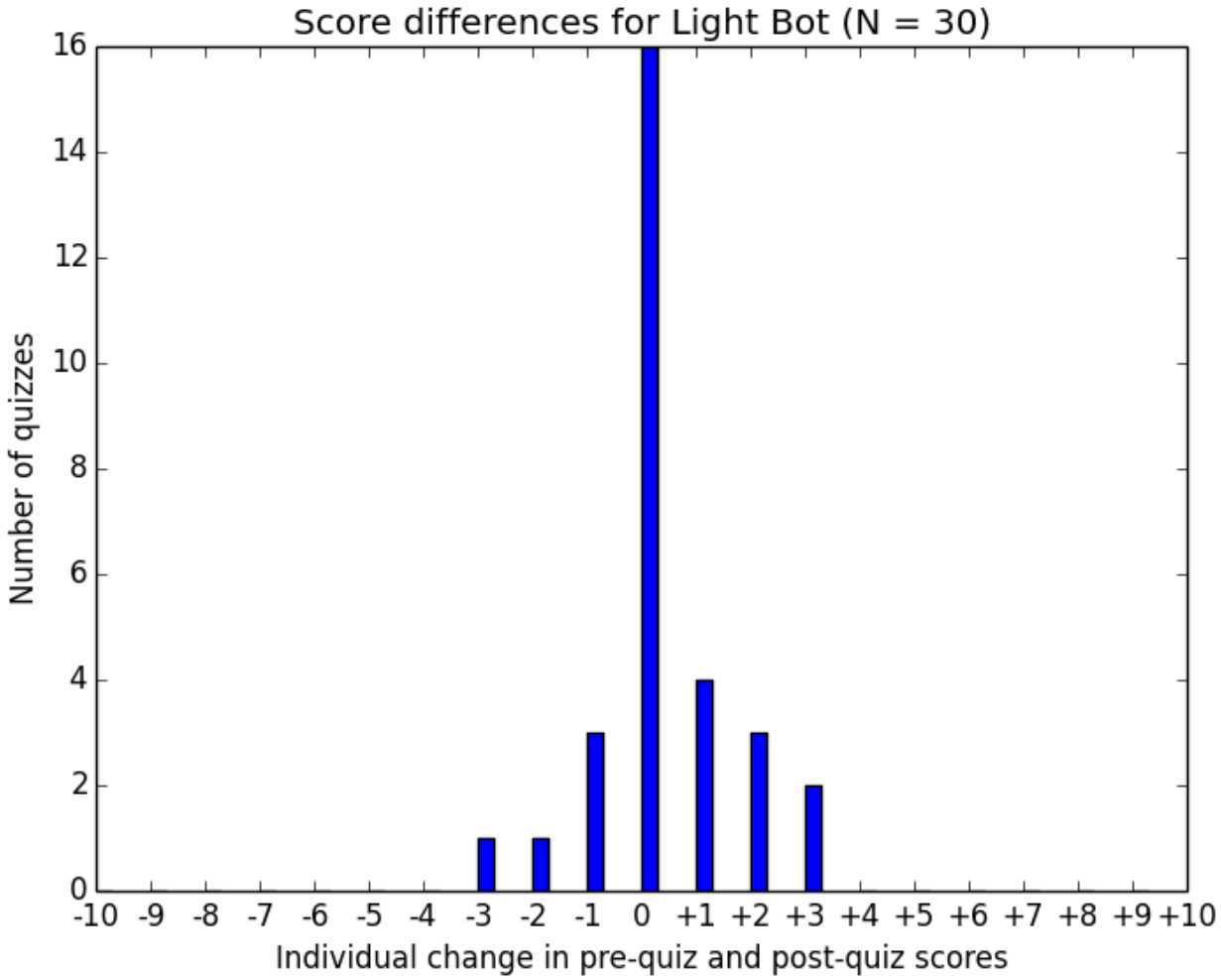

Figure 6.16: The score differences for Light Bot.

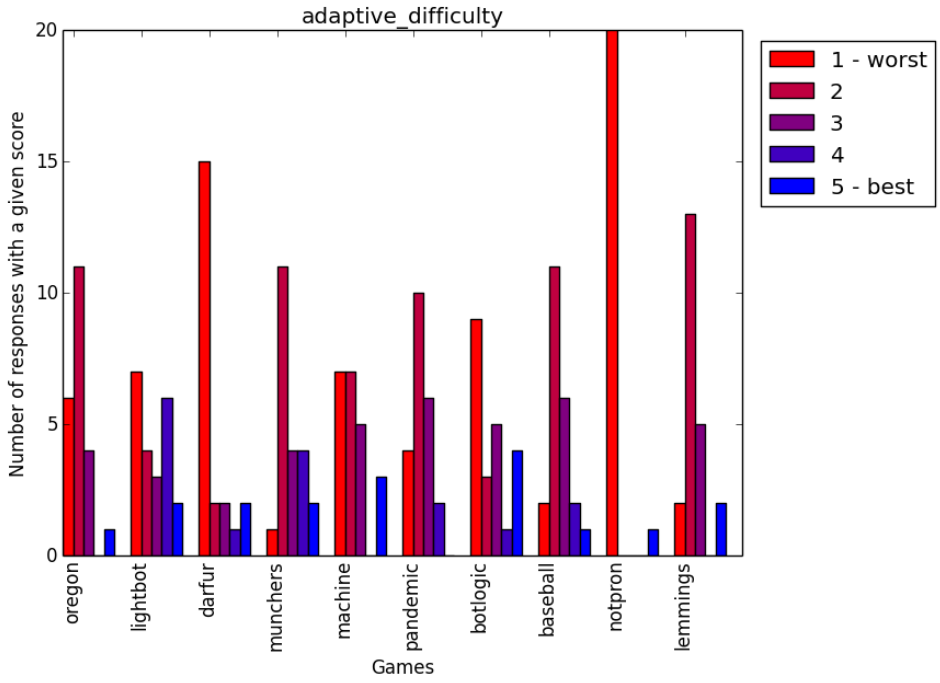

Figure 6.17: Adaptive Difficulty

6.3 Rubric Scores

This section contains the scoring results for each rubric item. For each rubric item, a bar graph is given for each game. These graphs are shown in Figures 6.17 through 6.29. The bar graph contains the scores that the game recieved for that rubric item.

An alternate visualization of this data, where each game has a bar graph for all of the rubric items, can be found in the next section.

6.3.1 Adaptive Difficulty

The most striking part of this graph is the obvious lack of adaptive difficulty present in *Notpron*, and to some degree, *Darfur is Dying*. This seems straightforward, as both *Notpron* and *Darfur is Dying* have no option to set difficulty. There are no games with large positive degrees of adaptive difficulty.

6.3.2 Checkpoint Frequency

For checkpoint frequency, response was universally negative. This is concerning, because I was confident that some games (*Notpron, The Incredible Machine, Light Bot*) had frequent checkpoints. This may imply that the prompt or selections were misleading.

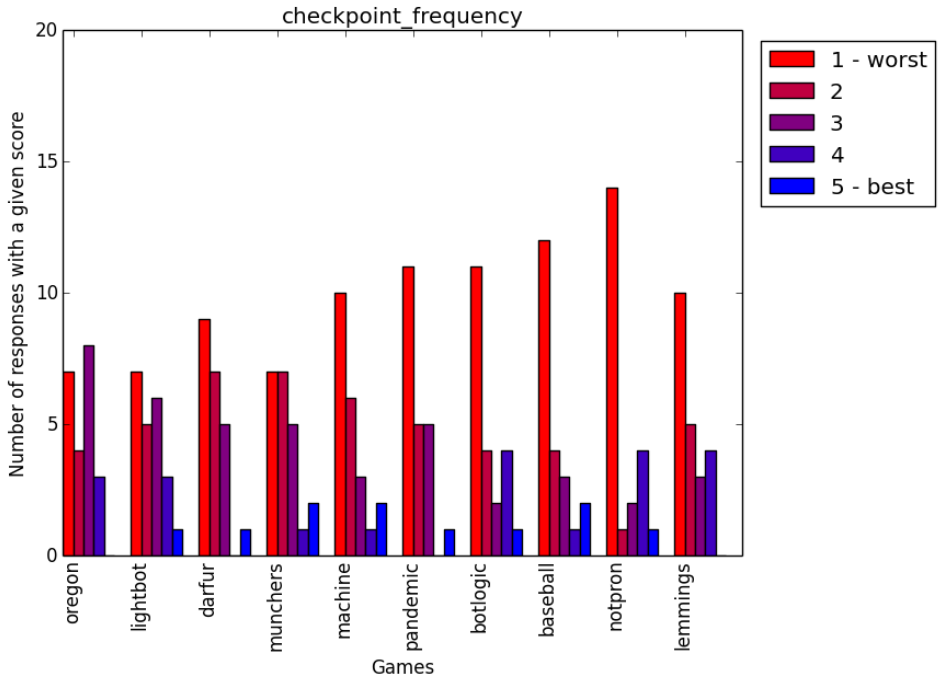

Figure 6.18: Checkpoint Frequency

6.3.3 Contextual Tutorials

Response for contextual tutorials within the game was universally negative, with the most notable instance being *Math Baseball*. This is expected; Math Baseball is a simple arithmetic game that provides no tutorials at all.

6.3.4 Encyclopedia Content

For the content of the game's encyclopedia, both *The Oregon Trail* and *Darfur is Dying* scored relatively highly. It's expected that both of these games (focused around historical/current events) seem to directly benefit from tangential learning.

6.3.5 Encyclopedia Location

The only game that scored moderately well on the location of the game encyclopedia was *The Oregon Trail*. This is expected; the introduction sequence of the game provides a lot of useful and historical information.

6.3.6 Freedom of exploration

The only game that scored relatively well on freedom of exploration was *Pandemic 2*. Players likely considered the 'open world' format of the game as part of this rubric item.

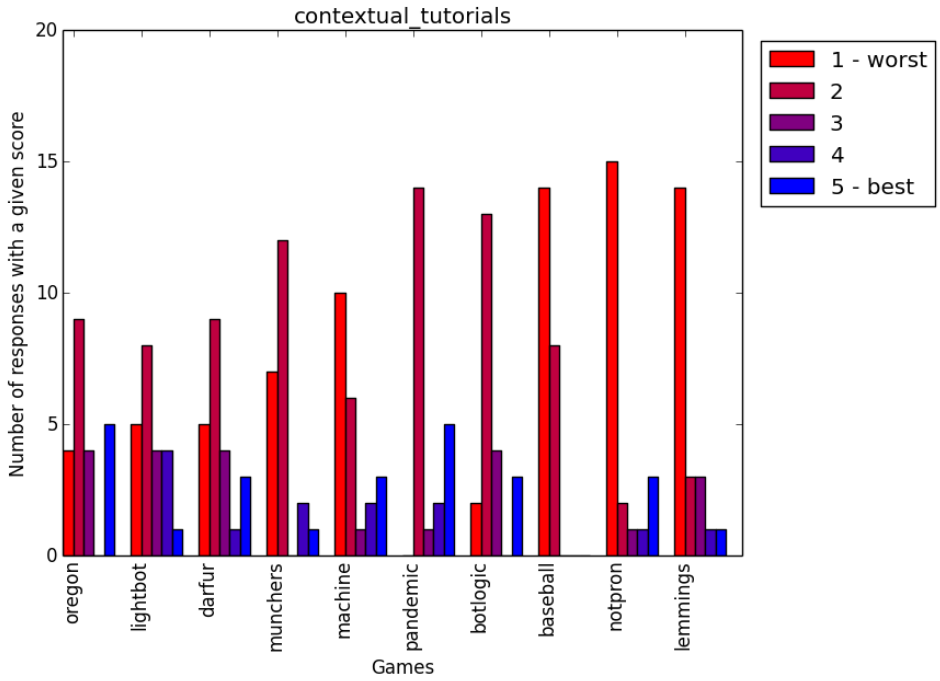

Figure 6.19: Contextual Tutorials

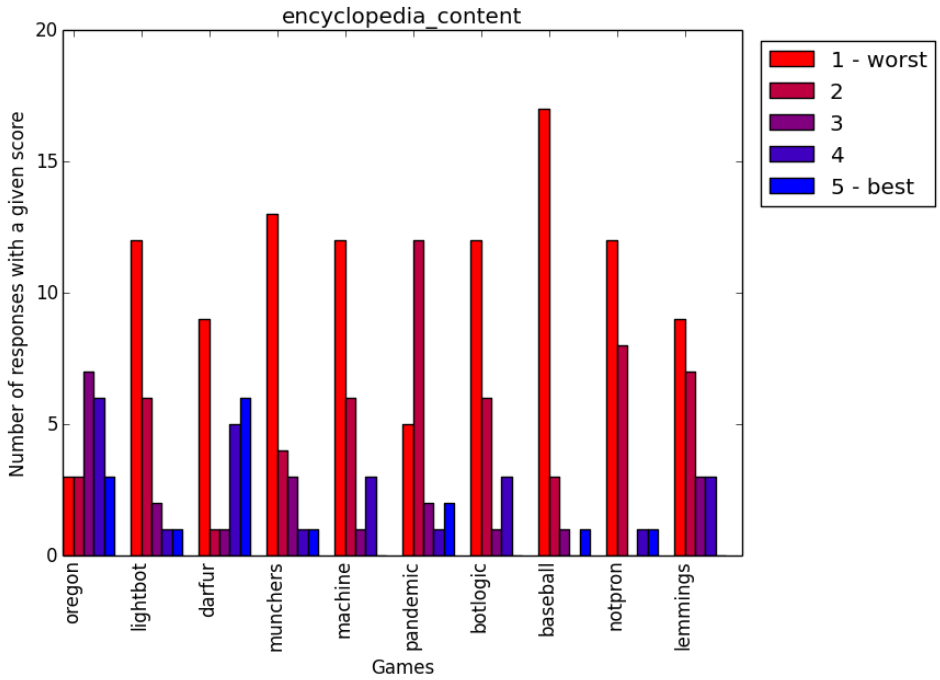

Figure 6.20: Encyclopedia Content

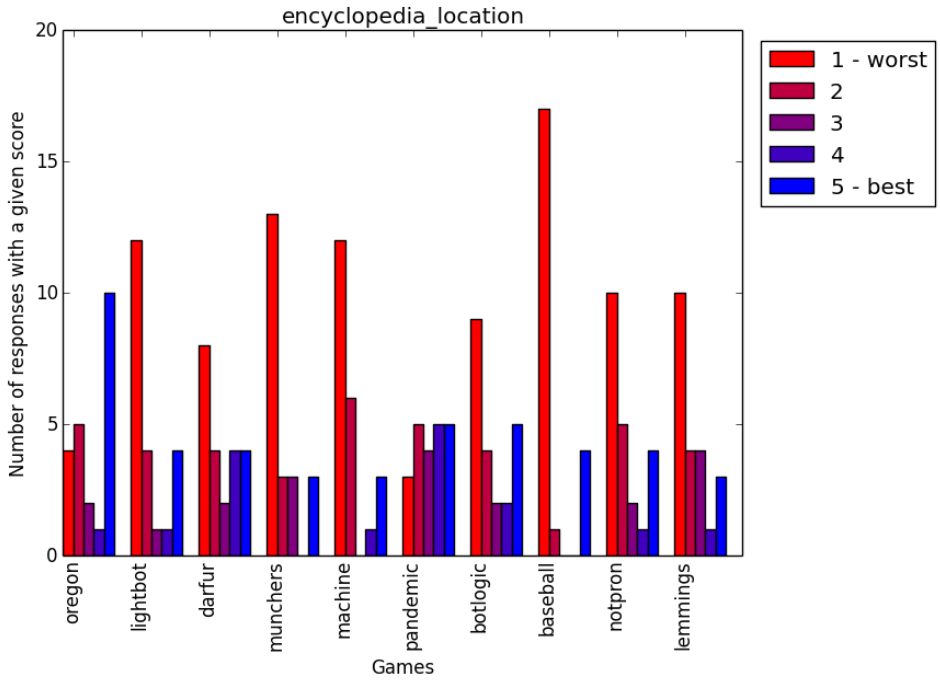

Figure 6.21: Encyclopedia Location

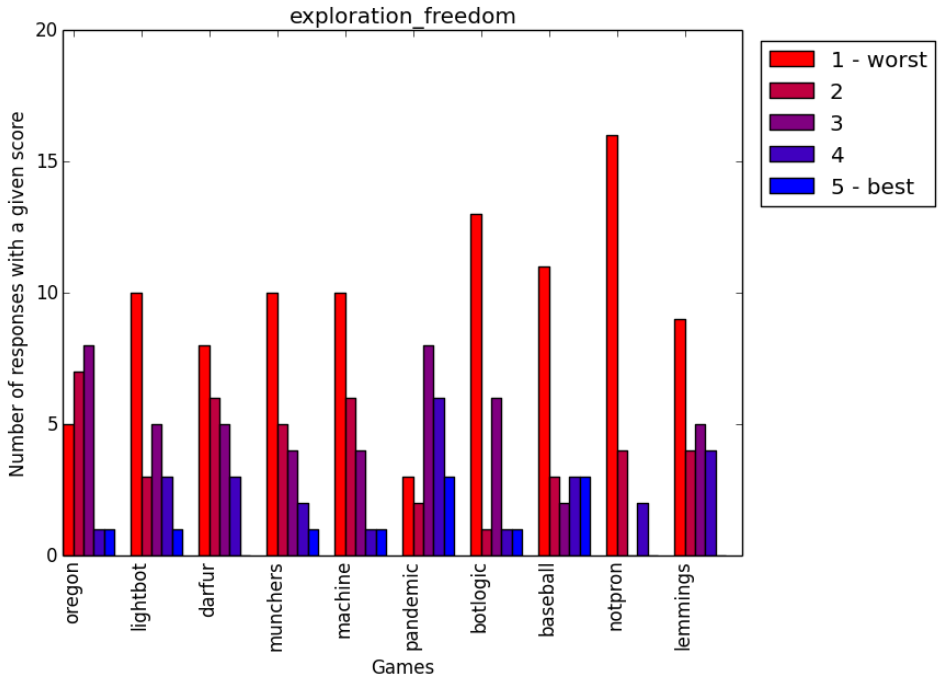

Figure 6.22: Freedom of exploration

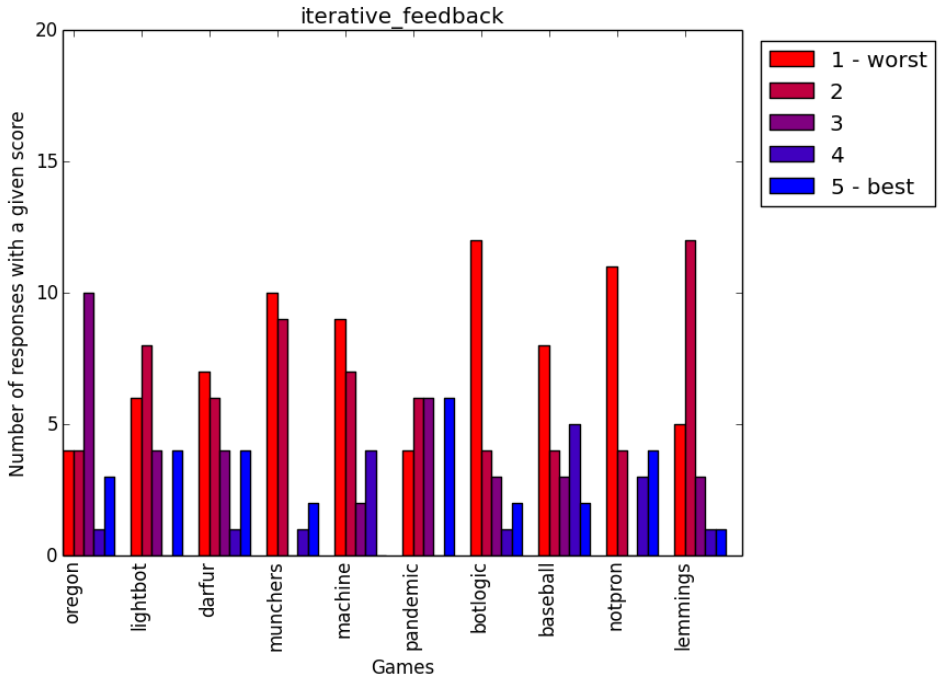

Figure 6.23: Iterative feedback

6.3.7 Iterative Feedback

Scores for games on iterative feedback were mixed at best. For *The Oregon Trail*, players may have considered their frequent check-ins while along the trail useful, where they could ask for advice and make decisions.

6.3.8 Unorthodox problem-solving

Scores for unorthodox problem-solving were mixed; no significant exceptions to note.

6.3.9 Amount of referential material

Several games scored well on the amount of referential material in them. *The Oregon Trail* and *Darfur is Dying* were both expected, as well as *Pandemic* to a degree. *Notpron* was expected to do well, given the knowledge (web technology) that players are expected to know to be able to complete the game. However, *Notpron* only scored in the middle of the range.

6.3.10 Popularity of referential material

Math Baseball's popular referential material is considered basic arithmetic, something that is ubiquitous outside the game. It scored very highly. However, *Number Munchers*, which covers extremely similar material, scored very poorly on

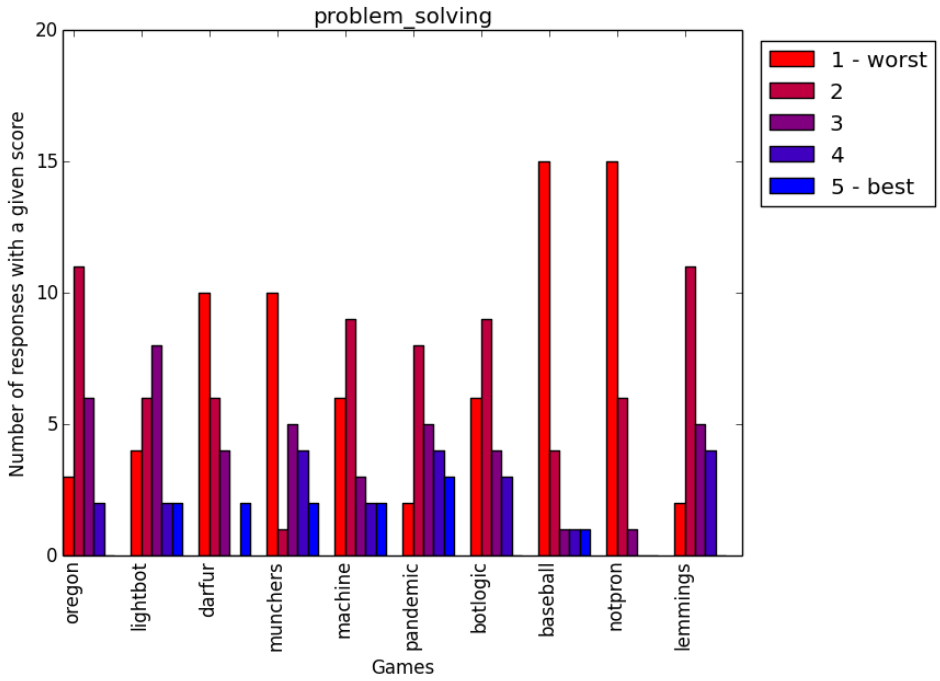

Figure 6.24: Unorthodox problem-solving

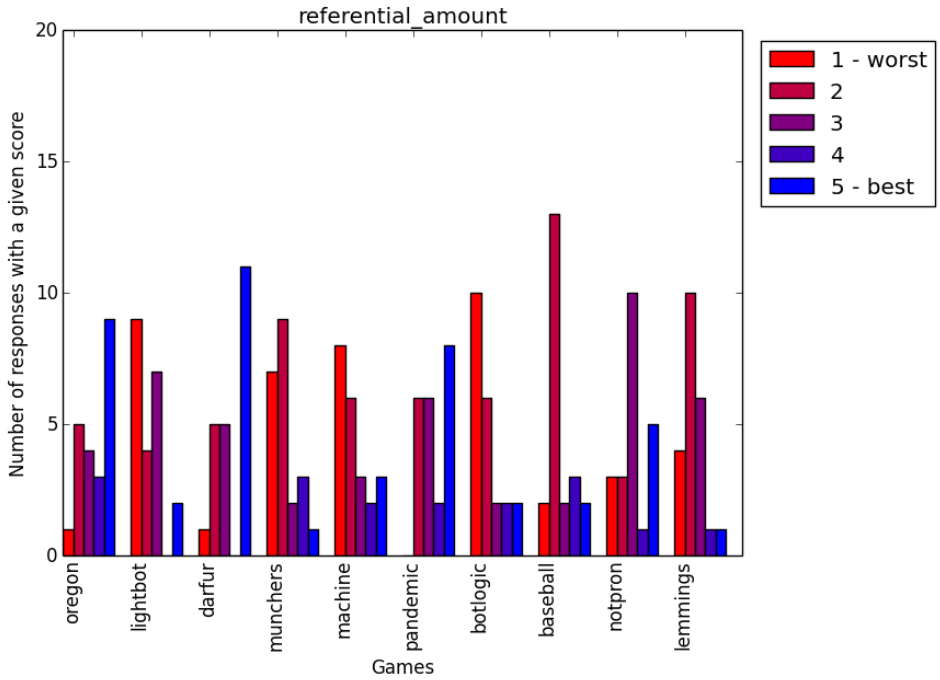

Figure 6.25: Amount of referential material

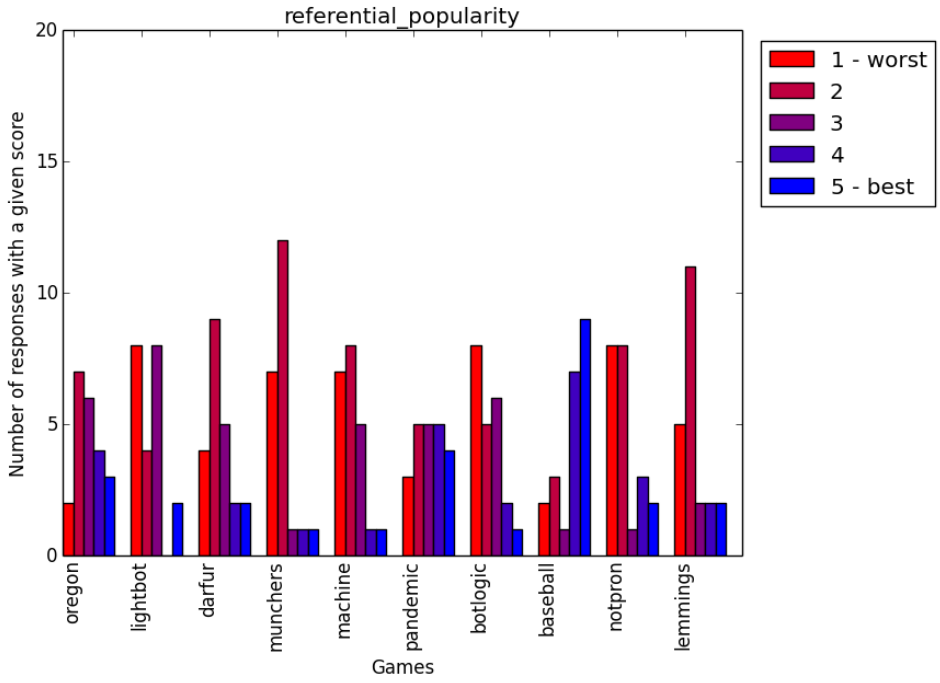

Figure 6.26: Popularity of referential material

this rubric item. This leads us to think that the rubric may be inconsistent, or additional research may need to be done.

6.3.11 Rewards for knowing referential material

Darfur is Dying was the only game with a significant positive score for referential rewards. This is expected, as the game contains situations where awareness of the conflict in Darfur directly affects gameplay (e.g. children should fetch the water instead of adults).

6.3.12 Reset time penalty for failure

In contrast to most of the other rubric items, the reset time penalty for failure recieved universally postitive scores for all games. This means that all the games were very effective about using the player's failure time wisely, or that the prompt or choices biased players towards positive responses.

6.3.13 Game resource penalty for failure

The resource penalty for failure received some strong positive and negative ratings for games. *The Oregon Trail, Number Munchers, and Pandemic* received negative ratings, which is somewhat expected; in the games, if the player fails, they usually have to restart the game. However, for *Light Bot, Botlogic, and Notpron*, players can restart a failed level or puzzle immediately, usually with no penalty.

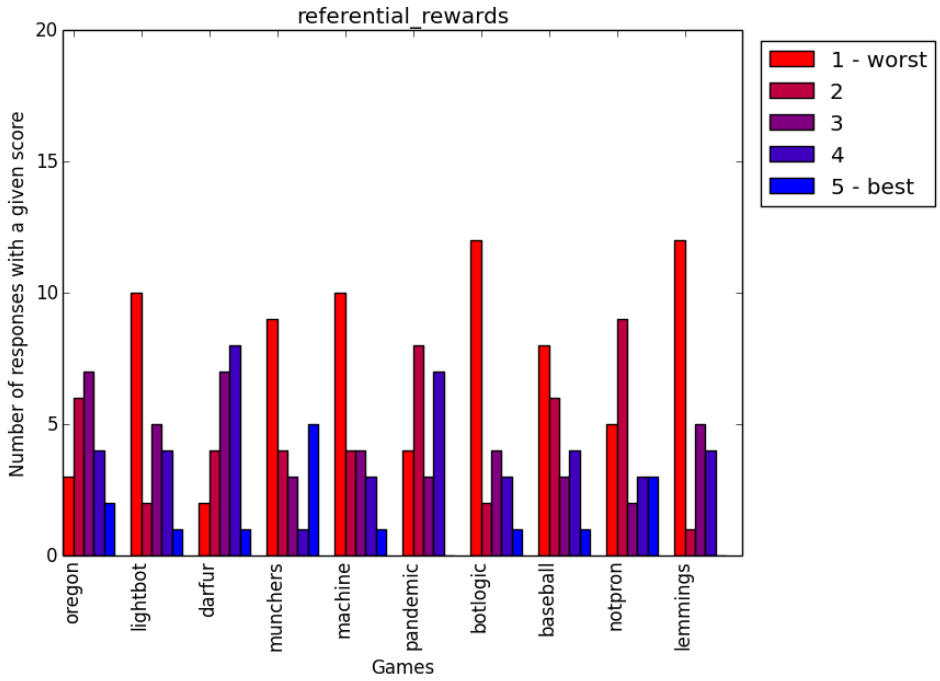

Figure 6.27: Rewards for knowing referential material

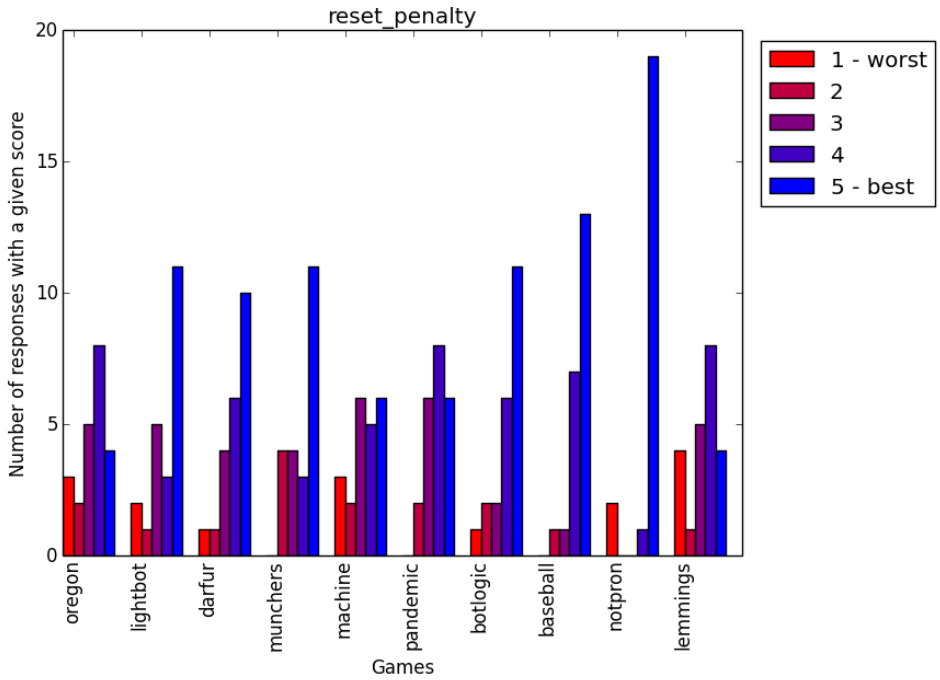

Figure 6.28: Reset time penalty for failure

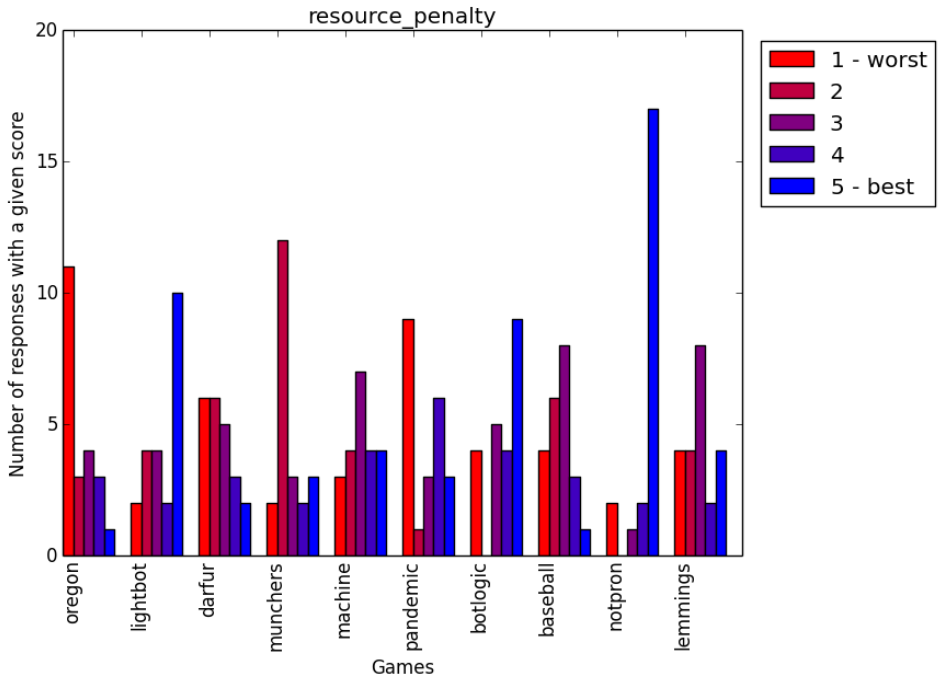

Figure 6.29: Game resource penalty for failure

It's important to note that raters may have confused game resource penalty with reset time penalty. It warrants further investigation, either into the survey prompts and choices, or into the games themselves.

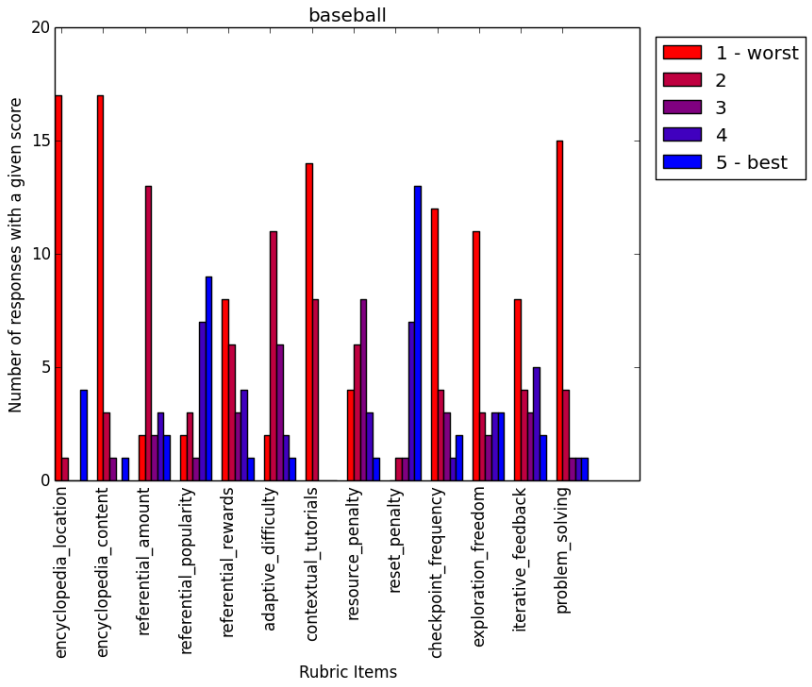

Figure 6.30: Math Baseball

6.4 Game Scores

This section contains the scoring results for each game. For each game, a bar graph is given for each rubric item. These graphs are shown in Figures 6.30 through 6.39. Each bar graph contains the scores that the rubric item recieved for that game.

An alternate visualization of this data, where each rubric item has a bar graph for all of the game, can be found in the previous section.

6.4.1 Math Baseball

Math Baseball scored well on the Popularity of Referential Material and the Reset time penalty. Popularity of Referential Material is easily explained, due to the prevalence of simple arithmetic in everyday life. The reset time penalty for *Math Baseball* is also minimal; it takes seconds to set up a new game.

6.4.2 BotLogic

BotLogic scored well on both the Game resource penalty and the Reset time penalty. In the game, if the player has a poorly written program, they only need to press the reset button to start the level over again, keeping all of their previous work. The high scores for both items are expected.

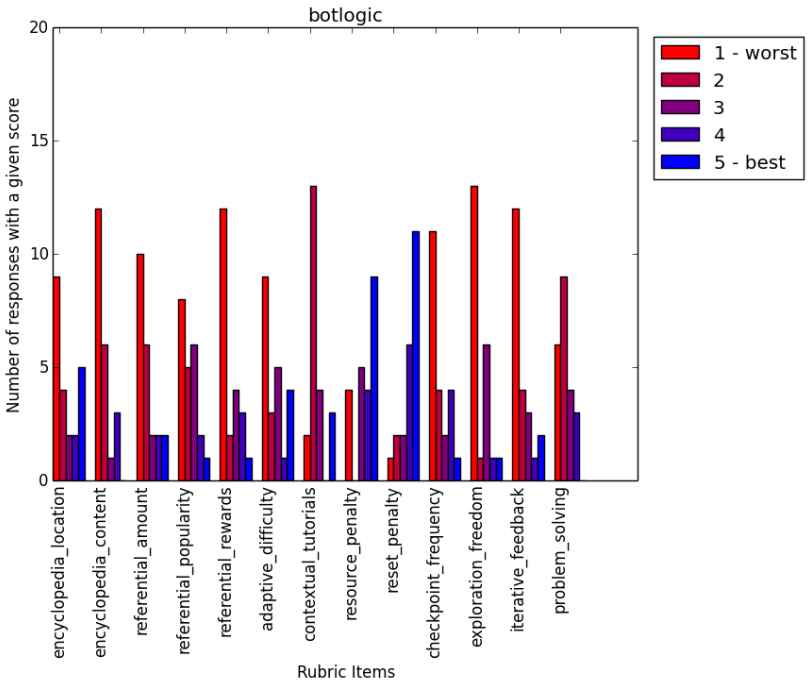

Figure 6.31: Botlogic

6.4.3 Darfur is Dying

Darfur is Dying scored well in the Amount of Referential Material, which is expected due to its purpose of advocacy. It also scores well on Reset Time Penalty; if players fail in the game, they're immediately taken back to a selection screen where they can start the challenge again.

6.4.4 Lemmings

Lemmings didn't score well in any items except marginally in Game Reset Penalty. This is expected, as level selection doesn't take very long.

6.4.5 Light Bot

Light Bot scored well on Game Resource Penalty and Reset Time Penalty. When a player has a faulty program, they can reset the level easily, keeping all of their existing work.

6.4.6 The Incredible Machine

The Incredible Machine did not score exceptionally well on any rubric items.

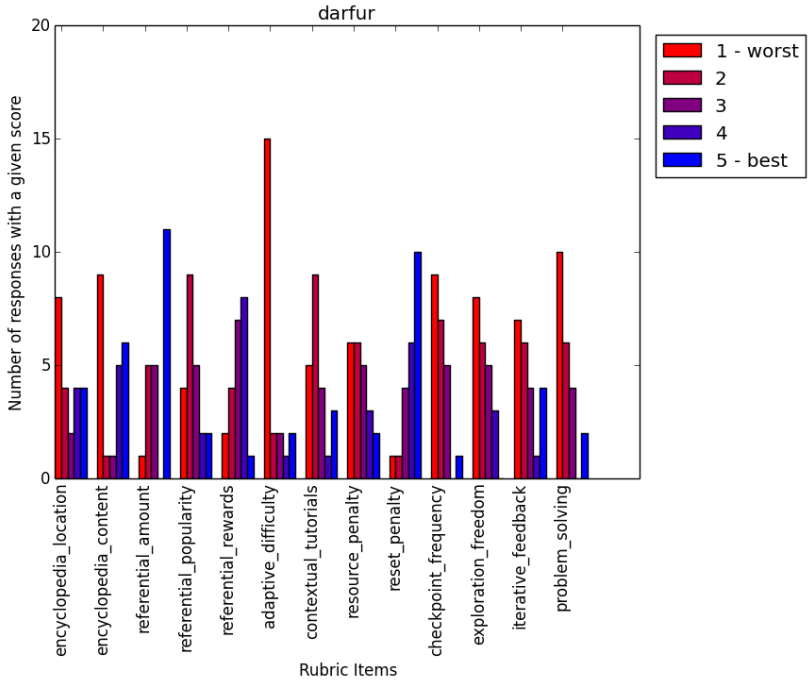

Figure 6.32: Darfur is Dying

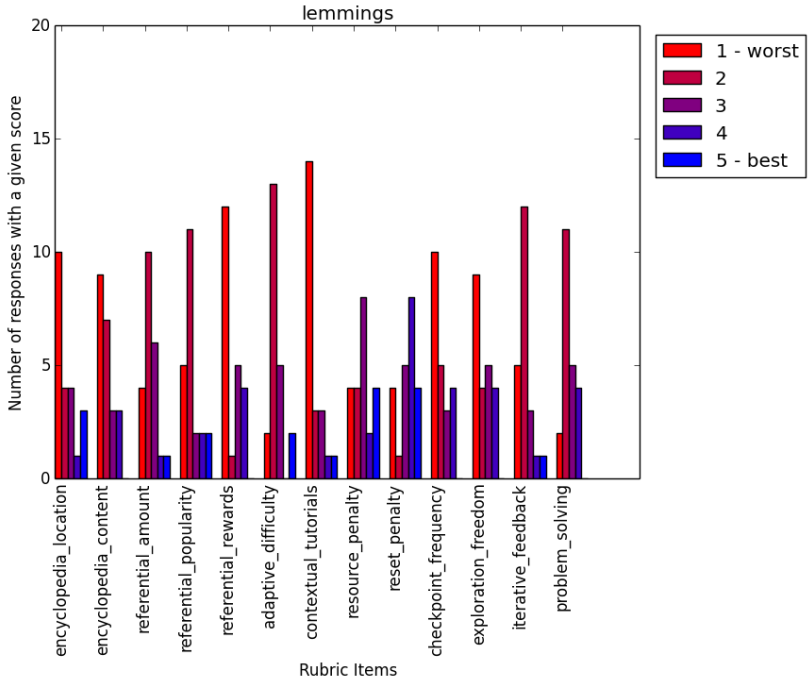

Figure 6.33: Lemmings

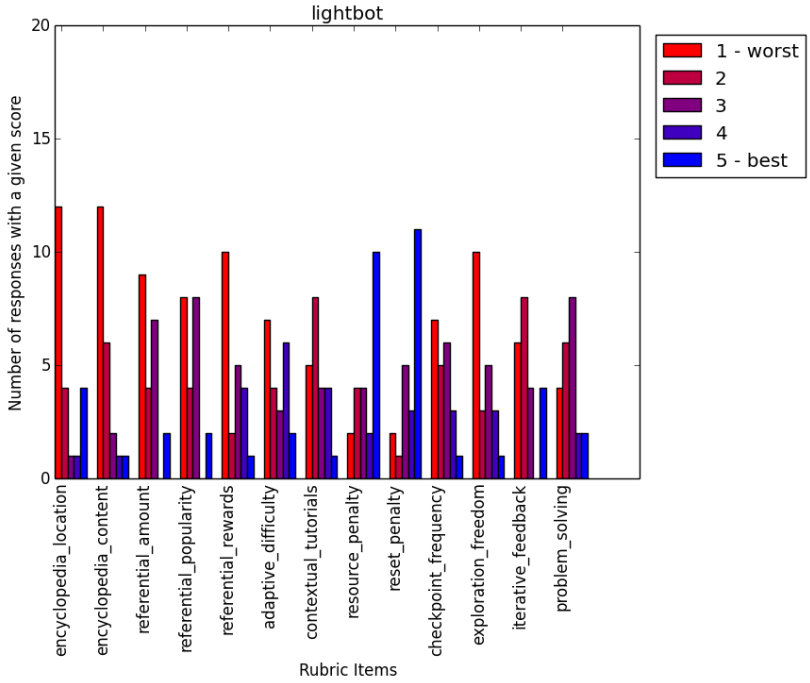

Figure 6.34: Light Bot

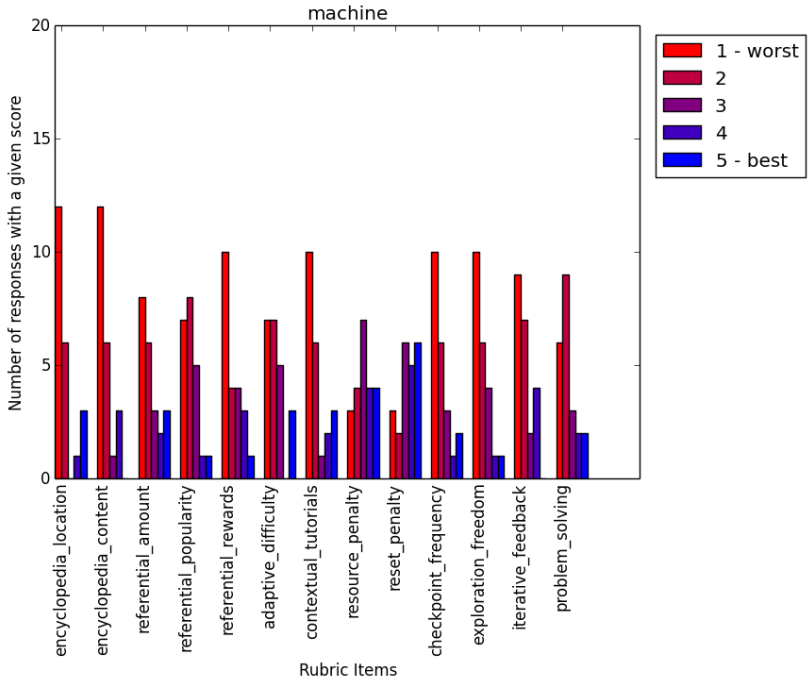

Figure 6.35: The Incredible Machine

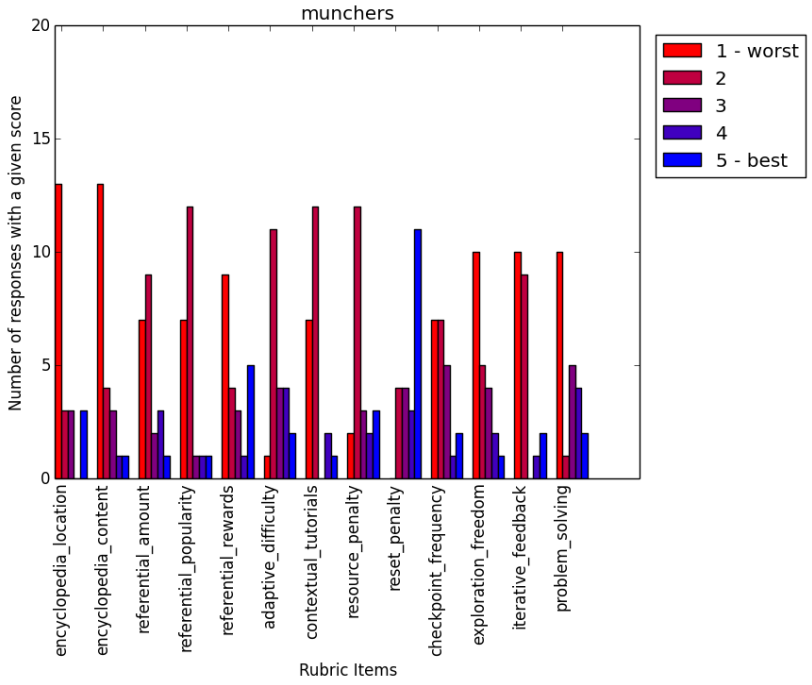

Figure 6.36: Number Munchers

6.4.7 Number Munchers

Number Munchers only scored well in the Reset Time Penalty, which is expected for a level selection screen.

6.4.8 Notpron

Notpron scored very well in both the Game Resource Penalty and the Reset Time Penalty. Failing a challenge means that a player cannot continue until they've completed the challenge, and they keep all existing work and level progress.

6.4.9 The Oregon Trail

The Oregon Trail scored highly on the Amount of Referential Information in the game, as well as the Location of the Game Encyclopedia. At the beginning of the game, players are taken through a large amount of historical and factual information about the Oregon Trail and pioneer life, so this result is expected.

6.4.10 Pandemic 2

Pandemic 2 had very mixed views for most of the rubric items. This could indicate that the nature of education within the game is something that we haven't thought of yet, or that the rubric was inaccurate at addressing these items.

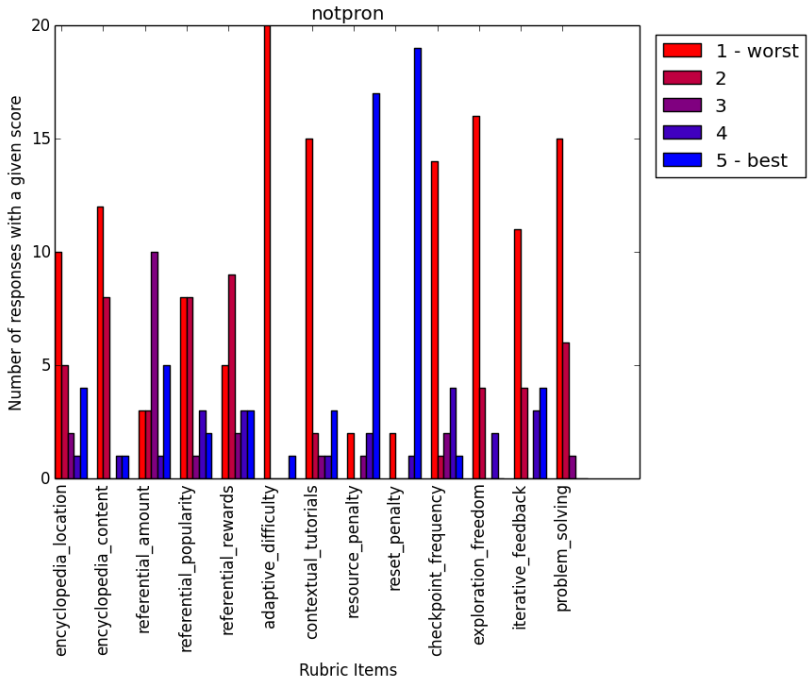

Figure 6.37: Notpron

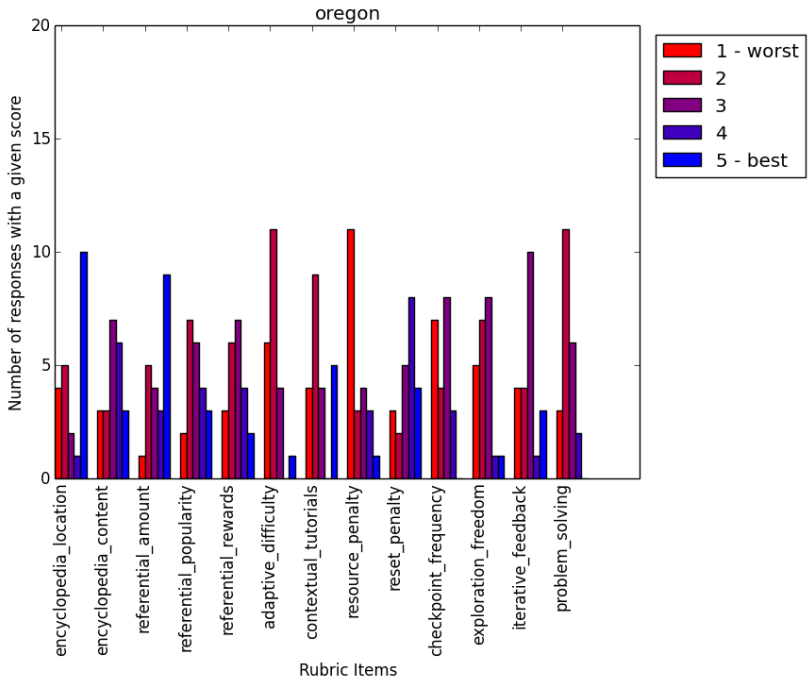

Figure 6.38: The Oregon Trail

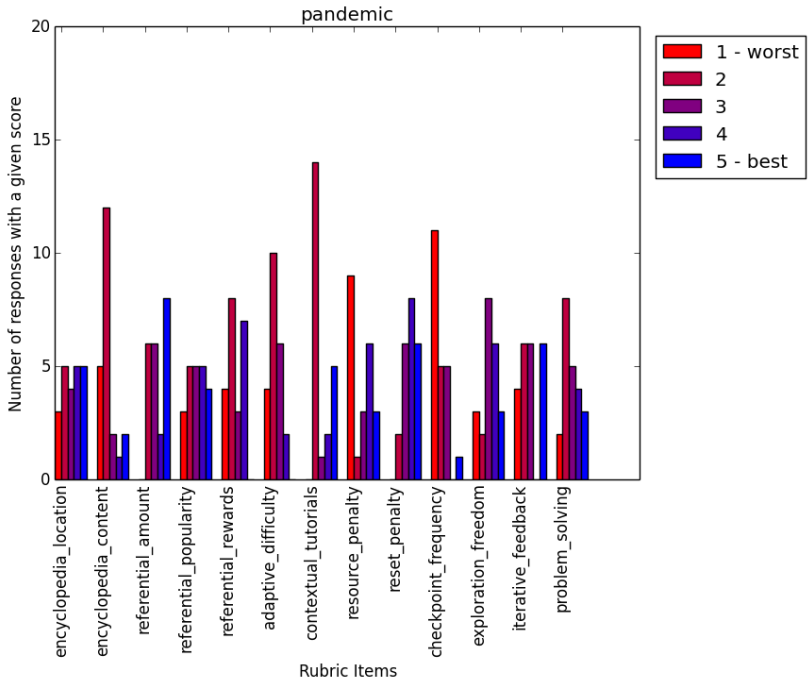

Figure 6.39: Pandemic 2

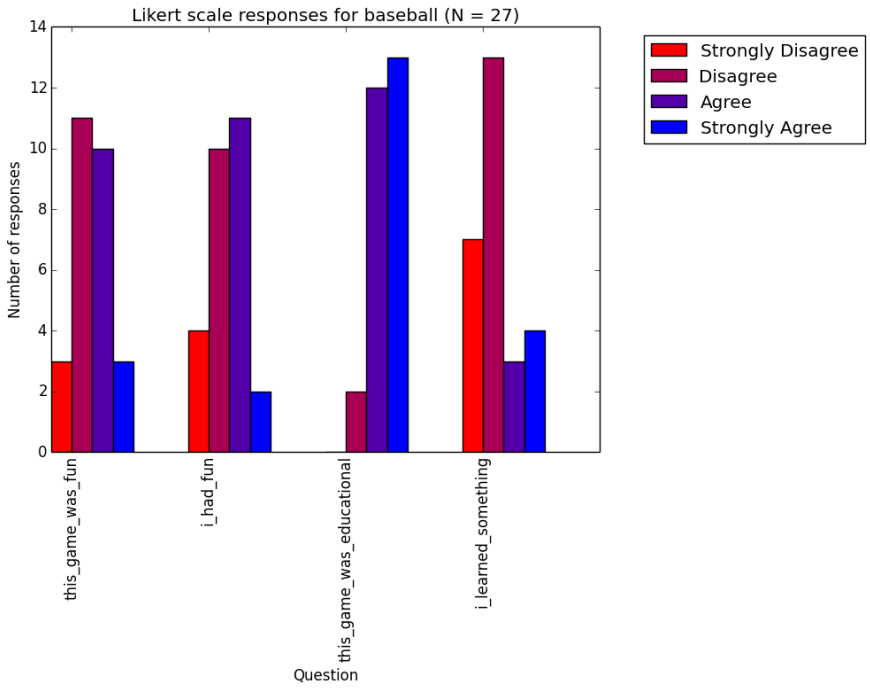

Figure 6.40: Math Baseball

6.5 Opinions on the fun and educational value of each game

At the end of the rubric section of the survey, players are asked four opinion questions; if the game was fun, if it was educational, if they had fun playing it, and if they learned anything while playing it. The responses were in Likert scale format.

6.5.1 Math Baseball

Players disagreed on whether or not *Math Baseball* was fun, but agreed that it was definitely considered an educational game. However, most players didn't learn anything from this game, presumably due to the low grade level. It's also important to consider that the game doesn't teach anything new; it is designed to test students on material they are already familiar with.

6.5.2 Botlogic

Players were largely ambivalent about the fun or educational content of *Botlogic*, but did not learn much from the game. This could be due to the low grade level the game aims to teach content at.

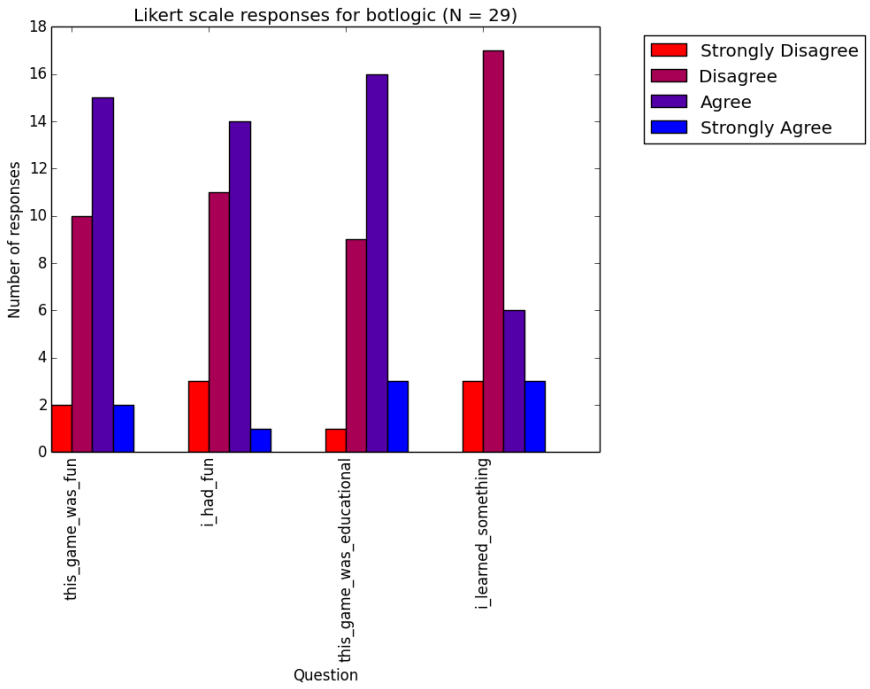

Figure 6.41: Botlogic

6.5.3 Darfur is Dying

Players agreed that *Darfur is Dying* was not a fun game, but it was an educational one, and one that they learned something from. It was effective at being educational, but likely not something players would willingly play.

6.5.4 Lemmings

Players agreed that *Lemmings* was fun, but had less agreement on the educational content of the game. As a primarily non-educational game, this is expected.

6.5.5 Light Bot

Players agreed that *Light Bot* was fun, and that they had fun playing it, but were very conflicted on whether or not it was an educational game.

This is unusual, because *Light Bot* was considered one of the more educational games selected. It's possible that most players didn't reach the 'educational' part of *Light Bot* (e.g. the functions and loops sections) due to their short play time.

6.5.6 The Incredible Machine

Players were ambivalent about *The Incredible Machine* being fun, as well as whether they had fun or learned anything while playing it. However, they did largely agree that it is an educational game.

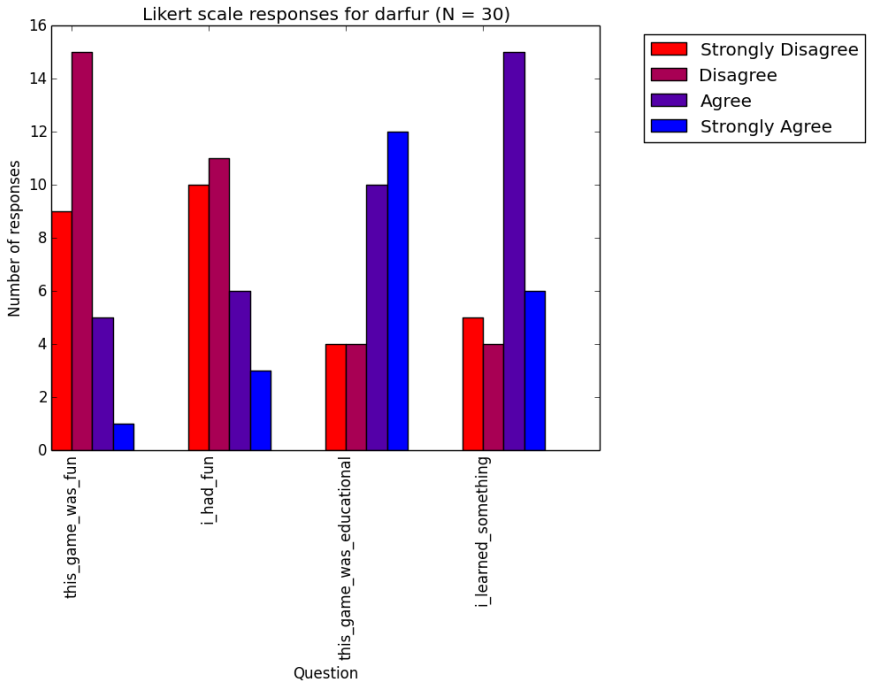

Figure 6.42: Darfur is Dying

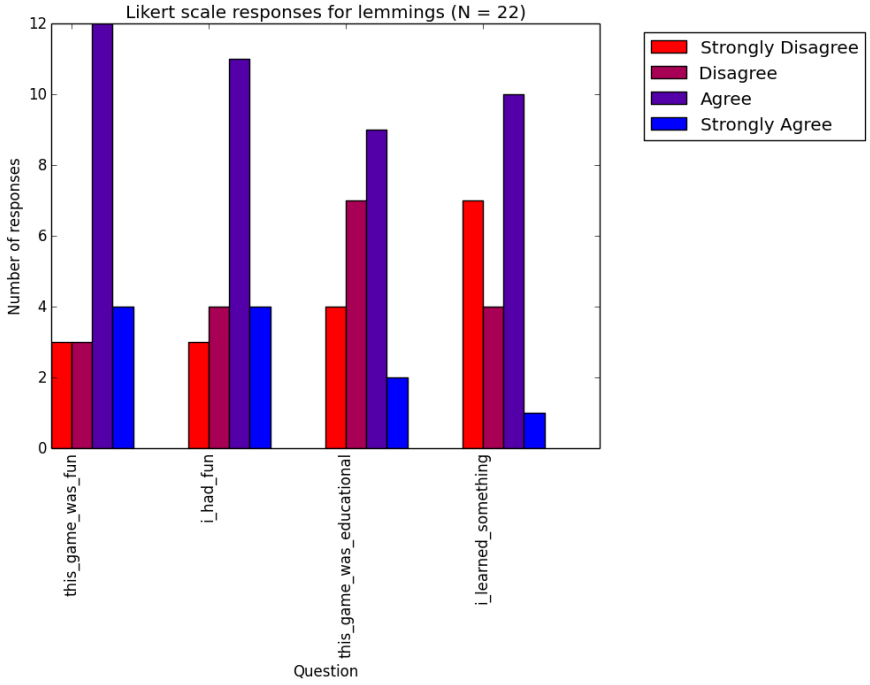

Figure 6.43: Lemmings

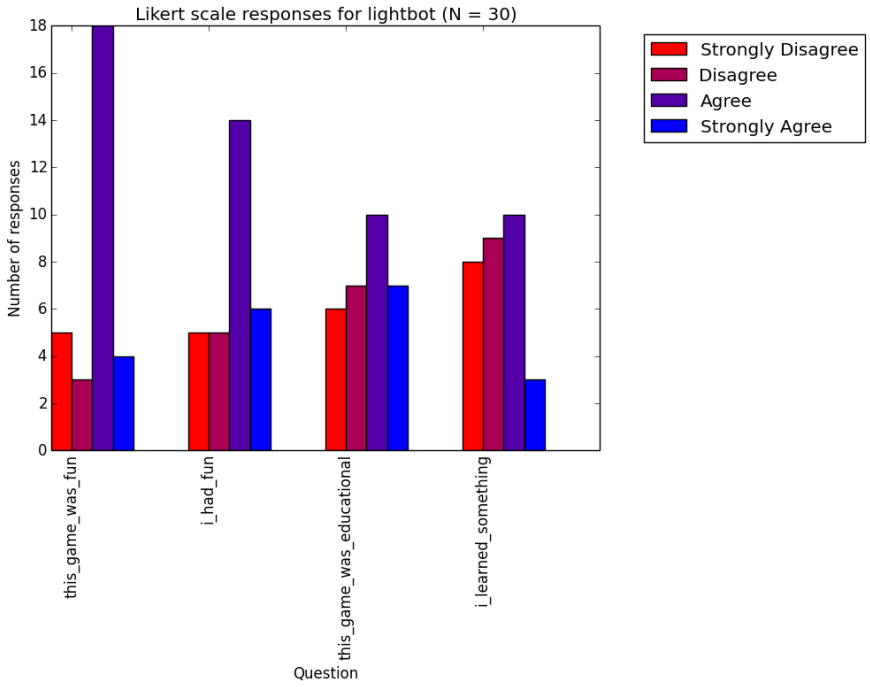

Figure 6.44: Light Bot

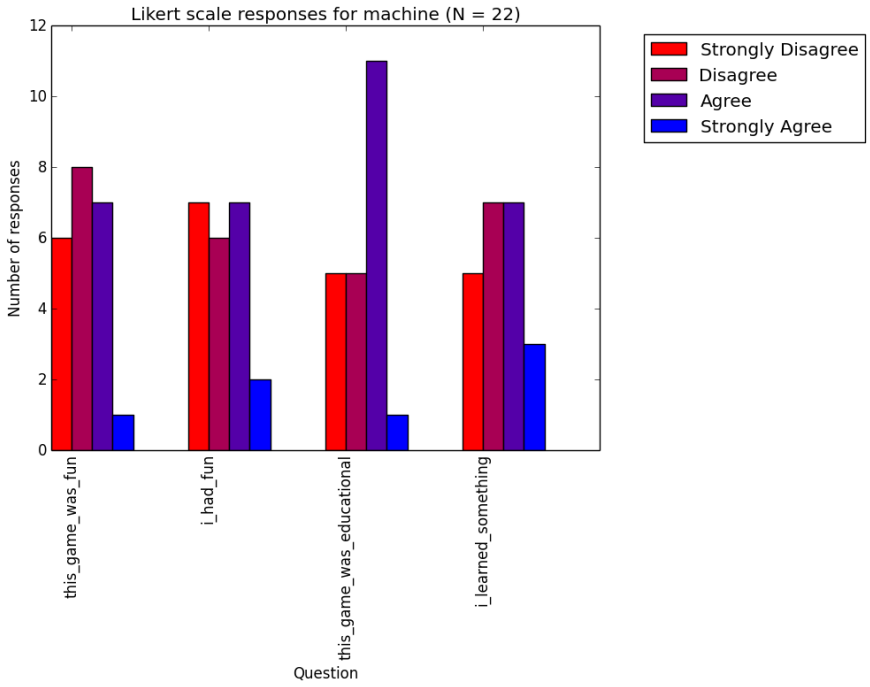

Figure 6.45: The Incredible Machine

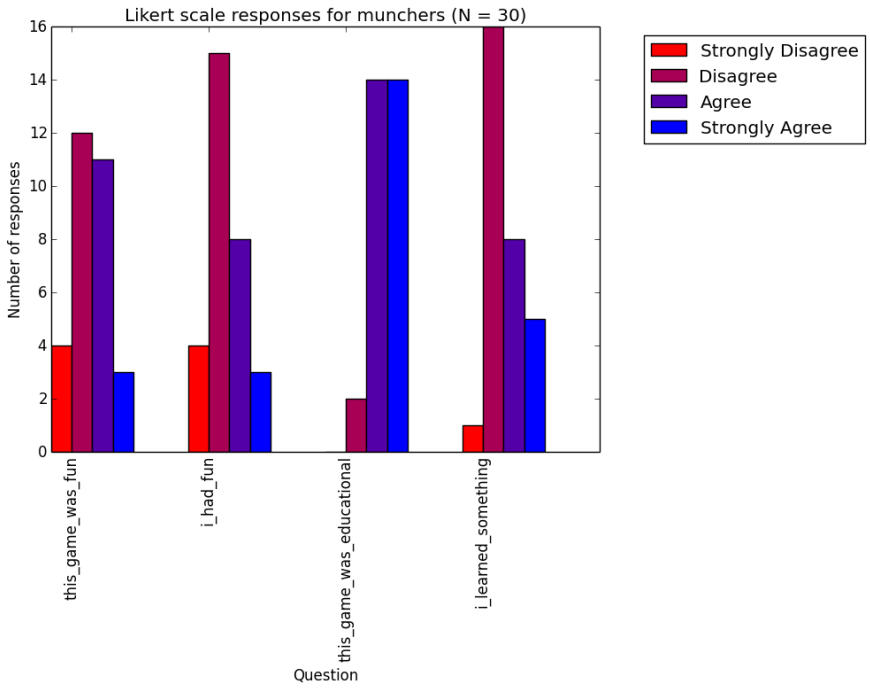

Figure 6.46: Number Munchers

6.5.7 Number Munchers

Players didn't have fun and didn't learn anything in *Number Munchers*, but they agreed that it is an educational game. *Number Munchers* teaches grade level math, so low scores on learning is expected.

6.5.8 Notpron

Players didn't have fun or learn anything in *Notpron*, and didn't consider it a fun game. However, they were somewhat conflicted as to whether or not it was educational.

6.5.9 The Oregon Trail

Players generally liked *The Oregon Trail*; they had fun playing it, and learned something while doing so.

6.5.10 Pandemic 2

This was one of the more unexpected results was with Pandemic 2. Across the board, players lightly agreed that *Pandemic 2* was a fun and educational game, and that they both had fun and learned something while playing it. The fun responses to the game were expected, but the educational ones were not; the game was selected due to its potential for tangential learning.

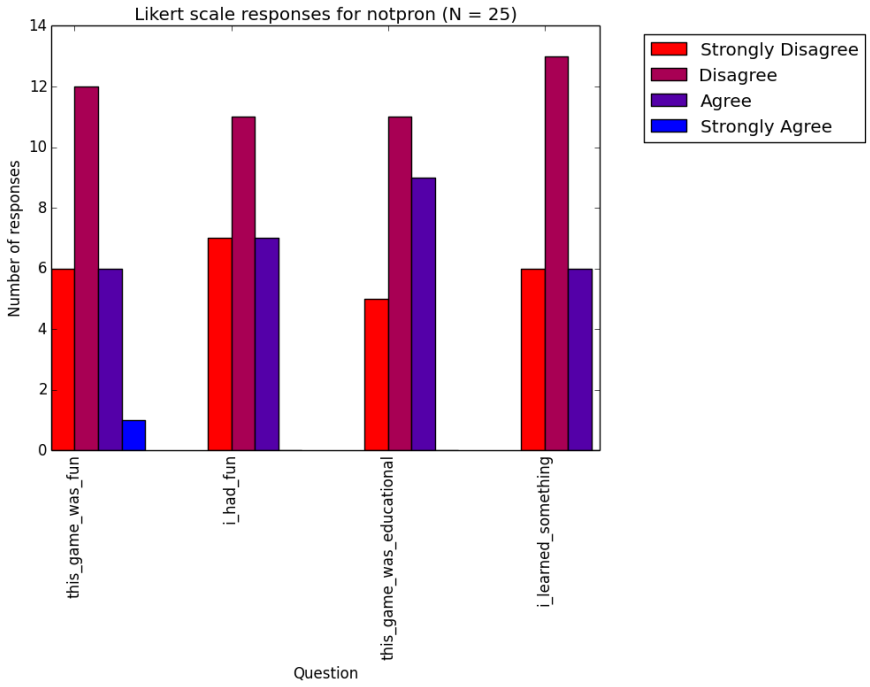

Figure 6.47: Notpron

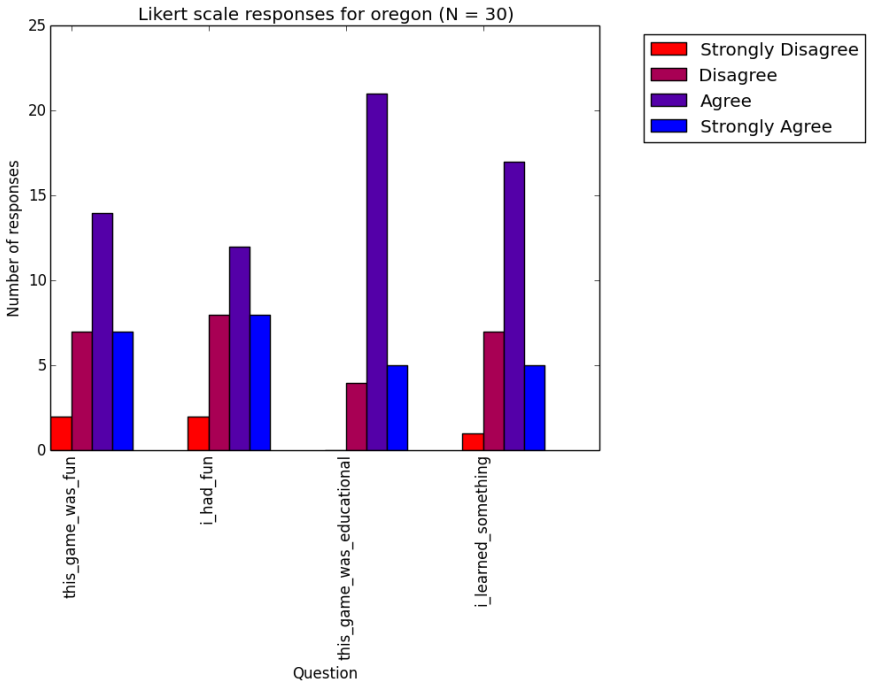

Figure 6.48: The Oregon Trail

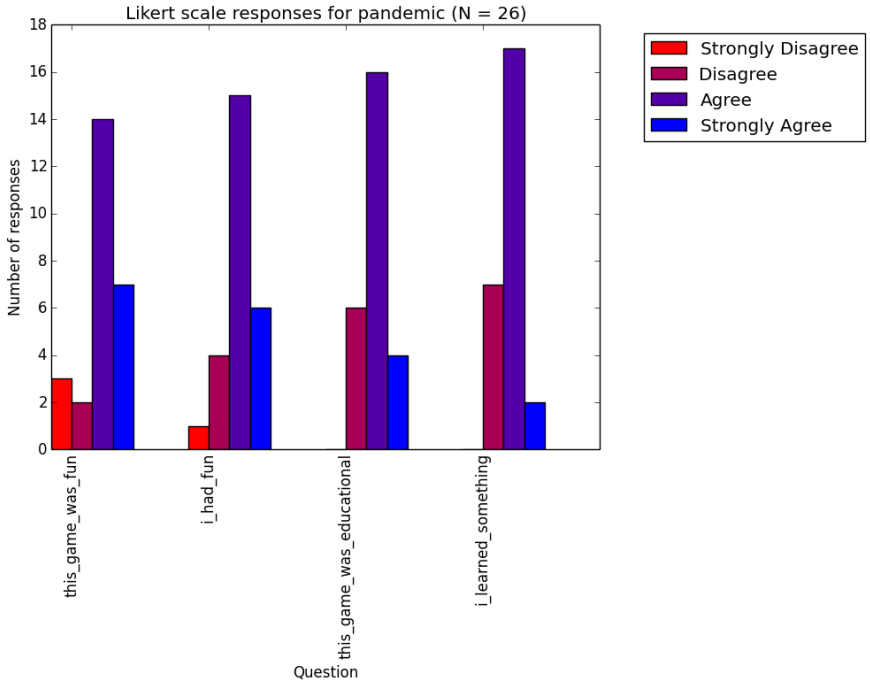

Figure 6.49: Pandemic 2

Chapter 7

Analysis

7.1 Quiz Scores

There are 4 games with quizzes, each with 30 responses each, 120 responses total. The mean quiz differences were analyzed for statistical significance using a two-tailed t-distribution.

7.1.1 How analysis using a t-distribution works

When evaluating statistical significance for paired samples, we have to use a test called a two-tailed t-distribution. More information can be found here [18] and here [19].

First, we take our paired samples. For us, these are our pairs of pre-quiz and post-quiz scores for each worker. We find the difference for each pair, which gives us a change for each worker.

Then, we calculate the mean and standard deviation of these quiz differences. Using either a 90% or 95% probability rating, we plug our number of raters, mean, and standard deviation into a two-tailed t-distribution. Increasing the number of samples will tend to reduce the uncertainty limits of the distribution. I used the number of raters minus one as our degrees of freedom.

The error bars displayed are our confidence limits - we are 90% or 95% sure that the true mean lies within these limits.

The null hypothesis, that our mean = 0, means that it's possible that our quiz had absolutely no affect. In order to invalidate that hypothesis, and prove that our quiz had a measurable effect, we want the error bars to not be touching zero; that way, the possible true mean cannot equal zero.

7.1.2 Analysis of Quiz Scores using t-distributions

Aggregated

If we aggregate the score differences across all quizzes, our results show that the true mean cannot equal zero, and thus say that our improvement in quiz scores

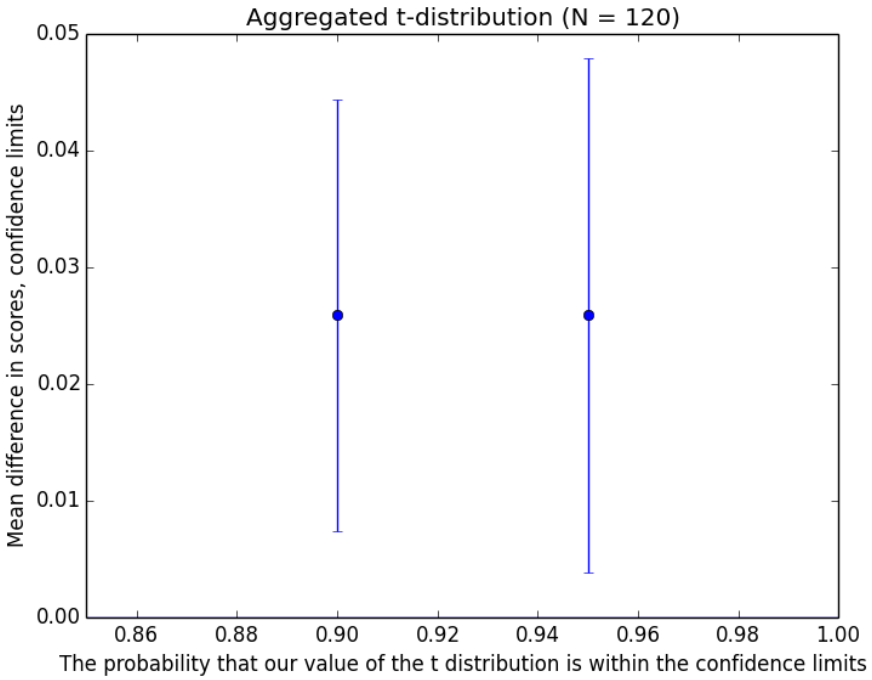

Figure 7.1: Aggregated tdist

are statistically significant. However, the quiz for each game contains different content and is laid out in a different manner, so we can't use the data aggregated across quizzes.

Darfur is Dying

We cannot eliminate the null hypothesis for our *Darfur is Dying* quiz, even if we reduce the probability to 0.9. This means that our quiz results for *Darfur is Dying* are not statistically significant.

The Oregon Trail

If we use p = 0.95, we cannot eliminate the null hypothesis for *The Oregon Trail*. However, if we reduce p to 0.9, we can eliminate the null hypothesis, and say that our quiz results for *The Oregon Trail* are statistically significant.

It's important to note that even if our results are statistically significant, the amount by which they are significant is extremely small; somewhere between a 2% to 10% increase in score.

Light Bot

We cannot eliminate the null hypothesis for *Light Bot*, even by reducing p. Our quiz results for *Light Bot* are not statistically significant.

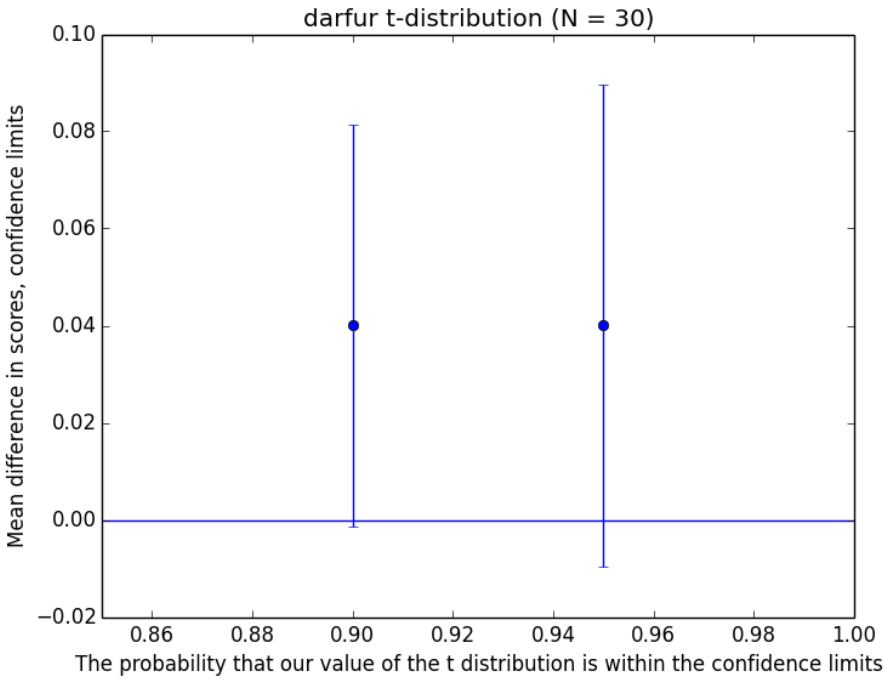

Figure 7.2: Darfur tdist

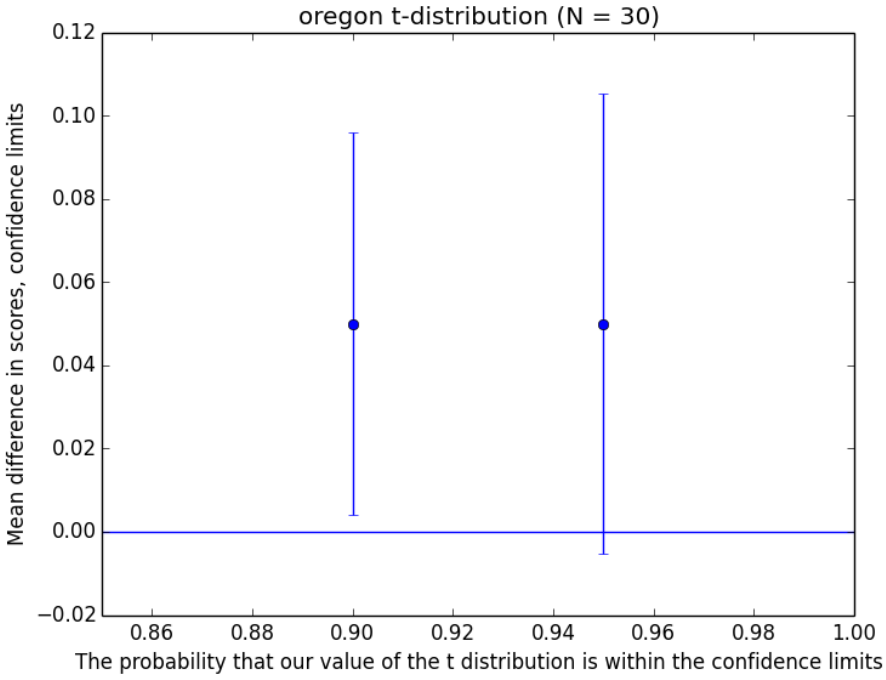

Figure 7.3: Oregon tdist

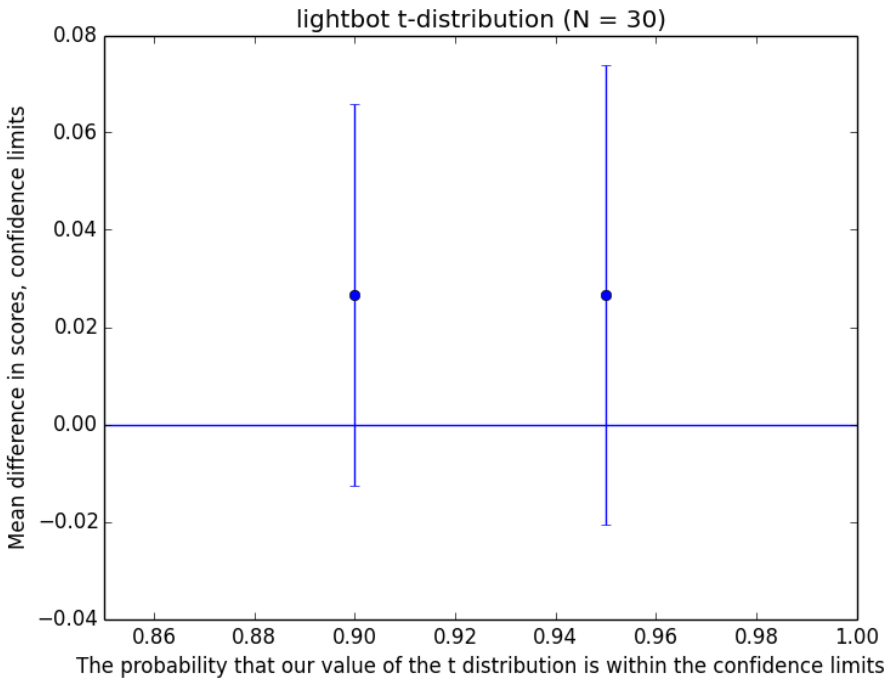

Figure 7.4: Lightbot tdist

Number Munchers

If we reduce p to 0.9, we can eliminate the null hypothesis from our *Number Munchers* results. However, it's very small, and it's actually in the negative direction; somewhere between a 0% and 2.5% decrease in score.

7.2 Rubric Inter-rater Reliability

7.2.1 How analysis using inter-rater reliability works

In order to ensure that the rubric I've built is consistent, we need to measure the responses we recieve using inter-rater reliability. More info can be found here and here.

Inter-rater reliability gives us a value (kappa) to indicate how consistent a given set of questions are. This is done by getting the results from several workers answering the questions, then analyzing how much they agreed on their responses.

For example, let's say I had a quiz with 50 True/False questions. If I had 2 people take the quiz, and they each got every single question correct (e.g. they agreed on every single question), the quiz would have a kappa value of 1, perfect agreement. If only one person got every single question correct, and the other got 0 questions correct (e.g. they disagreed on every single question), then the quiz would have a kappa value of -1, perfect disagreement. The values in between 1 and -1 indicate various levels of agreement, with 0 being complete randomness.

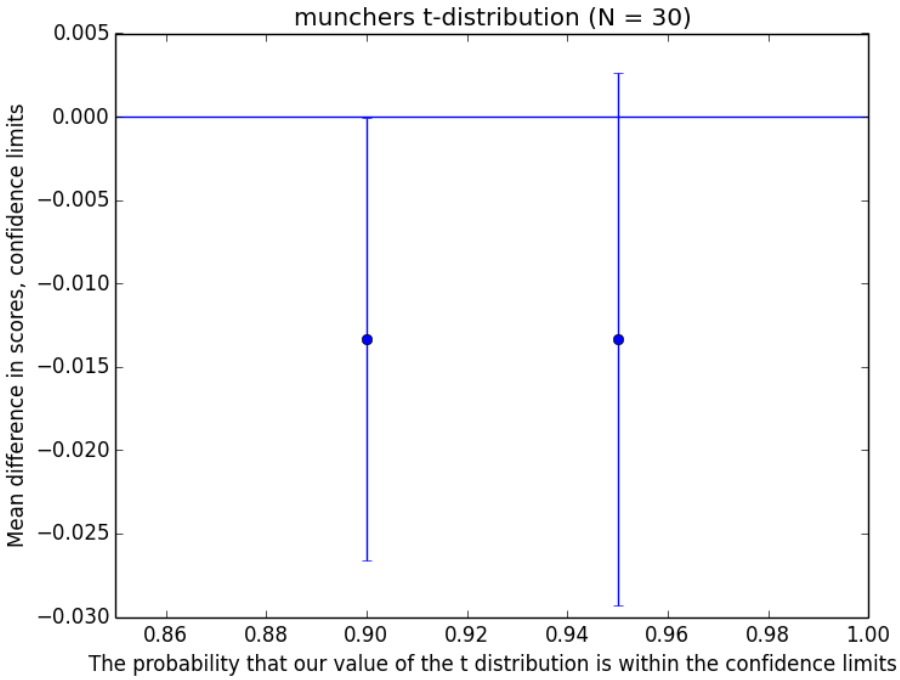

Figure 7.5: Munchers tdist

From Landis and Koch [20], here are the typical benchmarks for inter-rater reliability.

-1.0 Perfect Disagreement

-1.0 to -0.8 Almost Perfect Disagreement

-0.8 to -0.6 Substantial Disagreement

-0.6 to -0.4 Moderate Disagreement

-0.4 to -0.2 Fair Disagreement

-0.2 to 0.0 Slight Disagreement

0.0 Completely random

0.0 to 0.2 Slight Agreement

0.2 to 0.4 Fair Agreement

0.4 to 0.6 Moderate Agreement

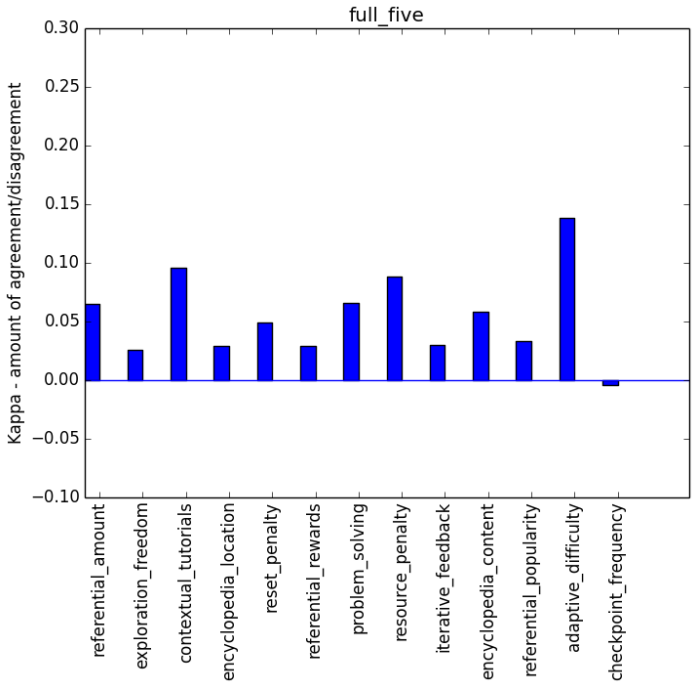

Figure 7.6: Inter-rater reliability

0.6 to 0.8 Substantial Agreement

0.8 to 1.0 Almost Perfect Agreement

1.0 Perfect Agreement

We want to use inter-rater agreement to evaluate the consistency of our design rubric. We'll use the method to evaluate multiple subjects with multiple raters and multiple-choice questions. Instead of looking at the design rubric as a whole document, we can actually examine each rubric item individually, and look at the responses it received across all games.

Think of it as if we are evaluating 13 different quizzes, one for each design rubric item. On each quiz, there are 10 questions, one for each game to be evaluated. Each question has 5 single-selection options, where each option corresponds to a level on the scale of the rubric item.

For each graph, each bar indicates a rubric item, with it's height above or below the x-axis indicating its kappa value (up to +1.0 for agreement, up to -1.0 for disagreement).

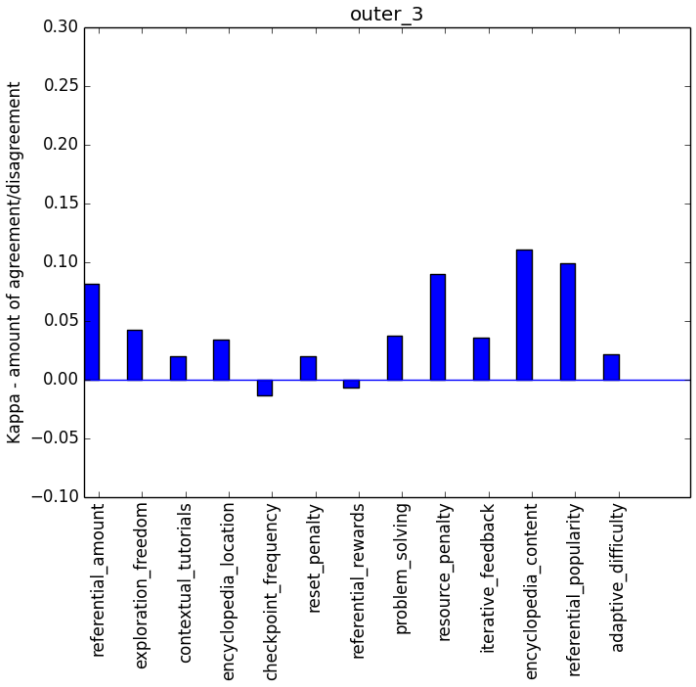

Figure 7.7: Inter-rater reliability with outer pairs combined

7.2.2 Analysis for Rubric items using inter-rater reliability

The kappa values for each rubric item were very small. All but one of the rubric items fell in the "Slight Agreement" category, with Adaptive Difficulty being the highest with a kappa value of 0.15. Checkpoint Frequency was the only negative rubric item.

This was unexpected and disappointing. After some further consideration, I decided to process the data slightly differently, to see if it would give me a different result.

Inter-rater reliability metrics commonly depend on a low number of available options (2 to 3) for workers to choose when taking a quiz. In our above example, our 50 question True/False quiz only contained 2 options for workers to choose from on each question; the rubric that we gave our workers contained 5 options for each question.

In order to achieve a different result, I ran two additonal inter-rater reliability calculations. With one, I summed the outside pairs together (e.g. [1, 2, 3, 4, 5] became [sum(1, 2), 3, sum(4, 5)]) to give us a result with only 3 choices that workers chose from. For the other calculation, I summed the inner 3 choices(e.g. [1, 2, 3, 4, 5] became [1, sum(2, 3, 4), 5]), to again give a result with only 3 choices.

Summing the outer pairs of options doesn't change our scores significantly, but summing the inner scores does. By considering the middle 3 options as a single option, Adaptive Difficulty reaches the "Fair Agreement" category with

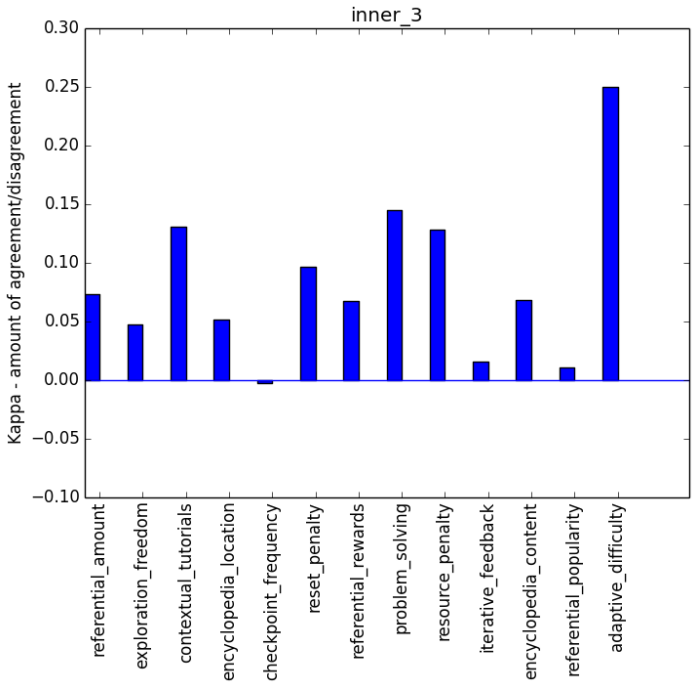

Figure 7.8: Inter-rater reliability with middle 3 options combined

a kappa value of 0.25. Contextual Tutorials, Unorthodox Problem-solving, and Game Resource Penalty all still fall in the "Slight Agreement" category, but they have improved kappa values of above 0.10.

Chapter 8

Conclusion

For our four games with quizzes (*Darfur is Dying, The Oregon Trail, Light Bot,* and *Number Munchers*), only *The Oregon Trail* and *Number Munchers* had statistically significant results. The Oregon Trail's was positive, with a mean improvement of 5%. *Number Munchers*' was negative, with a mean difference of -1%.

For our design rubric, the four design elements with the highest value of kappa (i.e., inter-rater reliability) were adaptive difficulty, Contextual Tutorials, Unorthodox Problem-Solving and Game Resource Penality. The value of kappa for Adaptive Difficulty fell into the category of "Fair Agreement," while the others fell into the category of "Slight Agreement". For Adaptive Difficulty, Contextual Tutorials, and Unorthodox Problem-solving, the scores on each of the games were universally negative or mixed.

For the rubric item "Game Resource Penalty," three games (*Notpron, Botlogic and Light Bot*) had relatively positive scores, and three games (*The Oregon Trail, Number Munchers, and Pandemic 2*) had relatively negative scores.

From these results we can conclude the following. *The Oregon Trail* was the only game that had a statistically significant improvement in quiz scores. Of the four design elements that had the highest inter-rater reliability, all had mixed scores except for Game Resource Penalty, which had primarily negative scores. Therefore, we could conclude that an educational game should have the maximum amount of Game Resource Penalty to be the most effective.

Punishing players for failure seems very contradictory to existing research. This conclusion comes with numerous caveats; the small difference in quiz scores, the extremely small inter-rater agreement values, and other unknown game design elements contribute to this being a result of no confidence. There are many unknown variables; the format of the game, quiz, survey, and administration process can all contribute to inconclusive work.

It is my opinion that we can derive no significant results from the data without further analysis. However, there are numerous learnings considering the methods we used to obtain these results, and how we can optimize them to create better results.

8.1 What worked well

8.1.1 Web-based games

Using web-based games to test worked relatively well; while we were not able to find original versions of some educational games, online emulators and generic versions of the games let us avoid installing software on the workers' machines.

8.1.2 Inter-rater Reliability

Inter-rater Reliability turned out to be a very effective way to measure how well the rubric was designed. Even if it was not designed very well, it was fortunate that I found papers (cite, city) detailing exactly what was needed for this research; an extension of the inter-rater reliability metric, using multiple raters across multiple games. It applied perfectly to my data.

8.1.3 Mechanical Turk: Volume

Using Mechanical Turk to get a large number of responses worked very, very well. It allowed us to carefully manage the number of responses we received, as well as incrementally process results to make sure our system was working correctly.

It was extremely cheap to get a large number of responses; I paid a total of $60 for 300 responses, and further research could gain more responses per dollar by limiting the number of games and the size of the task.

8.2 What did not work well

8.2.1 Quiz Design

There are a number of reasons why the results may have fallen so flat.

The quiz itself may have been faulty; perhaps the questions were too hard, too easy, or covered content that was not taught as part of the game. To combat this, the quizzes were designed to cover material that players were supposed to learn in the game, but it is possible that the multiple-choice format wasn't the best way to do that.

It is also possible that the games were not actually effective in teaching the players anything at all. Because of how the games were designed, players did not retain any relevant information about what the game was trying to educate them on.

8.2.2 Rubric Elements

It appears that there was little agreement on the scores rubric elements. This indicates that the prompts and descriptions were vague or confusing, and that scores would not be similar if contributed by multiple workers.

The two elements that must be considered are the writing and coherence of the rubric, and the quality of the rubric items. My opinion is that the quality or writing style of the prompts was of poor quality, and caused the majority of disagreement between raters.

I think that the content that the rubric is derived from is sound. The rubric needs some adjustment, but there are good examples of games for each rubric item. In addition, there is clear (but not statistical) evidence that certain rubric items are clearly present in certain games; with adjustment of the descriptions and prompts, there could be greater agreement.

8.2.3 Mechanical Turk: Quality

Another option is that the platform on which the surveys were administered has faults. There are spammers on Mechanical Turk, and while it is possible to review and reject responses deemed unusable, it is impractical to filter out every single one. However, if a spammer is selecting random responses, it is highly unlikely that they'll end up with a higher or lower score than when they started.

It is also possible to filter out poorly rated workers with an "acceptance rate" criteria, where Workers that accept the task are required to have at least a 90% acceptance rate across all of their HITs, but that increases response times and payout amounts.

Finally, it is possible that some workers lost interest in the HIT once they had finished the first sections, resulting in shoddy work on the second. This would mean that some workers would guess at the questions in the post-quiz, resulting in worse performance across the board.

8.3 Personal learnings

8.3.1 Mechanical Turk

Amazon's Mechanical Turk is an extremely powerful tool. It allows us to take tasks that are difficult or impossible to do (like rating a game) and scaling them immediately and effectively. We can allocate for more humans to complete our tasks much how we would allocate more servers to handle more requests.

Mechanical Turk has a GUI, API, and command-line tools for requesters to access. Because the GUI is limited in functionality specifically for the type of research I was doing, I had to use the API and command line tools to build and test all the surveys. This involved extensive research on documentation, experimentation, and debugging on the Mechanical Turk platform.

8.3.2 LaTeX

As part of conducting analysis alongside the experiment, a way was needed to generate the thesis document programmatically (e.g. without re-inserting the

graphs every time the data changed). For this, LaTeXwas perfect; documents could be compiled via command line, and would use whatever versions of files were currently in the directory. I had never used it before, but the documentation and examples online were extensive.

LaTeXhad some problems. Throughout the course of this project, the large number of images and graphs wreaked havoc with the layout system. With the addition of each new set of images or graphs, there were extremely long compile times until the images could be wrangled into a format that LaTeXcould generate quickly. I feel like I've learned a great deal about LaTeX, both the generalized knowledge of how to create good documents, but also the inner workings of images, generation, floats, section structure and the bibliography.

8.3.3 What it is like to conduct research

One of the goals for choosing to do a Senior Thesis (instead of a Senior Project) was to evaluate my own desire to perform research or attend grad school. Because I had experience doing several large group projects, I wanted to test and see if I would like performing research at the graduate school level.

I've learned a tremendous amount about the nature of conducting research. Notably, the difference between generating a highly-opinionated (but nonetheless experimentally unfounded) report on the nature of educational gaming, and conducting actual research by developing an exploratory hypothesis and running some experiments.

8.4 Options for future research

8.4.1 An extension of this research

This research generated some interesting but ultimately unusable results. However, there were many learnings on the rubric items, as well as how to use Mechanical Turk effectively. If the number of games and rubric items were reduced, while increasing the quality of the quizzes as well as ensuring better quality responses in general, there could be a great increase in both the measure inter-rater reliability and the quiz scores.

For a future quiz/survey on the Mechanical Turk platform, I'd recommend both an "acceptance rate" criteria, as well as some form of automated validation that the worker is actually completing the survey. This may include a "dummy question," where the reviewer only has to click the right button, or some other kind of verification where they submit a screenshot of their highest score within the game. This will increase response times and require higher payout amounts, but if it is focused only on one game it won't be much.

Another route worth exploring is paying out "bonuses" to workers who show exemplary work in the quiz and survey. However, you'd have to make sure to

incentivize honesty and not greed; my hunch is that asking workers to not use the internet to find solutions may not work when there is extra payouts involved.

In addition, it is possible to create "qualifications" for Mechanical Turk workers. First, the workers must complete a test where they are trying to achieve the highest score possible. This test could consist of scoring a game where the correct answers to each question are very clear. If they pass the test, then they are considered "qualified" to score the rest of the games. This would filter out workers who obviously do not understand the questions or rubric.

8.4.2 Iterative scalable design testing for educational games

Mechanical Turk could be an extremely effective tool for anyone who is doing development or research on the development of educational games. For example, someone is creating a web-based educational game, and they have built in logging tools that automatically send player data (scores, actions, levels completed) to a server. They could cheaply and easily use Mechanical Turk to get workers to play the game for at least 10 minutes, or more; they have a way to track if the player actually completed the game, as well as how many questions they have gotten right, allowing them to track how much players have learned.

Bibliography

[1] Bloom, Benjamin S. "Taxonomy of Educational Objectives; the Classification of Educational Goals,". New York: Longmans, Green, 1956. Print.

[2] Breuer, Johannes, and Gary Bente. "Why so Serious? On the Relation of Serious Games and Learning." Eludamos. Journal for Computer Game Culture 4.1 (2010): 7-24. Eludamos. Journal for Computer Game Culture. Web. 02 July 2013.

[3] Buhrmester, Michael. "Amazon Mechanical Turk Guide for Social Scientists." Web. <http://homepage.psy.utexas.edu/homepage/faculty/gosling/reprints/MTurkhowto.pdf>.

[4] Coe, Robert. "It's the Effect Size, Stupid: What Effect Size Is and Why It Is Important." It's the Effect Size, Stupid: What Effect Size Is and Why It Is Important. British Education Index, 12 Sept. 2002. Web. 21 Nov. 2013.

[5] Csikszentmihalyi, Mihaly. "Finding Flow." Psychology Today. 14 June 2012. Web. 02 July 2013.

[6] Dondlinger, Mary Jo. "Educational Video Game Design: A Review of the Literature." Journal of Applied Educational Technology 4.1 (2007). Print.

[7] Floyd, Daniel, and James Portnow. "Tangential Learning." Season 2, Ep. 9 - Tangential Learning. Web. 02 July 2013.

[8] "Genocide in Darfur." Game. Fun Trivia. Web. 21 Nov. 2013. <http://www.funtrivia.com/playquiz/quiz2560151d4fc10.html>.

[9] Hallgren, Kevin A. "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial." NCBI. U.S. National Library of Medicine, 23 July 2012. Web. 21 Nov. 2013.

[10] Hyman, Paul. "Communications of the ACM." Software Aims to Ensure Fairness in Crowdsourcing Projects. Communications of the ACM, Aug. 2013. Web. 21 Nov. 2013.

[11] Kjell, Bradley. "Counting Loop Quiz." Counting Loop Quiz. Web. 21 Nov. 2013. <http://www.cs.iastate.edu/~honavar/JavaNotes/Notes/chap16/chap16quiz.html>.

[12] "The Measurement of Inter-rater Agreement." Web. <http://hpm.fk.ugm.ac.id/hpmlama/images/Biostatistik/Tutorial_4_AS/2-chapter18.pdf>.

[13] Paras, Brad, and Jim Bizzocchi. "Game, Motivation, and Effective Learning: An Integrated Model for Educational Game Design." Proc. of DiGRA 2005 Conference: Changing Views Worlds in Play. 2 June 2005. Web. 2 July 2013.

[14] "Pioneer Life." ThinkQuest. Oracle Foundation. Web. 21 Nov. 2013. <http://library.thinkquest.org/J001587/>.

[15] "Quiz: The Oregon Trail." Quia. Quia. Web. 21 Nov. 2013. <http://www.quia.com/quiz/462983.html?AP_rand=227719871>.

[16] "QuizMoz - Oregon Trail Quiz." QuizMoz - Oregon Trail Quiz. Web. 21 Nov. 2013. <http://www.quizmoz.com/quizzes/Interesting-Facts-Quizzes/o/Oregon-Trail-Quiz.asp>.

[17] "What Do You Know about the Darfur Genocide?" What Do You Know about the Darfur Genocide? ProProfs. Web. 21 Nov. 2013. <http://www.proprofs.com/quiz-school/story.php?title=what-do-you-know-about-darfur-genocide>.

[18] Ammann, Larry. "Hypothesis Tests to Compare Means of Two Populations - Paired Samples." Hypothesis Tests to Compare Means of Two Populations - Paired Samples. N.p., 20 Nov. 2013. Web. 05 Dec. 2013. <http://www.utdallas.edu/~ammann/stat3355/node35.html >.

[19] Lane, David M. "T Distribution." T Distribution. N.p., n.d. Web. 05 Dec. 2013. <http://onlinestatbook.com/2/estimation/t_distribution.html >.

[20] Gwet, Kilem L. "Benchmarking Inter-Rater Reliability Coecients." Handbook of Inter-Rater Reliability. N.p., 2012. Web. 5 Dec. 2013. <http://www.agreestat.com/book3/bookexcerpts/chapter6.pdf >.