Preferred Percent Correct in Novel Computer-Based Tasks

Thesis presented to the Faculty of

Notre Dame de Namur University

In partial fulfillment of the requirements for the degree of

Masters of Science in Clinical Psychology

by

Steven Riley

Approved by:

____

Michele Haley, Ph.D., Chairperson

____

Matthew Harris, Ph.D., Chairperson

____

Helen Marlo, Ph.D., Second Reader

____

1/1/2015

Abstract

This research investigated whether computer users preferred moderate levels of difficulty when encountering new tasks. This research attempted to quantify the preferred percentage correct (PPC) on three computer-based tasks (one verbal, one visual and one kinesthetic) for a sample of Internet users. PPC was defined as the point where participants gave no net preference to the change in difficulty level upon completion of ten trials of a computer- based task with fixed difficulty level. The sample included 102 individuals of mean age 24.1 who were recruited through posts on Reddit, Facebook and Craigslist. Most participants were White (75%) and low-income (66% had incomes of less than $25,000/year), and about half of them were female (46%). Sixty-two of these participants spent at least five hours per week on average playing video games. Two constructs were used to measure PPC with significant evidence for convergent validity. Correcting for guessing, the data yielded PPC values of 58.7%, 54.5%, and 56.7% on the visual, verbal and kinesthetic tasks, respectively, all with standard errors < 3%. These data fit the relationship PPC = .57 + .43*g, where g is the chance of a correct guess on a task.

Table of Contents

List of Tables

Chapter I: Literature Review

Preferred Difficulty Levels in Novel Computer-Based Tasks

The average American gamer spends approximately $140 per year on games, and one in seven Americans spends at least 5 hours per week playing video games (Takahasi, 2010). There is significant financial incentive for video game companies to attempt to "corner the market" on game-related fun. In pursuing the business of gamers, companies that provide consistent entertainment to the video gaming public have likely learned how to craft their games to suit gamers' desire for challenge.

Even some non-gaming businesses have begun to employ game-like paradigms in order to keep employees (Silverman, 2013) and customers motivated. Over the last few years, the technique of "gamification," the utilization of game-similar qualities to motivate employees and customers to engage in activities desired by the employer or corporations (Silverman, 2013), has become mainstream (Amos, 2013). "Gamification," as defined by Deterding, Dixon, Khaled, & Nacke (2011), is the technique of offering game-similar rewards (i.e., badges, levels and achievements) and making those rewards contingent upon the performance of the user. For instance, Duolingo (a language learning website; www.duolingo.com) rewards the user with points for every completed exercise. If the user achieves a certain point total, that user's level increases. When a user completes a block of lessons, a badge is awarded and a notification is posted to Facebook and the user gains public acknowledgment that he/she has reached a certain level of language proficiency. While the correlation between gamification of an activity and user motivation to engage in the activity seem to be positive, these results are highly dependent upon both the context and the user (Hamari, Koivisto, & Sarsa, 2014). This research sought to explicitly quantify some of the information that might be implicit in the designs of successful games and gamified systems. Specifically, it sought to answer the question "what is the level of difficulty preferred by users involved in gamified activities?"

Fun. What is meant by "fun?" This research defined "fun" as the level of enjoyment a participant reported while engaged in the execution of a task. This paper examined three different theories of fun or optimal experience and demonstrated that they predict a similar phenomenon: the preferred difficulty level chosen by a user should be high enough to be interesting but not so high as to be impossible. This research investigated participants' preferred levels of difficulty in terms of their percentage of correct responses on a task in order to determine how often individuals want to get items correct while doing a task. This information was used to inform conclusions as to how the behavior of an individual may change when given a task that is either too hard (participants make too many mistakes) or too easy (participants make too few mistakes).

The three theories of fun integrated into this study include one based on affect regulation, one based on learning, and one based on a combination of affect and learning. The model based on affect regulation will examine Csikszentmihalyi's "flow" theory (Csikszentmihalyi, 1990) and what it predicts about behavior. The model based on learning will look at Schmidhuber's formal theory of creativity, fun, and intrinsic motivation (Schmidhuber, 2010). The model that combines the two is based on Vygotsky's theory of zone of proximal development (Vygotsky, 1978) and the Yerkes-Dodson law (Yerkes & Dodson, 1908).

Csikszentmihalyi's flow. Csikszentmihalyi's conceptualization of enjoyment for an individual engaging in a task includes the following eight factors: one, the task can be completed; two, the individual can concentrate on the task; three, the task has clear goals; four, the task gives immediate feedback; five, while doing the task, the individual's attention focuses to the point of forgetting the stresses of daily life; six, the individual feels a sense of control during the task; seven, ego and self-concern disappear; and eight, the sense of duration of time alters (Csikszentmihalyi, 1990). Not all eight factors must be present for enjoyment to occur. When the majority of these factors are present, one can be said to be in a state of "flow."

One of the corollaries of these conditions is, in order to enter a flow state, an individual must experience a balance between the perceived difficulty of the task and the participant's self-perceived skill level (Csikszentmihalyi, Abuhamdeh, Nakamura, Flow & Dweck, 2005). Thus, flow states require the difficulty of a task to closely match the participant's preferred difficulty level for that task. If the task is too easy, boredom results. If it is too hard, anxiety results. In between areas of anxiety and boredom is the possibility for the participant to experience flow.

Haworth and Stephen (1995) demonstrated that participants rated tasks that are too easy as more enjoyable than tasks that are too hard. Additionally, participants desired a match between perceived personal skill and challenge level (Haworth and Stephen, 1995). These findings predicted that participants would enjoy tasks with a difficulty level where the majority of tasks would yield successful results. However, one of the conclusions of the flow theory is that task participants want challenges. If a participant wanted to get 100% correct on a task, an optimal strategy would be to decrease the difficulty to its lowest point. But, the theory of flow predicted that this would not happen, and thus players would risk getting some percent of the tasks wrong in order to maintain a sense of enjoyment. Thus, it was considered likely that PPC values would be above 50%, but less than 100%.

Schmidhuber's formal theory of fun. Schmidhuber (2007) defined beauty as the compressibility of a data set. He claimed that the beauty we see in a face, a poem, a sunset or a theory all stem from the same property - something about the observation allows us to store a large amount of information with a small amount of memory. Hudson (2011) tested this connection by comparing compression ratios of pieces of music that had been run through various lossless compression algorithms. He found that those pieces that were of a style considered typically “beautiful” (such as choral and orchestral masterpieces) where more highly compressible than pop, rock, techno and random noise.

Schmidhuber goes on to define interestingness, novelty, surprise, or fun as the change with respect to time of the required memory footprint for storing information about an object or idea. In other words, something is fun or interesting if it allows an agent to see beauty in a new place (i.e. to compress data which previously seemed incompressible). Well before Schmidhuber developed his theory, Davis (1971) ran a study attempting to figure out which theories people found “interesting”. Davis concluded that uninteresting theories confirmed an assumption of the participant, but interesting theories denied an assumption. Meaning, those ideas which participants found interesting were ones that allowed participants to recode their views of the world.

Schmidhuber says that in order to find something interesting (i.e. compressible), the observations an agent makes about the world must be neither predictably regular nor random. Instead, data from such observations must be “regular in a way that is new with respect to the observer's current knowledge, yet learnable” (Schmidhuber, 2010, “How the Theory Explains Art” section, para. 1). What does this say about an agent’s preferred difficulty level for a novel task? If fun is proportional to the ability to compress observational data, then in order to increase fun it is necessary to increase the rate at which an agent learns how to incorporate data into a new compression scheme. Meaning, the measure of fun is the rate at which the agent learns. If a series of tasks is so easy as to not require an agent to extend its capabilities, then that set of tasks will not be fun for the agent. Similarly, if the agent has so little knowledge or ability that the outcomes of a task seem random, there will be no compression of data and therefore no fun. Schmidhuber's theory of fun leads to the conclusion that a moderate level of difficulty is ideal to produce the maximal amount of fun.

Frustration, arousal, Yerkes-Dodson, and the zone of proximal development. While the idea of frustration might normally have negative associations, this paper will now argue that frustration can be an important aspect of fun. What is frustration? Dollard, Miller, Doob, Mowrer and Sears (1939) define frustration as the result of "an interference with the occurrence of an instigated goal-response." (Dollard, Miller, Doob, Mowrer and Sears, 1939, p.8) In other words, frustration is the result of failure at a task at which an individual desired success.

The Yerkes-Dodson law states that as arousal increases, the ability of an organism to complete a difficult task initially increases, then hits a maximum, then decreases (Yerkes & Dodson, 1908). The Yerkes-Dodson law has been applied in numerous environments, including predicting which obese children will lose weight (Johnston, Moreno, Regas, Tyler & Foreyt, 2012), describing how cognitive abilities vary in response to psychostimulants (Wood, Sage, Shuman & Anagnostaras, 2014), and explaining how the validity of witness testimony changes with the nature of the crime (Roberts, 2014).

As argued previously in the section on Schmidhuber's formal theory of fun, the interestingness of a task is related to the speed at which a person can develop a method of compressing information about the task (i.e., the speed at which the agent learns the task). According to Vygotsky's "zone of proximal development," learning takes place most effectively at the border between where an individual can complete a task on his or her own and where the individual cannot complete the task (Vygotsky, 1978). Since an individual will have his or her greatest level of performance on a task at a middling level of arousal, it follows that the difficulty level where the highest rate of learning would occur is where the amount of arousal generated equals this middling level. Thus, finding the difficulty level that corresponds to the peak on the Yerkes-Dodson curve amounts to finding the highest amount of fun for a task.

Since frustration can increase arousal levels, it follows that having some small amount of frustration can lead to higher performance, and therefore more fun. Meaning, a certain amount of frustration stemming from getting some percentage of items wrong on a task will lead to more fun than getting all the items right or all the items wrong. Too many items wrong increases frustration (and hence arousal) past the level of maximum ability on a task; too few wrong leads to decreased arousal below the peak. In conclusion, the relationship between frustration, arousal, the Yerkes-Dodson law and Vygotsky's zone of proximal development leads to the prediction that there exists a non-zero percentage of wrong answers that will lead to the most fun.

Neuroticism. Individuals with high Neuroticism respond poorly to stressors, exaggerate the threat of neutral situations, and are unable to deal with minor frustrations (Ormel, et al., 2013). Since frustration plays a role in arousal, and arousal affects performance (according to the Yerkes-Dodson law), individual responses to frustration might affect a participant's response to a task. Thus, it is logical to believe that the frustration aversion that we would expect from individuals with high Neuroticism would lead to an aversion to getting answers wrong on a task.

A link between Neuroticism and increased avoidance of ambiguous stimuli has been shown in the literature (Lommen, Engelhard, & van den Hout, 2010), at least in the case where there was risk of a small electric shock. Since high Neuroticism individuals show increased aversion to frustration, it would make sense that individuals with high Neuroticism traits would avoid frustrations from a computer game in the same way. Thus, this study expected high Neuroticism individuals to prefer fewer wrong answers and generate higher PPC scores.

Guessing. Percent correct on a task is a not just a factor of skill, but also how often a guess is correct. In other words, a person can either know the answer or guess right to get the task correct. Basic probability shows that P(A or B) = P(A) + P(B) – P(A and B). If A is “answer known” and B is “guesses correct”, then this formula becomes P(task correct) = P(answer known) + P(guesses correct) – P(answer known and guesses correct). For tasks with a fixed number of answers, the probability of a correct guess is constant and equal to $1/k$ where k is the number of choices available. Thus, P(task correct) = P(answer known) + (1 – P(answer known))*$1/k$.

Since the present study attempted to establish a universal value for preferred percent correct, it was important to determine which value was portable: percent of answers correct or percent of answers known? Is getting an answer correct by guessing equally as satisfying as getting an answer right because of skill? If so, it would likely be the case that raw PPC was portable; if not, guess-corrected PPC would likely be constant between tasks.

Methods of finding PPC. It was not possible to ask participants directly about their preferences for percent correct, as they might not know what they would enjoy. The best way to get a value for preferred percent correct was to set up tasks for the participants and to have them decide which difficulty level was most fun, then try to turn this preferred difficulty choice into a preferred percentage correct.

In order to turn a preferred difficulty choice into a preferred percentage correct, it was first necessary to find a function relating these two quantities. The present study employed two estimations for this function. They were derived from different data and different methods, but were each intended to measure the same value. It was important to include both of them because either one could have been wildly off the mark in making an estimation. If the two agreed, it would be unlikely that they were both far off the mark in the exact same way.

Individual PPC (IPPC). Each participant would leave a record of successes and failures at the tasks for various difficulty values. This research fitted a modified logistic curve to each participant's data set, and then used this curve to calculate the predicted probability that the participant would succeed at a trial at the participant's chosen preferred difficulty level.

Difficulty levels varied from trial to trial, and it was possible a participant would encounter a difficulty level only one time for a given task. As such, a large percentage of difficulty values might have only had one success or one failure, leaving *p*-values of 0 or 1. Using logistic regressions in such situations is problematic. Instead of using logistic regressions, this research used search methods to find the curve that would have the maximum likelihood of generating the given result set.

An additional problem in establishing logistic curves was that the minimum expected probability for the visual and verbal tasks was not zero. The visual task had three choices, so if a participant guessed randomly, the expected fraction of items correct would be 1/3. Similarly, the verbal task for a list of m words had m+1 answer choices, so a random guess would have a 1/(m+1) likelihood of being correct. Generally, the probability of getting a correct answer is P(x) = L(x) + (1-L(x))*G(x), where P(x) is probability of a correct choice, G(x) is the probability of a correct guess, and L(x) is the *guess-free success rate*. L(x) functions were constructed by searching to find the logistic curve which had the highest likelihood of generating the participant's data set. Plugging in the participant's post-test preferred difficulty choice into P(x) yielded the *raw* IPPC value for that participant and task. Plugging this difficulty choice into L(x) yielded the *guess-corrected* IPPC value.

Aggregate PPC (APPC). A second estimate of preferred percent correct was made using linear regression. A scatter-plot was made using the percent correct at test difficulty as the y-value and the change in difficulty between the pre-test and post-test choices as the x-value. Each participant generated a single point for each task. For a set of participants, the regression line through the scatter-plot would intersect the y-axis where the expected change in difficulty between the pre- and post-tests would be zero. This is the percentage correct which corresponds to an aggregate desire for no change in difficulty level, and the current research interpreted this intercept as the aggregate preferred percentage correct, or APPC.

Research Aims and Hypotheses

The main purpose of this study was to quantify the preferred percentage correct (PPC) for each task. PPC could not have been found directly, so it was inferred from participant choices. There were two measures of PPC. The first was a measure of individual preferred percent correct (IPPC) derived from fitting curves to individuals' complete data sets and plugging in the last difficulty choice made. The second was an aggregate preferred percent correct (APPC) score. Taking a single point for each individual, creating a regression line, and finding the y-intercept of this line generated the APPC value. These two measures were computed using different methods but were intended to measure the same value.

Hypothesis 1: Participants' selections for ideal difficulty level will have a normal distribution.

Hypothesis 2: The APPC and IPPC methods of calculating PPC will yield the same values.

Hypothesis 3: PPC values will not be dependent upon task type.

Research question 3A: Will guess-corrected or raw PPC values create a better fit for hypothesis 3?

Hypothesis 4: PPC values will not be dependent upon game usage type. That is, those participants who spend more than 5 hours per week on average playing games will not have significantly different PPC values than those who do not.

Hypothesis 5: Higher Neuroticism scores will correlate with higher PPC values.

Chapter II: Method

Participants and Design

Participants were advertised for online at Facebook. The flyer emphasized that the study involved playing three video-game-like tasks that would take approximately 15 minutes total, and that individuals with all levels of gaming experience are needed for the study (please see Appendix A). Would-be participants were directed to the website www.ndnuvideogamestudy.com in order to be part of the experiment.

Between 8/25/2014 and 11/23/2014, 102 participants took part in the study. The group was young (age range 18 - 42, $M = 24.1$), and consisted mostly of people who spent large amounts of time gaming ($n = 62$ in the "high" game usage segment, $n = 40$ in the "low" game usage segment, as described below). This split is vastly different from the general public, where approximately 14% of people qualify for the "high" game usage segment ("New Report", 2013). The average age of video game users is 31 (Entertainment Software Association, 2014), which is higher than the average age of the sample.

37 of the participants were asked about gender, income and race. The gender of these 37 participants was roughly evenly divided between male and female, with 17 females, 20 males, and 0 transgendered individuals. Seven of the female participants were in the high game usage segment, and ten were in the low game usage segment. 15 of the male participants were in the high game usage segment, and five were in the low game usage segment.

The distribution of race of the 37 participants who were asked skewed toward White Americans. Of the participants asked about race, 28 answered as "Caucasian/White", two answered "Asian American", one answered "Hispanic", one answered "Asian", and five answered with two or more races. Of the five who reported to be two or more races, four included “Caucasian/White” as one of their races. Overall, 78% of the respondents answered as only White, and 89% were only White or White and another race. According to the 2010 Census (Hixson, Hepler and Kim, 2011, p.3), for the USA as a whole, these numbers were 64% and 75%, respectively.

The distribution of income of the respondents skewed toward low-income. Twenty-five of the 37 respondents answered as having annual incomes of less than $25,000, and seven answered as having annual incomes of between $25,000 and $50,000. This skew in income distribution was to be expected due to the distribution of ages of the participants; according to the U.S. Census (2011), the median annual income of individuals from age 15-24 was less than $14,000. 22 of respondents were in the 18-24-age range, and 19 of those reported annual incomes of less than $25,000.

Participants were divided into two groups based on the amount of experience they had with computer gaming. These levels were called “low” and “high” game usage. In order to define these groups, the present study relied on the definitions of The NPD Group (https://www.npd.com/wps/portal/npd/us/home/), a consumer market research firm that studies the gaming market. NPD research classified a “core” gamer as anyone who played video games for five or more hours per week (“New Report”, 2013). According to The NPD Group’s market research, this constitutes about 14% of the population (“New Report”, 2013). The current study adopted this definition of a “core” gamer to represent the “high” game usage segment. The sample consisted mostly of people who spent large amounts of time gaming (n = 62 in the “high” game usage segment, n = 40 in the “low” game usage segment). This split is different from the general public. The average age of video game users is 31 (Entertainment Software Association, 2014), which is higher than the average age of the sample.

This research defined the “low” game usage group to include those individuals who were unlikely to play games in any given week, or who were unlikely to play more than five hours of games per week. In order to sort participants into these categories, we asked participants to answer two Likert-scale questions about game usage habits as described in the “procedure” section below.

The design of this study was a 2x3 mixed-subjects quasi-experiment. The quasi-independent variables were “game usage” and “task type”. Levels for game usage were “low” and “high”, as described above. Task type levels were “visual”, “verbal” and “kinesthetic”, as described in the materials and apparatus section below. Every participant fell into exactly one of the two game usage levels, so game usage was a between-subjects variable. Every participant completed all task types, so task type was a within-subjects variable. For each task type, there were three dependent variables. The first was “pretest chosen difficulty”, which was a difficulty level choice made by the participant after he or she completed 15 practice runs of the task. The second was “post-test chosen difficulty”, which was a difficulty level choice made by the participant after the participant completed an additional ten test runs at the chosen pretest difficulty level. The third dependent variable was “score at chosen difficulty”, which was a percentage from 0 to 100%.

For a given participant and a given task, the participant was asked to complete least 25 trials of that task across three different phases. Phase one was the tutorial, where the participant completed five tasks of increasing difficulty with the help of detailed instructions. Phase two was practice, where the difficulty level varied according to the participant’s record; the algorithm increased the difficulty when a participant responded correctly and decreased the difficulty when the participant responded incorrectly. Before phase three started, the participant was asked to set the difficulty to the level he or she discovered to be the most enjoyable during the second phase. This value was inputted with a slider and confirmed with a dialogue box. The difficulty was fixed at this chosen level during phase three. Phase three was the test phase, where the participant completed ten trials at this self-selected difficulty level. Each trial from the second and third phases generated a data point (x = difficulty level of task, y = correct/incorrect response). When the participant completed the third phase, he or she was asked what difficulty level would be chosen if the participant did the task again. The participant inputted the answer to this question on a slider and confirmed the choice with a dialogue box.

Materials and Apparatus

Participants accessed the three tasks on a computer. These tasks were the visual, verbal, and kinesthetic tasks described below. Each task was introduced by a short tutorial describing how to input answers. Scoring and score reporting were automated.

Each of these tasks had to fulfill the following characteristics in order to work for the design of the current experiment. One, it must have been clear when there had been a success or failure on the task. Two, there must have been some easily adjustable component of the task which scaled either directly or inversely with difficulty. Three, the task must have been short enough so that 25 or more repetitions could have been completed in five minutes or less, ideally. Four, the task must have been novel but quickly learnable, meaning that it should have been composed of activities which were familiar to the participant but which were combined in a manner with which the participant was not familiar. Five, answers must have been able to be inputted with a short series of keystrokes or mouse movements.

For the visual task, the participant was presented with two triangles. The participant was asked to visualize the circles that circumscribed each triangle. The participant then decided whether these two circles were the same size by means of a series of buttons marked "left is bigger," "same size," and "right is bigger" (Appendix B). The participant had no time limit for this task. During all phases, after the participant selected an answer, the display revealed the circles and their relative sizes (Appendix C). One third of the time the circumscribed circles were the same size. Two thirds of the time the circles were of different sizes, which were equally split between having the larger circle on the left and having it on the right. On items where the circles were not the same size, the closer these two circles were to the same size, the harder it was to discriminate between equal-sized and unequal-sized pairs. When the difference in size between the circumscribed triangles fell below the participant's discrimination threshold, the participant should have been no better than random chance at deciding between similar and different sizes. The slider associated with this task adjusted the difference in radius of pairs of unequal size. The range of values on the slider was "0" to "50," measured in pixels of radius difference.

For the verbal task, the participant was presented with a target word and then a list. The participant had four seconds to determine how many of the words in the list rhymed with the target word. The list contained words which were chosen from one of four categories: rhyming lookalikes ("our" and "sour"); non-rhyming lookalikes ("our" and "pour"); rhyming non-lookalikes ("our" and "flower"); and non-rhyming non-lookalikes ("our" and "grower") (Appendix D for the entire list of words; Appendix E for an example). The number of rhyming words was a uniform random variable that varied between zero and the number of words on the list. Rhyming words were chosen with equal probability from lookalikes and non-lookalikes; similarly, non-rhyming words were chosen with equal probability from lookalikes and non-lookalikes. The selected words were sorted randomly and presented. When the participant was prompted with the word list, the clock started ticking down from four. The participant was instructed to click on a button indicating the correct number of rhyming words in the list. If the timer hit zero before the participant typed an answer, the target word and the word list disappeared, and were replaced by a prompt to choose an answer. When the participant entered an answer, the correct set of answers was revealed (Appendix F). The more words in the list, the harder it was to complete the task in four seconds. The slider in this task adjusted the number of words in the list. The slider's range extended from one word to ten words.

For the kinesthetic task, participants were asked to use the arrow keys to guide a ball through a randomly generated maze to a blue end square (Appendix G). Success meant getting the ball to the square without hitting a wall; if the participant collided with a wall, it was considered a failure. When a wall was hit, the ball disappeared and the wall blinked red where the ball hit the wall. After a failure, the ball reappeared at the beginning of the same maze so the participant could try again. After a success, a new maze was generated. The slider adjusted the radius of the ball from 1 pixel to the half the width of a corridor, roughly 60 pixels.

Procedure

Before this study began, a proposal was submitted to the Notre Dame de Namur University Institutional Review Board (IRB). Drafts of this proposal were given to the second reader, Dr. Helen Marlo, and the original thesis advisor, Dr. Nusha Askari. The thesis advisor, the second reader, and the IRB each signed off on the study before any data were collected. Sixty-five participants took part in the study under this arrangement. Dr. Askari left NDNU in summer 2014. Dr. Michelle Haley and Dr. Matthew Harris became the new thesis advisors for this experiment and asked for a redesign so that extra demographic information and a Neuroticism inventory could be collected from participants. A second IRB proposal with the redesigned study was submitted and approved in October 2014. The new thesis advisors, the second reader, and the IRB all signed off on the new design before any new data was collected. Thirty-seven participants took part in the study under this new arrangement.

Ads were placed on Facebook, Reddit and Craigslist for volunteers. The ads emphasized that individuals who were 18 years or older with all levels of gaming experience were needed, and that the study involved playing games on a computer for approximately 10 to 15 minutes. Please see Appendix A for a copy of the ad.

In order to be enrolled in the study, prospective participants visited a website (www.ndnuvideogamestudy.com) containing a program which automated the study and sent data back to a server. Both the program's code and the server code will be made publicly available at the end of the study. The program removed identifying information from all data sent over the Internet. The program and its associated database were hosted on a public web server. No information that could be used to identify participants was or will be stored on the server.

The participant first encountered an age-verification prompt, and then an informed consent form. In order to proceed, the participant had to give a birthday consistent with an age of 18 or older, and then had to acknowledge informed consent. Failure to do either caused the program to terminate. Please see Appendix H for a copy of the informed consent form.

Once age was verified and informed consent was given, the participant answered a two-question survey about gaming habits. The first question was "in a given week, how likely are you to spend any time playing video games?" The participant was presented with a Likert scale where one = "very unlikely", two = "unlikely", three = "equally likely as not", four = "likely", and five = "very likely". If the participant answered with a one or a two, that participant was sorted into the low game usage category as described above. If the participant responded with a three, four or five, that participant was asked a follow-up question: “in a given week, how likely are you to spend AT LEAST FIVE HOURS playing video games?” Another Likert scale with the same ratings appeared, and the participant indicated a response. An answer of one or two put the participant in the low game usage category, and a response of three through five put the participant in the high game usage category.

After the gaming habits inventory, participants were asked about their races, genders and income levels. Please see Appendix I for a list of demographics questions and possible answers. A Neuroticism personality inventory was then given. Please see Appendix J for the Neuroticism inventory questions and how it was scored.

Once gaming habits, demographics and Neuroticism were recorded, the program guided the participant through the three tasks described above in the materials section. Task order was completely counterbalanced and randomized. That is, there was an equal chance of the participant being asked to complete the tasks in any of the six possible orders (*ABC*, *ACB*, *BAC*, *BCA*, *CAB*, and *CBA*, where *A* = verbal, *B* = visual, and *C* = kinesthetic).

Participants completed 30 items of each task. At the beginning of each task, there was a short (less than one minute) tutorial on the nature of the task and how to input answers. The first five items comprised the tutorial. Between the tutorial and the test, the participant practiced the task for 15 trials. An algorithm set the difficulty level of these trials based on how well the participant had done in previous trials. After the end of the practice session, the participant was prompted to pick the level of difficulty level that was the most enjoyable. Following this selection was the test session, which was composed of ten items at this chosen difficulty level. At the end of each task, the participant was asked, “if you were to do this task again, what difficulty level would you want it to be?" The answer to this question was recorded and used for the analysis described below.

After all tasks were completed, a debriefing form was shown with the suggestion that the participant save a copy on a local drive (please see Appendix K for debriefing form). After debriefing, the responses and any additional comments were coded into an HTTP-post request to a hidden page on the www.ndnuvideogamestudy.com website. The page that caught the HTTP-post request was written in PHP. Data was stored in a MySQL database. The program itself was written in Java, digitally signed using a code signing certificate from Comodo, and deployed from the website using Java Web Start. Code for the program, the Java Web Start launcher, and the HTTP-post catcher will be available online after the completion of the data gathering phase and closure of the database.

Chapter III: Analysis

All analysis was conducted using STATA.

Normal Distribution of Participants' Chosen Difficulty Levels

In order to test the hypothesis that difficulty choices were normally distributed, Skewness-Kurtosis tests (STATA's sktest command, based on Jarque and Bera (1987) with corrections by D'Agostino, Belanger, and D'Agostino (1990) and Royston (1992) were conducted on each of the six samples of difficulty choice. Results are summarized in the table below.

Table 1

Skewness-Kurtosis tests on samples of difficulty choice

Sample	χ^2	P
Visual pre-test	13.86	.001*
Visual post-test	19.46	< .001*
Verbal pre-test	.45	.798
Verbal post-test	.11	.945
Kinesthetic pre-test	.91	.634
Kinesthetic post-test	5.73	.057

Note. * = $p < .05$. $N = 102$ for all tests.

Difficulty choices for the visual task were significantly non-normal. This analysis did not support the hypothesis that participants' chosen difficulty levels were normally distributed. Given the lack of normality of difficulty choices for the visual task, subsequent statistical investigations were taken to explore the underlying causes of this deviation from the hypothesized distribution. Two reasons appeared to account for the discrepancy: round number preference and mismatch between choice median and choice domain median.

Round number preference. If there were no preference for round numbers, difficulty values that were divisible by 10 should have occurred roughly 10% of the time for the visual and kinesthetic pre- and post- test difficulty choices. In order to test whether there was a preference for round numbers, binomial tests-of-fit were run on the difficulty choices in various samples against an expected 10% occurrence rate.

Table 2

Tests of round number preference in difficulty choices

Difficulty choice sample	Percent of choices ending in 0	p
Visual pre-test	43.1%	< .001*
Visual post-test	36.3%	<. 001*
Kinesthetic pre-test	34.3%	< .001*
Kinesthetic post-test	33.3%	< .001*

Note: p-values are results of binomial tests of significance vs. 10% expected. * = $p < .05$. $N = 102$ for all tests. Probabilities are one-tailed.

This showed that participants had a significant bias toward choosing round numbers for difficulty values. Given that moderate round number difficulty values could have been chosen for only the visual and kinesthetic tasks, round number preference could explain a portion of the deviation from normal present in distributions of choices for these two tasks.

Median mismatch. A second possible explanation for the non-normal distribution in difficulty choices for the visual task was the mismatch between the median of choices made and the median of the set of possible choice values. For instance, in the visual task, difficulty choices for the pre-test and post-test had median values of 18 and 13 respectively, but the range of possible choices contained all integers from 0 to 50, with a median of 25. Data summarizing this relationship is presented in the table below.

Table 3

Sample medians, choice range medians, and normality of samples

Difficulty choice sample	Median choice	Median of possible choices	Difference of medians	Is sample normal?
Visual pre-test	18	25	7	No
Visual post-test	13	25	12	No
Kinesthetic pre-test	30	29	1	Yes
Kinesthetic post-test	30	29	1	Yes
Verbal pre-test	5	5	0	Yes
Verbal post-test	5	5	0	Yes

Note: normality based on Skewness-Kurtosis tests with significance at $p < .05$.

As shown in this table, large differences between median choice and median of choice domain were associated with non-normal choice distributions.

Estimates of PPC Using Two Different Techniques

This research used two methods for estimating PPC and tested the hypothesis that these two methods yielded similar answers. Once values and standard errors were generated, two-tailed t-tests were performed to determine if the values were significantly different.

Table 4

APPC and IPPC values comparison by task

Task	IPPC mean	APPC	T	p
Visual	.7169 (.0186)	.7612 (.0176)	1.73	.085
Verbal	.6303 (.0224)	.6355 (.0177)	0.18	.856
Kinesthetic	.5668 (.0286)	.5849 (.0236)	0.49	.623

Note: standard error reported in parentheses. $N = 102$ for all tests. Probabilities are two-tailed. Data are uncorrected for guessing.

The p-values for each of these tests were non-significant. This evidence supported the hypothesis that IPPC and APPC values were not different.

PPC Values and Tasks

The main goal of this research was to find what would be the preferred percent correct for participants All of the theoretical justifications for this experiment were agnostic as to task type, so it would be reasonable to expect no difference between PPC values for task types. A one-way, repeated measures ANOVA test was run to test the hypothesis that IPPC values were different for different tasks. For this test, $F(2, 101) = 1.43$, $p = .017$. The F-statistic for this ANOVA test was significant at the $p < .05$ level. Importantly, the F-statistic for the Task variable was large and significant, with $F(2) = 11.24$, $p < .001$. This result did not support the hypothesis that IPPC values are independent of task type.

A question in this research was whether correcting for guessing would yield more consistent results among tasks. In order to test this hypothesis, a one-way, repeated measures ANOVA test was run on the guess-corrected IPPC values. For this test, $F(2, 101) = 1.22$, $p = .121$. More important was the small contribution of the Task variable, with $F(2) = .59$, $p = .553$. This analysis showed no significant difference among IPPC values for the three tasks once those values were corrected for guessing. This supported the hypothesis that PPC was independent of task type, but only on the condition that PPC was corrected for guessing.

PPC and Game Usage Type

Two-tailed t-tests were run to test the hypothesis that PPC and Game Usage Type were not related. The results are detailed in the table below.

Table 5

IPPC mean values for game usage types

Task	Low game usage IPPC mean	High game usage IPPC mean	*t*	*p*
Visual	.7466 (.0270)	.7114 (.0251)	0.93	.357
Verbal	.6293 (.0329)	.6265 (.0288)	0.06	.950
Kinesthetic	.5628 (.0432)	.5767 (.0367)	0.24	.809

Note: standard error reported in parentheses. $N = 40$ for low usage group, $N = 62$ for high usage group for all tests. Data are uncorrected for guessing. Probabilities are two-tailed.

No significant differences were found between IPPC means for the high and low game usage types. This supported the hypothesis that game usage type is not related to PPC values.

Neuroticism and PPC

In order to test the hypothesis that PPC and Neuroticism were positively correlated, a linear regression was conducted between Neuroticism scores and IPPC values for all Neuroticism scale respondents for each task type. One-tailed t-tests were run on the coefficients of regression. The results of these tests are detailed in the table below.

Table 6

Neuroticism vs. IPPC regression coefficient t-tests

Task	Neuroticism regression coefficient	*t*	*p*
Visual	-.0503 (.0310)	-1.62	.943
Verbal	.0377 (.0367)	1.03	.156
Kinesthetic	.1039 (.0464)	2.24	.016*

Note: * = $p < .05$. Standard error reported in parentheses. $N = 37$ for all tests. Data are uncorrected for guessing. Probabilities are one-tailed.

A significant result was found for the relationship between kinesthetic task IPPC and the result on the Neuroticism inventory, with an increased IPPC of approximately 10% for each point increase in Neuroticism. The other two tasks showed no significant correlation with the

Neuroticism inventory score. This result provided support for the hypothesis that there is a significant correlation between Neuroticism and IPPC, but only in the case of the kinesthetic task.

Chapter IV: Discussion

The current study sought to find a universal preferred percentage correct (PPC) value for novel, computer based tasks. It also sought to study whether certain characteristics of participants, such as Neuroticism and amount of game usage, and certain characteristics of tasks, such as task type and correct guess rate, were associated with changes in PPC values.

Major Findings

Hypotheses.

Hypothesis 1 stated, "Participants' selections for ideal difficulty level will have a normal distribution." This hypothesis was not supported. Specifically, for the visual task, there was significant evidence of non-normality in the distribution of both pre-test and post-test choices.

Hypothesis 2 stated, "The APPC and IPPC methods of calculating PPC will yield the same values." This hypothesis was supported. Task by task comparisons of APPC and IPPC values did not show differences at the $p = .05$ level. It was reasonable to conclude from this that IPPC and APPC produce comparable measurements, and that any conclusions about IPPC means could be generalized to APPC values.

Hypothesis 3/research question 3A stated, "PPC values will not be dependent upon task type. Will guess-corrected or raw PPC values create a better fit for this hypothesis?" For raw IPPC values, this hypothesis was not supported. However, for guess-corrected IPPC values, this hypothesis was supported. This provides a definitive answer for the research question: guess-corrected PPC values (and not raw PPC values) are what remain consistent over different tasks.

Hypothesis 4 stated, "PPC values will not be dependent upon game usage type." This hypothesis was supported. Participants who spent at least five hours per week playing games showed no preference for different PPC values than did those who spent fewer than five hours per week on games.

Hypothesis 5 stated, "Higher Neuroticism scores will correlate with higher PPC values." This hypothesis was supported in the case of the kinesthetic task, but not for the other two tasks.

Non-normal distribution of difficulty choices. In two of the six samples of difficulty choices, data did not fit normal distributions. Two reasons appeared to be responsible for this: round number preference and mismatch of selection domain to user choices.

For the verbal task, difficulty choices were constrained between 0 and 10 words per list. However, for the visual and kinesthetic tasks, there were much larger domains of choices that participants could make; the set of possible choices for the visual and kinesthetic tasks spanned 50 and 58 values, respectively. As a result, in these two tasks, there were many more "round numbers" (numbers divisible by 10) which could be chosen. These numbers were indeed chosen much more often than would be expected by chance alone. As Pope and Simonsohn (2011) showed, preference for round numbers affects behavioral choices in multiple domains. Round number preference is strong enough to cause significant losses to some stock market investors; Bhattacharya, Holden and Jacobsen (2010) estimated that stock buyers lose hundreds of millions of dollars per year due to this cognitive bias (Bhattacharya, Holden and Jacobsen, 2010). Round number preference would cause "clumping" at round number values, changing the distribution away from what would be expected for normal data.

In addition to round number preference, mismatch between median choices and median slider values appeared to have an effect on the normality of difficulty choice distributions. Table 3 showed that for all distributions where the median mismatch was greater than 1, the distribution was significantly non-normal. This would make sense in terms of producing skewed distributions. If there were less room on one side of the median to make choices, then those choices would aggregate over a smaller set of values and create an asymmetric distribution, leading to skew and non-normality.

Equivalence of APPC and IPPC measures. For each of the tasks, APPC values and IPPC means were not significantly different from each other. It was important to measure PPC twice to add evidence that measurements were valid. Given that these two different methodologies produced results that were not significantly different, there is more assurance that the measurements corresponded to what they were intended to measure. Thus, these measures showed convergent validity.

PPC independent of game usage. Though there turned out to be no relationship between game usage type and PPC, it might have been reasonable to expect a relationship. Gamers spend their free time and money playing games, thus they must find something inherently enjoyable about them. That inherently enjoyable aspect could have been the challenge and difficulty associated with games. However, if gamers were gamers because they could handle getting items wrong more often, then there should have been a relationship between PPC and game usage type.

Given the lack of a relationship between game usage and PPC, it seems that gamers prefer something else in video games besides a high level of challenge. Perhaps it is the immediate feedback, the clear goals, or the required concentration. However, these items are all on the list of requirements for flow experiences (Csikszentmihalyi, 1990). In fact, good game design incorporates all eight of Csikszentmihalyi's aspects of flow, and the literature on video games commonly cites flow as a reason why people play (Klasen, Weber, Kircher, Mathiak & Mathiak, 2011; Gentile et al., 2011). If games are designed in such a way that makes flow experiences easy to achieve, then maybe the better question is not why some people are gamers, but why some are not. Perhaps there is something missing from Csikszentmihalyi's aspects of flow. Perhaps social consequences prevent non-gamers from playing, the type of goals leading to a flow state are more important than the clearness of the goals, and certain types of feedback are preferred by non-gamers. If it were possible to predict whether someone would enjoy being a gamer, then such a technique could be expanded to include indications for other types of flow experiences.

Guess-free PPC independent of task type. The most portable measurement between the three tasks was the guess-corrected PPC value. The average of the guess-corrected PPC values for the three tasks was approximately 57%. This yields the following expected raw PPC value equation: PPC = .57 + .43*g, where g is the chance of a correct guess. In a multiple-choice situation, G = 1/*k* where *k* is the number of choices available to the participant. The following table summarizes the expected relationship between PPC and number of choices available per task. This empirical relationship provides a ripe target for future experiments.

Table 7

Expected raw PPC in multiple choice environments

Number of choices	Expected raw PPC
2	79%
3	71%
4	68%
5	66%
6	64%

Note: based on a guess-corrected PPC value of 57%.

In order to enter a flow state, an individual must experience a balance between the perceived difficulty of the task and his or her self-perceived skill level (Csikszentmihalyi, Abuhamdeh, Nakamura, Flow & Dweck, 2005). Thus, what matters is not the number of items correct on a task, but the number of items known by the participant. From this statement, it might have been expected that raw PPC values would not be consistent, but that guess-corrected PPC values would be. The data from the present study reinforce the claim that self-perceived skill level (and not skill level data from other sources, such as score reports) must match perceived difficulty of the task in order for flow to take place.

Neuroticism and PPC. There was some limited evidence of a relationship between Neuroticism and PPC, but only in the case of the kinesthetic task. The small sample size (n = 37) of the Neuroticism test respondents likely decreased the power of the measure and prevented a true test of the hypothesis from occurring. However, given the significant result for the kinesthetic task alone, and the nearly opposite result for the visual task ($t = -1.62$, $p = .943$, from table 4), one might speculate as to a connection. The kinesthetic task involved many more actions per task than the visual and verbal tasks did. The visual task required one action per task, which was comparing the estimated size of circles. The verbal task required a number of rhyming comparisons per task equal to the size of the wordlist, which averaged approximately five words. The kinesthetic task required the participant to hold and release the appropriate key multiple times per maze. For instance, in the kinesthetic task prompt pictured in Appendix G, the participant would need to make eleven turns without running into a wall, which would require over 20 correctly timed key presses and releases. In other words, a successful trial in the visual task requires completing one action at moderate likelihood of success, but a successful kinesthetic task trial requires completing many actions at high likelihood of success. Perhaps those individuals with high Neuroticism are more pessimistic when estimating the likelihood of failure over a sequence of actions, when each of these actions has a high probability of success. Prospect theory (Kahneman & Tverskey, 1979) presents a possible rationale for this distortion in perceived probability. Prospect theory predicts that when making probabilistic decisions, people will commonly underweight the likelihood of highly likely but not certain events. For instance, an event that has a likelihood of 95% will subjectively be weighted as having a lower than 95% probability, but a certain event will subjectively be weighted as having a likelihood of 100%. Perhaps there is a link between the pessimism inherent in high Neuroticism individuals and increased underweighting of highly likely probabilities.

Anticipated Confounds

Neuroticism. This research cited frustration response and avoidance of negative stimuli in its theoretical justification. Thus, it was important to check for individual response to these two factors to see how much of an effect they would have on participant choices. According to Jeroniumus, Riese, Sanderman & Ormel (2014, p.3), "neuroticism is defined as the propensity to experience distress and negative emotions", making neuroticism a good proxy for reactivity to negative stimuli. Additionally, frustration tolerance is inversely correlated with the neuroticism trait (Kundu & Basu, 1991). Thus, assessing for the neuroticism trait in participants was an appropriate way to control for both of these factors.

Anticipating neuroticism as a possible covariate and testing for its effects on PPC values proved to be fruitful. Even though the number of participants who completed the neuroticism inventory was relatively small (n = 37), a significant correlation was found between neuroticism score and IPPC value on the kinesthetic task. Additionally, the effect of neuroticism on the visual task was nearly opposite, though it failed to achieve statistical significance due to the one-tailed statistical test (please see table 6). The interactions between neuroticism, PPC and task type indicate that more study is warranted.

Learning modalities. Schmidhuber's formal theory of fun states that enjoyment, fun and interest happen when some redundancy in the data for our world is illuminated and we are able to recode our worldview in a smaller footprint (Schmidhuber, 2010). That is to say, enjoyment happens when learning happens. Since this research explicitly used the equivalence between learning and enjoyment as part of its theoretical justification, it was important to make sure that any learning related confounds were minimized when trying to estimate PPC values. Specifically, slower learning might have increased the rate of getting answers wrong on a task. This would likely have increased frustration levels and decreased the level of difficulty with which a user felt most comfortable.

A possible factor affecting learning was the participant's preferred learning modality, or the mode through which the individual most readily absorbs and retains information. The current study adopted the set of learning modalities in the Swassing Barbe Modality Index (SBMI) (Barbe & Swassing, 1979). These learning modalities are "visual", "auditory" and "kinesthetic". There is evidence that learning rate is independent of specific types of learning-modality dependent performance across settings. For instance, one study found that memory performance in a sensory domain is not correlated with preferred learning modality (Krätzig & Arbuthnott, 2006). This research provided one game from each of the three SBMI learning domains in order to examine the possible effects of learning style on learning rate. However, there were no observed differences between guess-corrected PPC values for the three task types, indicating no overall preference for any of the three modalities.

Limitations

Indirect measurement of PPC. The present study did not directly measure the percentage preferences of participants. Instead, it inferred values from participants' win/loss records on tasks and their difficulty value preferences. As such, there is a possibility that the measures employed were not valid. In order to mitigate this possibility, two measurements were made using different techniques and different data. These two measurements (IPPC means and APPC) converged for all tasks (Table 4), which increased the likelihood of validity of the PPC measurements. However, there still exists the possibility that both PPC measurements were invalid.

Lack of consequences. Unlike many computer-based tasks, there were no consequences for a win or a loss in the current study's tasks. Some games have in-game consequences for outcomes, such as giving or taking away points, equipment, or currency. Others provide social consequences for success or failure, such as changing a participant's rank on a leader-board. This lack of incentive was by design, as the current study intended to measure preferred percentage correct isolated from other factors. However, as a side effect of this design choice, it is possible that the lack of substantive consequences for outcomes resulted in lower attention being paid to the tasks, leading to less effort given.

Demographics. There are three main reasons why this research expected the demographics of the sample to be biased. First, participants accessed the study via a webpage, which would likely cause the sample to skew older and more affluent (Zickuhr & Smith, 2013), and to include fewer minorities than would be expected (Prieger & Hu, 2008). Second, the primary method of participant recruitment was through online social networks, specifically Reddit, which would cause the sample to skew younger and more male (Duggan & Smith, 2013). Third, the study involved tasks that were similar to computer games, which would likely bias the sample towards males (Ogletree & Drake, 2007). In order to account for these possible selection biases, participants were asked for their age, gender, household income, race/ethnicity, and video game usage.

As expected, the sample was significantly younger than the population as a whole, with a mean age of 24.1. The number of respondents who declared their race to be Caucasian/White was high; all but four of 37 respondents checked White or White and another race. The amount of video game usage was also quite high, with over 60% reporting video game usage of five hours per week or more. Surprisingly, the sample was low-income, with a majority of respondents (25/37) reporting annual incomes of less than $25,000/year. The actual sample was balanced with respect to gender, with 17 female and 20 male respondents. Due to the sample's deviations from the general public in the dimensions of age, race/ethnicity, gaming habits, and income, the results of this study may not generalize to the population as a whole.

Accents. Some of the comments that the participants left indicated that their pronunciations of certain words were different from what the present study expected. Three of the participants cited not growing up in the United States as a reason why they thought they did poorly on the verbal task. One participant reported not knowing some of the words in the task (such as "agate"), though steps had been taken to ensure all of the words were commonly used. Given that there was an unexpected divergence in pronunciation, it might have been better to only allow individuals who grew up speaking English in the United States to take part in the verbal task. If this research is repeated or extended, such a precaution should be taken.

Preferred learning modalities. Even though no relationship between task type and guess-corrected IPPC was found, there exists the possibility that preferred learning modality affected individual IPPC values. If the participants in the sample were equally balanced among preferred learning modalities, then the current study would not have been able to detect any effects due to these preferences. A proper control of this variable would involve assessing participants for preferred learning modalities and attempting to find links to IPPC values.

Future Research

New studies. The empirical relationship PPC = .57 + .43g needs to be tested to see if it holds. One way of doing this would be to assign the verbal task but restrict the number of possible answers, which would increase the chance of guessing correctly. Comparing the PPC values of each group based on the value of g could establish whether this relationship is true or just a statistical artifact.

The tasks in this study lacked any real consequences, so participants were not invested in the outcomes. This was by design – the present study only sought to measure PPC values independent of other factors. Adding consequences to outcomes would likely increase the desired percentage correct of the tasks. Adding a leader board to the tasks could create social consequences, or adding a cash bounty for getting tasks correct could add monetary consequences. Studying the effects of adding these consequences could yield information about our relative valuations of these factors. Extending the number of trials of a task or having it run over multiple sessions might increase the participant's investment in the outcome, and PPC values might change over time.

A complicated relationship between neuroticism, PPC and task type was hinted at, thought the test was underpowered due to the small number of neuroticism inventory respondents. More study is warranted to pin down this relationship. The kinesthetic task was the only one that showed a significant relationship between PPC and neuroticism score, and it differed from the other two tasks in the number of actions that had to be performed per item. For participants to estimate how often they would succeed at the task, they had to estimate the probability of succeeding at several very easy actions in a sequence. If participants with high neuroticism were more pessimistic about the outcomes of these easy actions, or were more pessimistic about how high probability events chain into a successful sequence, it would explain the results in Table Six. Kahneman and Tversky (1979) conducted experiments showing how participants' subjective assessments of probabilities differed from objective values under decisions involving risk. Repeating these experiments while requiring participants to assess probabilities of success of sequences of highly likely events might show differences between those participants who have high and low neuroticism scores.

Changes in current study if repeated. If this experiment were to be repeated, there are a number of corrections that should be made. First, the sample should be homogenized with respect to accent for the verbal task. Three participants commented that they took issue with the pronunciations because they learned English as a second language or grew up in another English-speaking country besides the United States. Individuals who do not have traditional American accents should be excluded from the sample.

Second, the range of values available for difficulty choice in the visual task should be reduced. The median values for the pre-test and post-test difficulty choices in the visual task were 18 and 13, respectively. However, the median of possible values was 25, and the difference between these medians ended up skewing the difficulty choices of participants into non-normal distributions. In order to remedy this, the maximum difficulty value should be restricted to 30 pixels.

Third, something should be done to reduce divergence from normality due to round number preference in difficulty choices. There are a number of possible ways of doing this. One of the ways would be to take the numbers off of the slider completely, though that would likely increase confusion as to what choice was being made. In order to compensate for the lack of information, visual cues could be added to indicate to the participant what choice he or she is about to make. A second way would be to restrict choices to a smaller set of values. This would reduce the granularity of choices that participants could make, but lack of granularity did not seem to have a strong effect on the data in the verbal task.

Fourth, a true control on preferred learning modality should be imposed. In addition to having tasks associated with each of the three SBMI domains, an inventory assessing for preferred learning modality should be given to participants. Tests should be run to see if there is an interaction between preferred learning modality and PPC values on the three tasks.

Conclusions

This study attempted to quantify preferred percentage correct (PPC) for tasks that were novel and consequence-free. The current study showed that guess-corrected PPC estimations were not affected by method of measurement, or participant's task type or amount of game usage. Average PPC for the three tasks was approximately 57%, which yielded the empirical relationship PPC = .57 + .43g, where g is the percent of time a randomly input answer is correct. Evidence was found for a positive correlation between PPC and Neuroticism, but only for the kinesthetic task.

These results might supply game designers with a new way of thinking about design of games with respect to difficulty level. Currently, balancing difficulty levels in games is an expensive, time-consuming process that involves releasing the game to small segments of the target audience (a process known as "beta testing") and waiting for feedback. Instead, game tasks could be designed with a target percentage correct in mind. Data could be gathered and compared to this percentage automatically, and this might allow designers to let the difficulty of the game fluctuate programmatically as a result of player performance, but still maintain the interest of players. Research in developing AI techniques for optimally adjusting difficulty in games is well underway (Spronck, Sprinkhuizen-Kuyper & Postma, (2004); Chanel, Rebetez, Bétrancourt & Pun (2011). The results of this study can provide such attempts a target PPC value at which to aim.

In addition, the results from this experiment are relevant to optimal educational design. Currently, school systems that use the letter system of grading assign importance to getting 90% or more correct answer results. However, this study indicated that participants preferred difficulty levels where they achieved less than 60% correct when corrected for guessing. Perhaps schools would interest students more by increasing the difficulty level of tests and assignments and expecting a lower percent correct. The high school graduation rate stands at about 70%, and only 32% of students leave high school ready for college (Greene & Forster (2003), p.3). Raising the difficulty level and lowering the expected percent correct might yield the dual benefit of keeping more students interested and producing more college-ready graduates.

References

Amos, S. (2013). “Watch: Gamification goes Mainstream”. Retrieved November 26, 2013 from http://www.huffingtonpost.com/shawn-amos/watch-gamification-goes-m_b_3950834.html

Barbe, W. B., & Swassing, R. H. (1979). *Swassing Barbe Modality Index: SBMI*. Zaner-Bloser Incorporated.

Bhattacharya, U., Holden, C. W., & Jacobsen, S. (2012). Penny wise, dollar foolish: Buy-sell imbalances on and around round numbers. *Management Science, 58*(2), 413-431.

Chanel, G., Rebetez, C., Bétrancourt, M., & Pun, T. (2011). Emotion assessment from physiological signals for adaptation of game difficulty. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 41*(6), 1052-1063. doi:10.1109/TSMCA.2011.2116000

Csikszentmihalyi, M. "Enjoyment and the Quality of Life." *Flow: The Psychology of Optimal Experience*. New York: Harper & Row, 1990. 43-70. Print.

Csikszentmihalyi, M., Abuhamdeh, S., Nakamura, J., Flow, A. E., & Dweck, C. S. (2005). *Handbook of competence and motivation.* The Guilford Press, New York, 598-608.

D’Agostino, R. B., Belanger, A., & D'Agostino Jr, R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician, 44*(4), 316-321. doi:10.1080/00031305.1990.10475751

Davis, M. S. (1971). That’s interesting. *Philosophy of the Social Sciences*, *1*(2), 309-344. doi: 10.1177/004839317100100211

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011, September). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th International*

Academic MindTrek Conference: Envisioning Future Media Environments (pp. 9-15). ACM. doi:10.1145/2181037.2181040

Dollard, J., Miller, N. E., Doob, L. W., Mowrer, O. H., & Sears, R. R. (1939). Frustration and aggression. doi:10.1037/10022-000

Duggan, M., & Smith, A. (2013). 6% of online adults are Reddit users. *Pew Internet & American Life Project*, *3*.

Entertainment Software Association. (2014). *Essential facts about the computer and video game industry*. Retrieved from http://www.theesa.com/facts/pdfs/ESA_EF_2014.pdf, 11/23/14.

Gentile, D. A., Choo, H., Liau, A., Sim, T., Li, D., Fung, D., & Khoo, A. (2011). Pathological video game use among youths: a two-year longitudinal study. *Pediatrics*, 127(2), e319-e329. doi: 10.1542/peds.2010-1353

Greene, J. P., & Forster, G. (2003). *Public high school graduation and college readiness rates in the United States* (Vol. 3). New York, NY: Center for Civic Innovation at the Manhattan Institute.

Hamari, J., Koivisto, J., & Sarsa, H. (2014, January). Does Gamification Work?--A Literature Review of Empirical Studies on Gamification. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on* (pp. 3025-3034). IEEE. doi:10.1109/HICSS.2014.377

Haworth, J., & Evans, S. (1995). Challenge, skill and positive subjective states in the daily life of a sample of YTS students. *Journal of Occupational and Organizational Psychology, 68*(2), 109-121. doi: 10.1111/j.2044-8325.1995.tb00576.x

Hixson, L., Hepler, B. and Kim, M. (2011). *The White Population: 2010*. 1st ed. [ebook] US Census Bureau, p.3. Available at: http://www.census.gov/prod/cen2010/briefs/c2010br-05.pdf [Accessed 30 Nov. 2014].

Hudson, N. J. (2011). Musical beauty and information compression: Complex to the ear but simple to the mind?. *BMC research notes*, *4*(1), 9. doi:10.1186/1756-0500-4-9

Ivory, J. D., & Kalyanaraman, S. (2007). The effects of technological advancement and violent content in video games on players' feelings of presence, involvement, physiological arousal, and aggression. *Journal of Communication*, *57*(3), 532-555. doi: 10.1111/j.1460-2466.2007.00356.x

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, 163-172.

Johnston, C. A., Moreno, J. P., Regas, K., Tyler, C., & Foreyt, J. P. (2012). The Application of the Yerkes–Dodson Law in a Childhood Weight Management Program: Examining Weight Dissatisfaction. *Journal of pediatric psychology*, jss040. doi: 10.1093/jpepsy/jss040

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263-291. Stable URL: http://www.jstor.org/stable/1914185

Klasen, M., Weber, R., Kircher, T. T., Mathiak, K. A., & Mathiak, K. (2011). Neural contributions to flow experience during video game playing. *Social cognitive and affective neuroscience*, nsr021. doi: 10.1093/scan/nsr021

Krätzig, G. P., & Arbuthnott, K. D. (2006). Perceptual learning style and learning proficiency: A test of the hypothesis. *Journal Of Educational Psychology*, *98*(1), 238-246. doi:10.1037/0022-0663.98.1.238

Kundu, R., & Basu, J. (1991). Frustration reactions as predictors of academic achievement and personality dimensions of school children. *Psychological Studies, 36*(3), 146-155.

Lommen, M. J., Engelhard, I. M., & van den Hout, M. A. (2010). Neuroticism and avoidance of ambiguous stimuli: Better safe than sorry?. *Personality & Individual Differences, 49*(8), 1001-1006. doi:10.1016/j.paid.2010.08.012

"New Report from the NPD Group Focuses on Core Gaming and Core Gamers". (2013) Retrieved November 26, 2013 from https://www.npd.com/wps/portal/npd/us/news/press-releases/new-report-from-the-npd-group-focuses-on-core-gaming-and-core-gamers/.

Ogletree, S. M., & Drake, R. (2007). College students' video game participation and perceptions: Gender differences and implications. *Sex Roles*, *56*(7-8), 537-542. doi:10.1007/s11199-007-9193-5

Ormel, J., Jeronimus, B. F., Kotov, R., Riese, H., Bos, E. H., Hankin, B., & ... Oldehinkel, A. J. (2013). Neuroticism and common mental disorders: Meaning and utility of a complex relationship. *Clinical Psychology Review, 33*(5), 686-697. doi:10.1016/j.cpr.2013.04.003

Pope, D., & Simonsohn, U. (2011). Round Numbers as Goals: Evidence From Baseball, SAT Takers, and the Lab. *Psychological Science (Sage Publications Inc.), 22*(1), 71-79. doi:10.1177/0956797610391098

Prieger, J. E., & Hu, W. M. (2008). The broadband digital divide and the nexus of race, competition, and quality. *Information economics and Policy*, *20*(2), 150-167. doi:10.1016/j.infoecopol.2008.01.001

Roberts, N. (2014). The Reliability of Eyewitness Testimony.

Royston, P. (1992). Comment on sg3. 4 and an Improved D'Agostino Test. *Stata Technical Bulletin, 1*(3).

Schmidhuber, J. (2007, January). Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In *Discovery Science* (*pp*. 26-38). Springer Berlin Heidelberg. doi:10.1007/978-3-540-75488-6_3

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *Autonomous Mental Development, IEEE Transactions on*, *2*(3), 230-247. doi:10.1109/TAMD.2010.2056368

Schmidhuber, J. (2010). Formal Theory of Fun and Creativity Explains Science, Art, Music, Humor. Retrieved December 1, 2014, from http://people.idsia.ch/~juergen/creativity.html

Silverman, R.S. (2013). "Latest Game Theory: Mixing Work and Play". Retrieved November 26, 2013 from http://online.wsj.com/news/articles/SB10001424052970204294504576615371783795248

Spronck, P., Sprinkhuizen-Kuyper, I., & Postma, E. (2004). Difficulty scaling of game AI. In *Proceedings of the 5th International Conference on Intelligent Games and Simulation (GAME-ON 2004)* (pp. 33-37). doi:10.1.1.386.7736

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The five factor model of personality. *Journal Of Experimental Psychology: Applied*, *17*(2), 71-96. doi:10.1037/a0024170

Takahasi, D. "Americans spend $23.5B each year on video games". (2010). Retrieved November 26, 2013 from http://venturebeat.com/2010/05/09/americans-spend-25-3b-each-year-on-video-games/

U.S. Census Bureau. (2012) Selected Characteristics of People 15 Years Old and Over by Total Money Income in 2011, Work Experience in 2011, Race, Hispanic Origin, and Sex. [table]. Retrieved November 29, 2014 from http://www.census.gov/hhes/www/cpstables/032012/perinc/pinc01_1.xls.

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

Wood, S., Sage, J. R., Shuman, T., & Anagnostaras, S. G. (2014). Psychostimulants and cognition: A continuum of behavioral and cognitive activation. *Pharmacological reviews*, *66*(1), 193-221. doi:10.1124/pr.112.007054

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit□formation. *Journal of comparative neurology and psychology*,*18*(5), 459-482. doi:10.1007/978-0-387-79948-3_1340

Zickuhr, K., & Smith, A. (2013). Home broadband 2013. *Washington, DC: Pew Internet & American Life Project: Pew Research Center*.

Appendix A

Recruitment Flyer

Seeking adult participants to play video games for a study
No gaming experience necessary

Research is focused on finding the best difficulty level of three different types of games

Participants may be of any gender or race and of any level of gaming experience

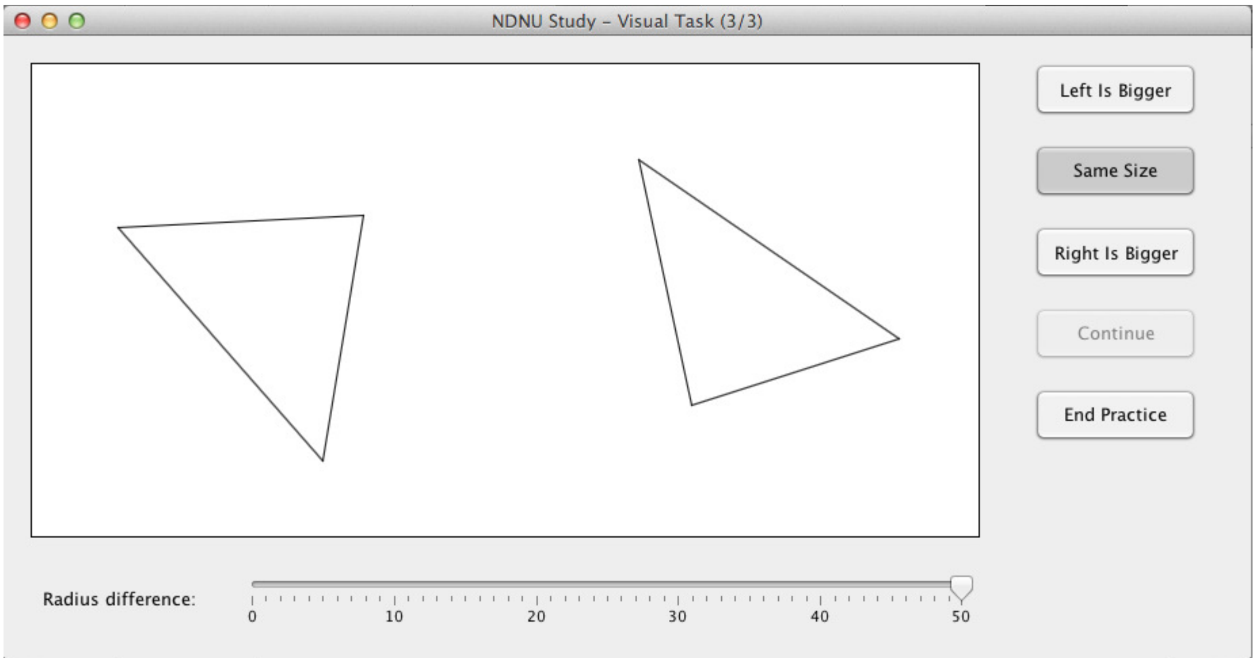
Participants must be 18 years of age or older to participate

Completion of the study involves playing games for 15 to 20 minutes and can be completed at any internet-capable computer with a mouse

To take part in this study, please visit
www.NDNUvideogamestudy.com

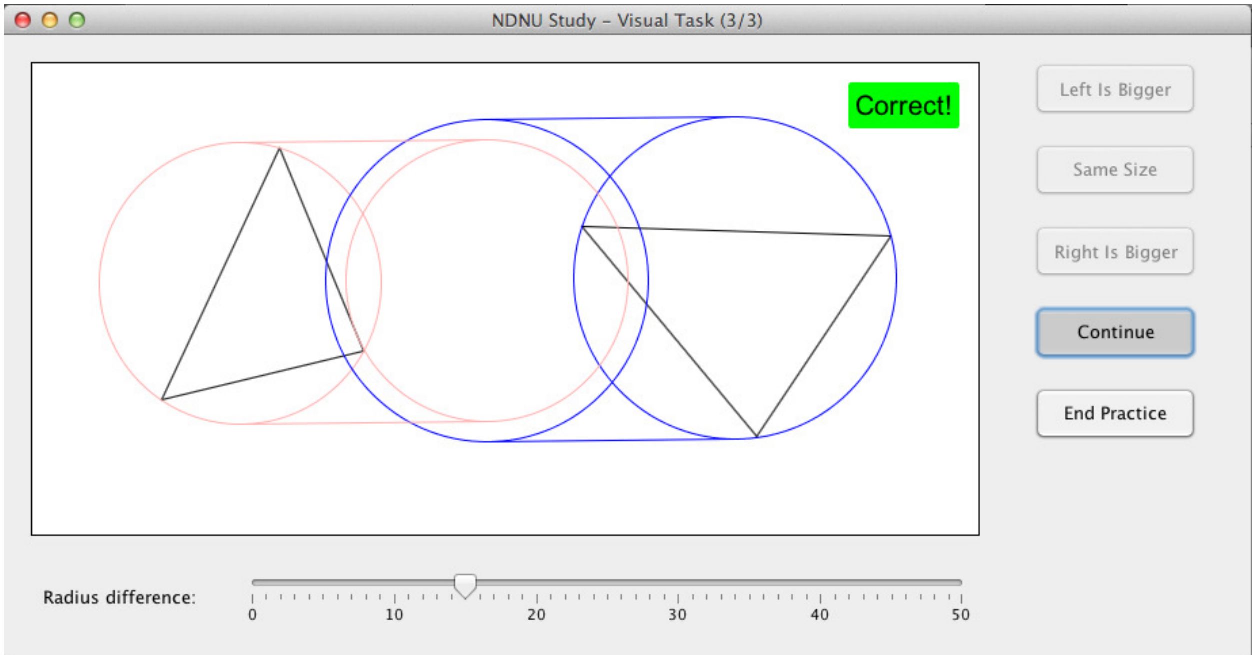
Appendix B

Visual Task Prompt



Appendix C

Visual Task, Answer Revealed



Appendix D

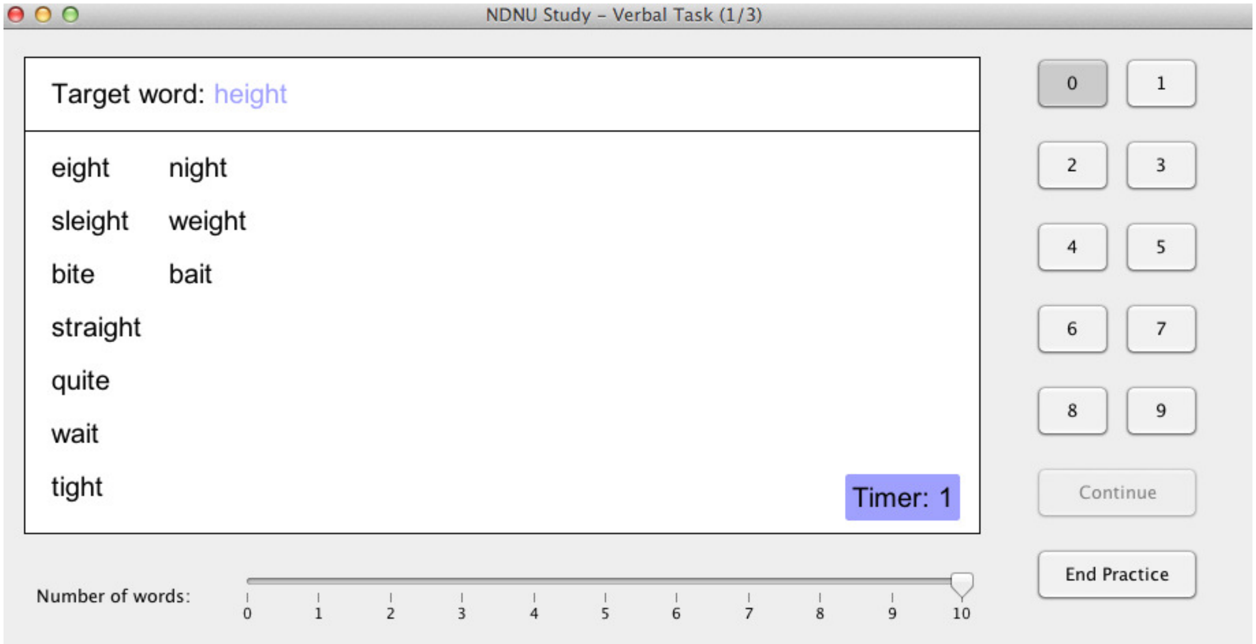
Verbal Task Word List

Target Word	Rhyming Lookalikes	Rhyming Non-lookalikes	Non-rhyming lookalikes	Non-rhyming non-lookalikes
ace	lace, face, mace, pace, race, grace	base, case, chase	terrace, menace, surface	ease, ice
band	gland, grand, stand, land	canned, banned, manned	wand	donned, grinned, gunned
bee	flee, dee, glee	flea, brie, sea	puree, melee, entree	adonis, tennis
bit	zit, pit, tit	Mitt	bait, gait, suit	late, mate, date
blew	new, hew, crew, drew	shoe, sioux, true, you	sew	thou, toe
brad	chad, glad, grad	Add	bead, dead, quad, goad	odd, rudd
by	my, dry, fly, cry	aye, buy, eye, high, sigh	any, joy, soy	neigh, weigh
cassette	corvette, dinette, baguette	beget, cadet, duet	ballet, chalet, fillet	latte, butte, matte
cough	trough	off, doff, scoff	bough, dough, enough, rough	blow, bro, tow,
cow	pow, now, how	ciao, thou, mao	low, blow, grow, mow	you
cue	blue, hue, sue, undue	you, few, view, zoo, do, to	pique, rogue, segue	thou, sew, go
good	hood, stood, wood	could, should, would	food, mood	crude, chewed, dude, lewd, rude
grave	brave, crave, shave	they've, waive	eave, have	sleeve, eve, cleave, weave
grey	prey, they, fey	stay, fillet, neigh	key, joey, abbey	basket, bonnet
hare	dare, rare, care	bear, pear	are	dear, fear, gear, hear
height	sleight	quite, bite, night, tight	weight, eight	bait, wait, straight
hurt	blurt, curt, spurt	dirt, flirt, pert, shirt, skirt	court	fort, port, quart

Target Word	Rhyming Lookalikes	Rhyming Non-lookalikes	Non-rhyming lookalikes	Non-rhyming non-lookalikes
kite	quite, smite, white	sight, wright	suite, elite, petite	bought, weight
love	glove, shove, above	of	clove, drove, move	aloof, proof, hoof
mice	price, slice, twice	vise, precise	juice, voice, bodice	bruise, wise
paid	laid, maid, raid	blade, glade, grade	said	bed, bread, head
plate	kate, date, fate, gate, hate, late, mate	eight, weight	agate, karate, pirate	bright, wight, tight
shed	bed, bled, fled, sled, sped	head, bread, dread, dead, tread, spread	aced, axed, seed, sued, iced	bead, mead, knead
sour	dour, hour, our, flour	how're, cower, flower	tour, pour, four	fewer, lower, sewer
stone	bone, lone, throne	blown, groan, own	none, done, gone, one	clown, gown, drown
that	chat, hat, shat	brat, flat, blat	what	beat, goat, swat
war	oar, boar, roar	for, your	bar, far, par, star	are
water	alma mater	daughter, slaughter	eater, cater, later, tater	laughter, freighter, traitor
womb	tomb	fume, plume, hume, flume	bomb, comb, aplomb	mom, palm, calm, prom
wreath	heath, sheath, beneath	keith, teeth	death, breath	meth, seth, beth

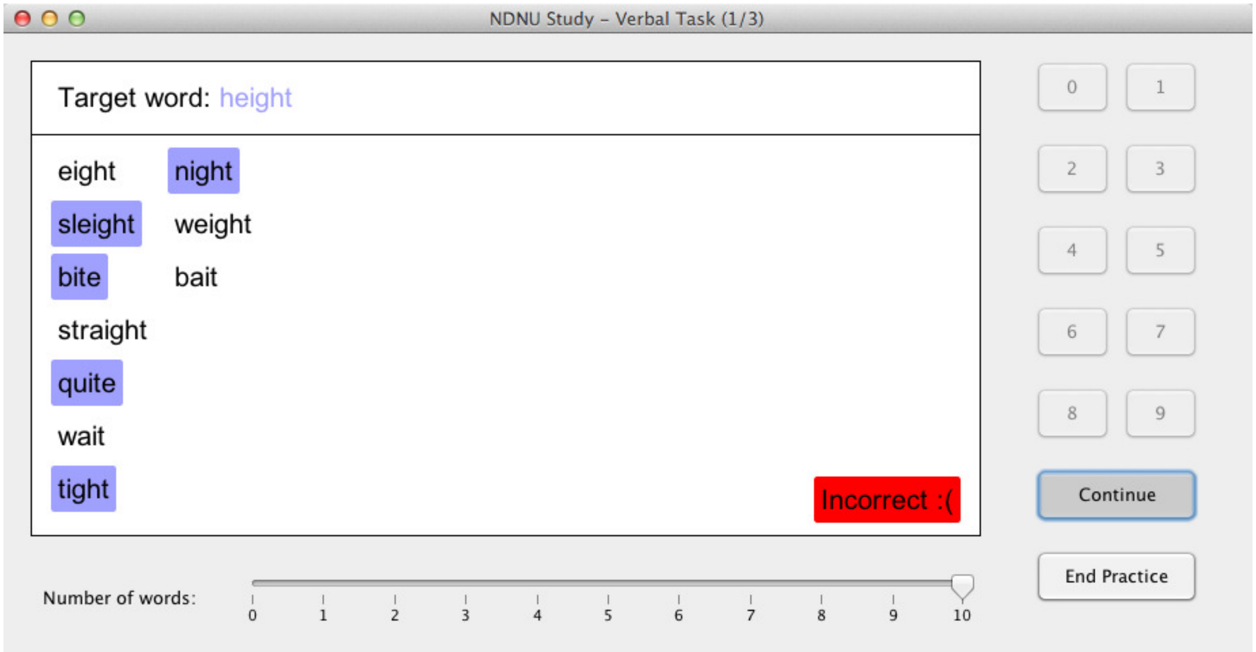
Appendix E

Verbal Task Prompt



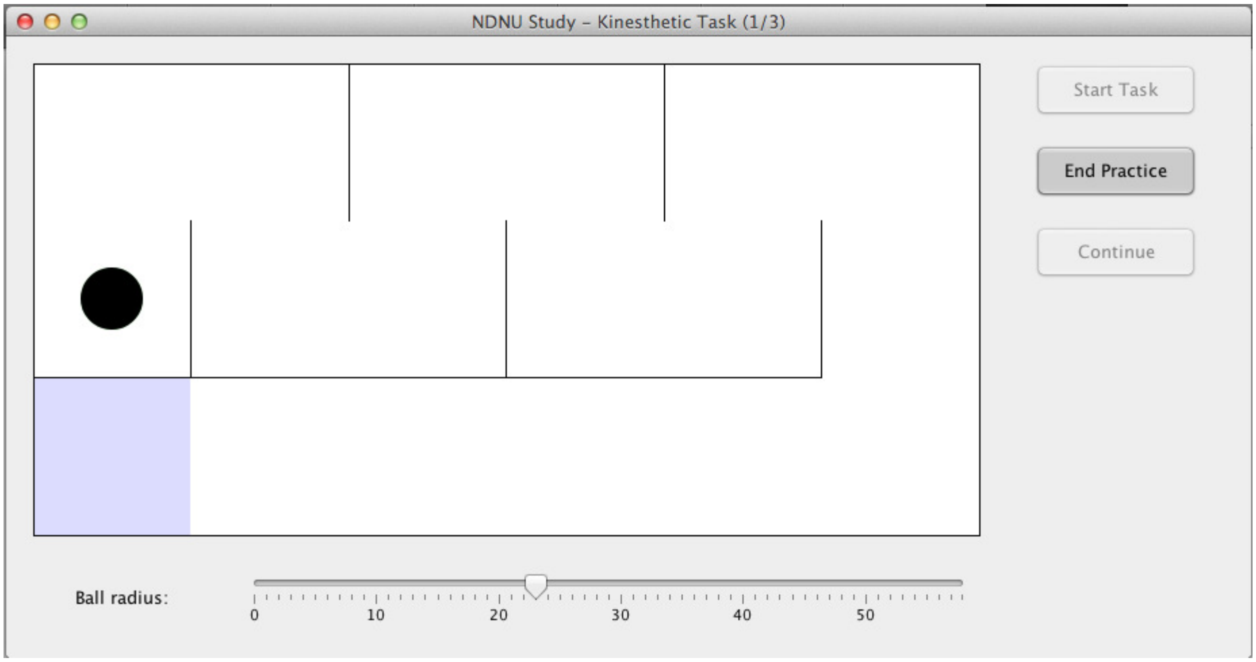
Appendix F

Verbal Task, Answers Revealed



Appendix G

Kinesthetic Task Prompt



Appendix H

Informed Consent Statement

Preferred Difficulty Levels in Novel Computer-Based Tasks

You are invited to participate in a research study, the purpose of which is to help understand the factors contributing to preferred difficulty levels in novel computer-based tasks.

INFORMATION
To help us in this task, we will ask you to participate in a study which lasts approximately fifteen (15) to twenty (20) minutes. You will be asked to complete three different types of tasks and to adjust the difficulty level of those tasks to be the most enjoyable for you. There will be three different types of tasks: visual, verbal, and kinesthetic.

For the visual task, you will be asked to mentally draw a circle which goes through all three tips of a triangle. You will be asked to compare the sizes of two of these circles.

For the verbal task, you will be asked to decide how many words in a list rhyme with a target word. You will have ten (10) seconds to complete this task. When you have made your decision, please type in a single number from zero (0) to nine (9) to indicate your choice. For instance, if the target word is “our”, and the list contains “power”, “pour”, “grower” and “sour”, you would answer by pressing ‘2’ on your keyboard (since “power” and “sour” rhyme with “our”, but “pour” and “grower” do not).

For the kinesthetic task, you will be asked to “erase” a line with a virtual “eraser” by moving your mouse in a single stroke. Please do not deviate from the line when doing this.

For each of these types of tasks, you will be able to adjust the difficulty level between trials for the first fifteen (15) trials. You will then be asked to set the difficulty level to a constant value for the next fifteen (15) trials. Please choose carefully when doing so, as once you make your choice this value cannot be altered for the following trials.

A target of 60 subjects has been set for this trial, but that number may vary based upon interest expressed by the population.

RISKS
There are no foreseeable risks associated with this study.

BENEFITS
The results of this study will help benefit scientific understanding of what people want to

experience when interacting with computer-based tasks. Such information will be beneficial to the design of both educational and recreational video games.

CONFIDENTIALITY

The data obtained in this study will be treated as confidential and will be stored securely on a server. You will not be asked to contribute any identifying information. Your responses will be encrypted before being sent to the server, and no personal information about you or your computer will be sent with your responses. The applet with which you are about to interact will not place any cookies on your computer and will not transmit data about the session to any third parties. Data from your responses will be saved for at least three (3) years after completion of the study. Data will be destroyed no more than ten (10) years after completion of the study.

COMPENSATION

You will receive no monetary compensation for participating in this study.

CONTACT

If you have any questions at any time about the study or its procedures, you may contact the principal investigator, Steven Riley, at 619-757-8799 or at NDNUVideoGameStudy@gmail.com. If you feel you have not be treated according to the descriptions in this form, or your rights as a participant have been violated during the course of this study, you may contact the Research Integrity Officer (RIO) at the office of the Provost at NDNU by calling (650) 508-3494.

PARTICIPATION

Your participation in this study is voluntary. You may decline to participate and have your data withdrawn at any time before, during or after completing the tasks in this applet. If you decide to participate, you may discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled.

CONSENT

If you are not eighteen (18) years of age or above, you are not able to give consent to complete this study. By clicking “I agree”, you state that you have read this consent form and that you understand the above information. Please print out this page for your records.

Appendix I

Demographics Questions and Possible Answers

What is your gender?

Female
Male
Transgendered: Female to Male
Transgendered: Male to Female

What is your race/ethnicity? (Check all that apply.)

African American/Black
Asian American
Caucasian/White
Latino/Latina
Pacific Islander
Other (Please specify)

What is your current income?

0 - 25,000
25,000- 50,000
50,000-75,000
75,000-100,000
100,000- and up

Appendix J

Big Five Personality Inventory

Retrieved from http://www.ocf.berkeley.edu/~johnlab/pdfs/BFI.doc

Bold items comprise the neuroticism scale

Non-bold items were not used in the study

How I am in general

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who *likes to spend time with others*? Please write a number next to each statement to indicate the extent to which **you agree or disagree with that statement.**

5 Agree strongly	**4** Agree a little	**3** Neither agree nor disagree	**2** Disagree a little	**1** Disagree Strongly

I am someone who…

1. _____ Is talkative

2. _____ Tends to find fault with others

3. _____ Does a thorough job

4. _____ Is depressed, blue

5. _____ Is original, comes up with new ideas

6. _____ Is reserved

7. _____ Is helpful and unselfish with others

8. _____ Can be somewhat careless

9. _____ **Is relaxed, handles stress well.**

10. _____ Is curious about many different things

11. _____ Is full of energy

12. _____ Starts quarrels with others

13. _____ Is a reliable worker

14. _____ Can be tense

15. _____ Is ingenious, a deep thinker

16. _____ Generates a lot of enthusiasm

17. _____ Has a forgiving nature

18. _____ Tends to be disorganized

19. _____ Worries a lot

20. _____ Has an active imagination

21. _____ Tends to be quiet

22. _____ Is generally trusting

23. _____ Tends to be lazy

24. _____ Is emotionally stable, not easily upset

25. _____ Is inventive

26. _____ Has an assertive personality

27. _____ Can be cold and aloof

28. _____ Perseveres until the task is finished

29. _____ Can be moody

30. _____ Values artistic, aesthetic experiences

31. _____ Is sometimes shy, inhibited

32. _____ Is considerate and kind to almost everyone

33. _____ Does things efficiently

34. _____ Remains calm in tense situations

35. _____ Prefers work that is routine

36. _____ Is outgoing, sociable

37. _____ Is sometimes rude to others

38. _____ Makes plans and follows through with them

39. _____ Gets nervous easily

40. _____ Likes to reflect, play with ideas

41. _____ Has few artistic interests

42. _____ Likes to cooperate with others

43. _____ Is easily distracted

44. _____ Is sophisticated in art, music, or literature

SCORING INSTRUCTIONS

To score the BFI, you'll first need to **reverse-score** all negatively-keyed items:

Extraversion: 6, 21, 31
Agreeableness: 2, 12, 27, 37
Conscientiousness: 8, 18, 23, 43
Neuroticism: 9, 24, 34
Openness: 35, 41

To recode these items, you should subtract your score for all reverse-scored items from 6. For example, if you gave yourself a 5, compute 6 minus 5 and your recoded score is 1. That is, a score of 1 becomes 5, 2 becomes 4, 3 remains 3, 4 becomes 2, and 5 becomes 1.

Next, you will create scale scores by ***averaging*** the following items for each B5 domain (where R indicates using the reverse-scored item).

Extraversion: 1, 6R 11, 16, 21R, 26, 31R, 36
Agreeableness: 2R, 7, 12R, 17, 22, 27R, 32, 37R, 42
Conscientiousness: 3, 8R, 13, 18R, 23R, 28, 33, 38, 43R
Neuroticism: 4, 9R, 14, 19, 24R, 29, 34R, 39
Openness: 5, 10, 15, 20, 25, 30, 35R, 40, 41R, 44

Appendix K

Debriefing Form

Thank you for your participation in this research on the preferred difficulty level of computer based tasks. Novel computer-based tasks were used for all participants in this study. The goal of the study was two-fold: to gather information on computer-user's preferred task difficulty, and to attempt to validate an algorithm that predicts difficulty level. It was hypothesized that participants would all prefer approximately the same percentage of correct trials for a task, and that each would set difficulty levels in such a way as to achieve this percentage correct. If you would like to learn more about design of computer based tasks, please see the references listed below.

Current research has found that participants prefer a moderate level of difficulty for tasks. Your participation was important in helping researchers find that level of difficulty and predict the optimal level for future tasks.

Final results will be available from the investigator, Steven Riley, by 12/15/2014. You may contact me at misterriley@gmail.com to receive an email copy of the final report. All results will be grouped together; therefore individual results are not available. Your participation, including your name and answers, will remain absolutely confidential, even if the report is published.
If you have any additional questions regarding this research, please contact me at misterriley@gmail.com.