Preferred Difficulty Levels in Novel Computer-Based Tasks

Steven Riley

Notre Dame de Namur University

Abstract

According to several theories, including Csikszentmihalyi’s theory of flow, the Yerkes-Dodson law, and Vygotsky’s zone of proximal development, we hypothesized that computer users would prefer moderate levels of difficulty when encountering new tasks, and that the preferred percentage correct for these tasks would be between 80% and 95%. In order to address this hypothesis, the preferred difficulty levels of a young (*n* = 10, *μage* = 29.3) group of Facebook users were analyzed using three computer-based tasks (one visual, one verbal, and one kinesthetic). Due to the small dataset there were problems in drawing robust conclusions. However, there was growing evidence of two important conclusions: one, learning takes place slowly in the assigned tasks; and two, participants would correct the difficulty level when the percentage correct on a task was too high or too low. These findings would imply that the tasks are well suited to study in this manner and that users indeed prefer a moderate level of difficulty. This report describes the pilot study; the full study will be published in Fall 2014.

Preferred Difficulty Levels in Novel Computer-Based Tasks

The search for interactive fun is a serious pastime of Americans. Blockbuster video games can gross upwards of $1 billion (“Top 10 highest”, 2012), and the average American gamer spends approximately $140 per year on games (Takahasi, 2010). There is obvious financial incentive for video game companies to “corner the market” on fun.

Corporate America has noticed this trend, and has begun to apply the lessons to keeping employees and customers motivated. Over the last few years, the technique of “gamification” has become mainstream (Amos, 2013). Gamification involves adding the trappings of games, such as points, badges, levels and achievements, to actions and relationships, which ordinarily lack such trappings. For instance, the website [www.duolingo.com](http://www.duolingo.com), a free resource for language learning, rewards a user with points for every completed exercise. If a user achieves a certain point total, that user’s level increases. Some companies use gamification to create incentives and increase the productivity of employees (Silverman, 2013).

However, this begs the question - what do we mean by “fun”? This paper will examine three different theories of fun and optimal experience, and it will show that all three predict a similar phenomenon: preferred difficulty level expressed by the user should be high enough to make a task interesting, but not so high as to make it impossible. In particular, all three will agree on the prediction that someone having fun at a task is getting at least some number of items wrong. This paper will then develop a study that will assess participants’ preferred levels of difficulty in terms of their percentage of correct responses on a task. The conclusions of this study will be statements about how often individuals desire to get items correct while doing a task, and statements about how the behavior of an individual changes when given a task that is too hard vs. one that is too easy.

The three models that this study will adopt include one based on affect regulation, one based on learning, and one based on a combination of these two factors. The model based on affect regulation will examine Csikszentmihalyi’s “flow” theory and what it predicts about behavior. The model based on learning will look at Schmidhuber’s formal theory of creativity, fun, and intrinsic motivation. The model that combines the two is based on Vygotsky’s zone of proximal development and the Yerkes-Dodson law. These theories will be fully described below.

As shown below, there is a close relationship between fun and learning. Because of this relationship, a possible confound on establishing a preferred difficulty value is the participant’s preferred learning modality, which is the mode through which the learner most readily absorbs information. This paper will adopt the set of learning modalities in the Swassing Barbe Modality Index (SBMI). These learning modalities are “visual”, “auditory” and “kinesthetic” (Barbe & Swassing, 1979). There is some disagreement over how preferred learning style affects learning rate (Dunn, 1993; Krätzig & Arbuthnott, 2006). Ideally, this study would test users for learning style and try to find a relationship based on these preferred orientations. However, it appears that several of the more popular learning style inventories disagree and are in fact measuring different things (Ferrell, 1983). In order to examine the possible effects of learning style on learning rate, this study will incorporate data from one visual, one verbal, and one kinesthetic task.

**Three Models of Fun**

**Range effects and flow**

In order to demonstrate a logical basis for the following experiment, this paper will establish an approximation of the function that relates task enjoyment and task score. This approximation is based on the concepts of range effects and flow. Flow can be thought of as the peak level of enjoyment that an individual experiences when the difficulty of a task exactly matches that individual’s preferred difficulty level for that task. The theory of flow states that the amount of enjoyment an individual experiences is a strictly decreasing function of the difference between the difficulty level of the task and the individual’s ideal difficulty level (Csikszentmihalyi, 1990; Csikszentmihalyi, 1997).

A function that represents flow according to the above description would relate difficulty to enjoyment, with difficulty on the *x*-axis and enjoyment on the *y*-axis. This function would have a peak at a specific *x*-value and decay to zero for large and small *x*-values. Such a monomodal (one-peaked) function might look like the one found in Figure 5.

Range effects include ceiling and floor effects, which effectively limit how well or poorly a person can score on a task, respectively. Range effects occur when the difficulty of a task is either too high, in which case the individual scores extremely low, or the difficulty is too low, in which the individual scores extremely high. In both of these cases, scores plateau outside of a sensitive range. When range effects are present, the expected score a person will achieve on a task is only sensitive to changes in difficulty over a relatively small domain of difficulty values. Any difficulty values outside this domain result in scores close to the minimum or the maximum.

A function that represents range effects would relate difficulty to score, with difficulty on the *x*-axis and score on the *y*-axis. This function would have a steep decline only on a small domain of *x*-values. For *x*-values less than this sensitive range, the function would be flat and have *y*-values close to the maximum score, corresponding to a ceiling effect. For *x*-values greater than this sensitive range, the function would be flat and have *y*-values close to the minimum score, corresponding to a floor effect. A class of smooth functions that fit this definition is referred to as “sigmoid”. Such a sigmoid function might look like the one found in Figure 7, which is from the class of sigmoid functions known as logistic functions.

For these two functions described above, both have difficulty as the *x*-axis. For a given difficulty, the first function associates a measure of enjoyment, and the second a score. For each difficulty it is therefore possible to relate the score to the level of enjoyment, thus creating a composite function with score and enjoyment on the *x*- and *y*-axes, respectively. The resulting function might look like the one found in Figure 8. The exact shape of the function is not important. Variations in the type of sigmoid or monomodal function yield compositions with similar properties, namely a restricted domain and a single peak. Based on the above assumptions, peak enjoyment happens at an expected score somewhere below the maximum, and that having an expected score at the minimum or maximum is a thoroughly unpleasant experience. Please see Appendix H for a more rigorous argument.

**Schmidhuber’s formal theory of fun**

In the field of artificial intelligence, one of the main goals is to get artificial agents to interact with and learn from the environment the way humans do. Alan Turing, the father of modern computing, discussed the importance of simulated learning as far back as the mid-20th century (Turing, 1950). In order to get artificial agents to learn the way humans do, it would be necessary to first understand what motivates humans to learn, and to then implement some form of this motivation in a manner a computer could understand.

Schmidhuber (2010) discusses such a way. If an agent (human or artificial) has an internal model of the world, it requires a certain amount of memory to hold all of the information pertaining to that model. Once in a while a discovery might take place that would illuminate a redundancy in the data. This illuminating discovery would allow the agent to recode the information for the world model into a smaller footprint. For instance, Newton’s discovery of the law of gravity gave a common explanation for apples falling from trees and planets orbiting the sun.

Schmidhuber defines beauty as the compressibility of a data set (Schmidhuber, 2007). He claims 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 that allows us to store a large amount of information with a small amount of memory. 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).

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.” 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 and the Yerkes-Dodson law**

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? The Merriam-Webster online dictionary defines frustration as “a feeling of anger or annoyance caused by being unable to do something.” In other words, frustration is the result of failure at a task at which one wanted to succeed.

Frustration can increase levels of arousal. What is arousal? Merriam-Webster online defines arousal as “a state of heightened physiological activity”. Studies show that physiological arousal increases with frustration. For instance, Doob & Kirshenbaum (1971) found that exposure to frustration leads to higher blood pressure. Similarly, Hokanson & Burgess (1964) used frustration to increase the heart rates of participants by 20 beats per minute.

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). This relationship is commonly represented with an inverted *U*-shaped curve with “arousal” on the *x*-axis and “ability” or “performance” on the *y*-axis, as shown in Figure 13. As argued above 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 person 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 a percentage of wrong answers exists that is non-zero that will lead to the most fun.

**Hypotheses**

The main purpose of this study is to pioneer a method to measure the relative desires of frustration and accomplishment in individuals who encounter novel computer-based tasks. The following four hypotheses lead to direct tests of the theories and models mentioned above. The first hypothesis of this paper is that for each individual, there is a curve representing enjoyment as a function of score on a task, that this curve has a single peak, and that this curve has minimal values for scores of 0% and 100% (please see Figures 8-12 for possible examples). The second hypothesis is that a participant would be willing and able to adjust a task to match this difficulty level if given the means to do so. The third hypothesis is that the peak of this function (and thus peak enjoyment) would be in the 80%-95% score range. The fourth hypothesis is that the percentage correct at peak enjoyment would not be dependent upon task type or amount of gaming experience. The fifth hypothesis is that the a user’s sensitivity to negative stimuli (as measured by self-report on a neuroticism inventory) would predict his or her preference for percent correct on a task – specifically that higher neuroticism scores would correlate with higher preferred percent correct.

**Method**

**Participants and design**

Participants were recruited online and on the NDNU campus. Online recruitment occurred via ads posted at Facebook, Reddit and Craigslist. Flyers were posted around the NDNU campus. The ads and flyers emphasized that the study involved playing three video-game-like tasks that would take approximately 10 minutes total, and that individuals with all levels of gaming experience were needed for the study (please see Appendix F). Would-be participants were directed to the website [www.ndnuvideogamestudy.com](http://www.ndnuvideogamestudy.com) in order to be part of the experiment.

Between 3/18/2014 and 3/25/2014, 11 participants took part in the study. Of these, data from the first 10 were included in the pilot data (*N* = 10). The group was young (age range 18 - 36, *μ* = 29.3), and overwhelmingly consisted of people who spent large amounts of time gaming (*n* = 9 in the “high” game usage segment, *n* = 1 in the “medium” game usage segment, as described below). For the full study, the author will take steps to reduce this skew, as described in the “discussion” section below.

Participants were divided into three groups based on the amount of experience they had with computer gaming. These levels were called “low”, “medium” and “high” game usage. In order to define these groups, we relied on the definitions of The NPD Group, 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. 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. We defined the “low” usage group to include those individuals who were unlikely to play games in any given week. Correspondingly, the “medium” group consisted of those who played somewhere between zero and five hours of games during a typical week. Volunteers were sorted into “high”, “medium” and “low” game usage groups based on this definition. 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 3x3 mixed-subjects quasi-experiment. The quasi-independent variables were “game usage” and “task type”. Levels for game usage were “low”, “medium” 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 three 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 two dependent variables. The first was “score at chosen difficulty”, which is a percentage from 0 to 100%. The second was “chosen difficulty after the test”. These two variables are described below.

For a given participant and a given task, the participant was asked to complete 25-30 trials of that task across three different phases. Phase one was the tutorial, where the participant completed four or five tasks of increasing difficulty with the help of detailed instructions. Phase two was practice, where the user completed 15 tasks and the difficulty level varied between tasks as decided by an algorithm. Phase three was the test phase where the user completed 10 tasks at a user-selected difficulty level. 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. The difficulty was fixed at this chosen level during phase three. 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: “if you were to do this task again, what difficulty would you prefer?” The difference between this value and the previously expressed preferred difficulty value was recorded as the second dependent variable described above.

**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 and adjust the difficulty level. 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 be 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 a reasonably short period of time (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 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” (Please see Appendix A for example). 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 (Please see Appendix B for example). 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 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 randomly 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”) (please see Appendix C for example). 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 (please see Appendix D for example). 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 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 (Please see Appendix E for example). 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, and then the ball reappeared at the beginning of the maze. 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 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.

Ads were placed on Facebook for volunteers. The ad 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 F 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 G 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 medium game usage category, and a response of three through five put the participant in the high game usage category.

After gaming habits were recorded, the user was prompted to answer three more demographics questions about gender, race/ethnicity, and income. Then the user completed a neuroticism inventory. (Start here)

Once gaming habits 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).

At the beginning of each task, there was a short (less than one minute) tutorial on the nature of the task, how to input answers, and how to alter the difficulty level via the slider. Participants were guided through at least 20 items of each task. The first five items comprised the tutorial. Between the tutorial and the test, the user was able to set the difficulty level on each item and practice for as long as he or she wished. After the participant chose to end the practice session, he or she was prompted to pick the level of difficulty level that was the most fun. Following this selection was the test session, which was composed of 15 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.

Once all three tasks were completed, the responses and any additional comments were coded into an HTTP-post request to hidden page on the [www.ndnuvideogamestudy.com](http://www.ndnuvideogamestudy.com) website. The page that caught the request (and stored the data in a MySQL database) was written in PHP. 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, the HTTP-post catcher, and the database extraction (also written in PHP) will be available online after the completion of the data gathering phase and closure of the database.

**Analysis**

The main analysis of this experiment focused on establishing estimated preferred difficulty values for each of the nine combinations of task type and game usage. For each participant, three estimates of this measure were possible from the data. Two are described here, and one is described in the secondary analyses section. The first estimate of preferred difficulty is the score that the participant received on the fifteen test items of each task. The reason that this is an estimate of preferred difficulty is that the user chose the difficulty level as the most fun. The second estimate can be interpolated from the combination of percentage correct on the task and the desired change in difficulty level after the task is complete.

The first estimate is a measure of prospective preferred difficulty. That is, it is a measure of what a participant would like for the near future. This can be calculated by a simple percentage of correct answers for the test phase of the task. Each participant generated one such percentage for each task. Averages per user varied from a minimum of .467 to a maximum of .933, with a mean of .723 and a standard deviation of .129. The Kinesthetic task produced the lowest average at .61, and the Verbal task produced the highest average at .8. The mean average of a task was .723 with a standard deviation of .100. These data deviate significantly from the hypothesized value of 80% to 95%. Please see Figure 14 for this data.

The second estimate was a measure of reflective preferred difficulty. That is, it measured the participant’s attitude toward his or her experience with the recent past. For each task, a participant generated a score and a preferred change in difficulty, or Δdiff, which was described in the method section. These pairs generated a scatter plot with score on the *x*-axis and Δdiff on the *y*-axis. These Δdiff scores were only comparable for a given task type. By looking at different subsets of the data, we can generate twelve such scatter plots. These are: nine for each cell in the 3x3 grid of data; and three for the total set of data in each of the three task types. However, given the skew in game usage type found in the pilot study, this analysis will only focus on the total data set for each task type.

For each scatter plot, a regression can be performed. According to the theories set out in the introduction, higher scores on tasks should lead to desires for increased difficulty (positive Δdiff), and lower scores will lead to desires for decreased difficulty (negative Δdiff). Thus, we would expect there to be a negative slope to the regression line. (Since for the Kinesthetic and Visual tasks, high slider values indicated easier tasks, we would expect this result to be reversed on those tasks). We would expect the *x*-intercept of this line to fall somewhere between the minimum and maximum score for the task. This *x*-intercept represents the best estimate of the population’s desired difficulty level. That is, this point on the regression line is where the predicted Δdiff equals zero, and where a participant would likely want to keep the difficulty the same.

In order to test the hypothesis that lower scores led to higher Δdiff values on the Verbal task, a regression was performed. The result approached significance, with *β* = .604, *R*2 = .364, *p* = .065, *p* > .05. In order to test the hypothesis that lower scores led to lower Δdiff values on the Kinesthetic task, a regression was performed. The result was not significant, with *β* = -.322, *R*2 = .104, *p* = .364, *p* > .05. In order to test the hypothesis that lower scores led to lower Δdiff values on the Visual task, a regression was performed. The result approached significance, with *β* = -.584, *R*2 = .341, *p* = .076, *p* > .05.

To test whether the *x*-intercept of the regression line on the Verbal task was within the hypothesized range, the regression line was constructed. The equation of this regression line was *y* = -2.97 + 3.46*x*. The *x*-intercept had a coordinate of .859, which was within the hypothesized range of .8 to .95. To test whether the *x*-intercept of the regression line on the kinesthetic task was within the hypothesized range, the regression line was constructed. The equation of this regression line was *y* = 2.2 - 5.54*x*. The *x*-intercept had a coordinate of .397, which was outside of the hypothesized range of .8 to .95. To test whether the *x*-intercept of the regression line on the Visual task was within the hypothesized range, the regression line was constructed. The equation of this regression line was *y* = 5.6 - 7.89*x*. The *x*-intercept had a coordinate of .710, which was not within the hypothesized range of .8 to .95.

To test the hypothesis that there was no difference between preferred difficulty levels for the three different tasks, within-subjects, repeated measures ANOVA test was run on the outcome percentages of each participant. For this ANOVA, *F*(2, 8) = 2.174, *p* = .176, *p* > .05. This result did not achieve significance, which was in agreement with hypothesis four.

Another question of importance for this study is whether learning continued to take place during the test session, and if so, at what rate? Ideally no learning would take place, since that could confound the relationship between percentage correct and Δdiff as described above. In order to test for the presence of learning, a linear regression was run on a scatter plot of percentage correct vs. item number on test phase data. If there was learning taking place, then one would expect the slope of the linear regression to be significantly greater than zero.

To test whether learning occurred in the kinesthetic task, a plot of index vs. outcome was generated for all kinesthetic test data, and a regression was run. No significant relationship was found, with *β*= .003, *R*2 = .000, *p* = .969, *p* > .05. To test whether learning occurred in the Visual task, a plot of index vs. outcome was generated for all Visual test data, and a regression was run. No significant relationship was found, with *β*= -.051, *R*2 = .003, *p* = .539, *p* > .05. To test whether learning occurred in the Verbal task, a plot of index vs. outcome was generated for all Verbal test data, and a regression was run. No significant relationship was found, with *β* = -.322, *R*2 = .104, *p* = .371, *p* > .05.

**Discussion**

Though data was sparse for this experiment, there were some interesting patterns that emerged. First of all, since the *p*-values for the beta coefficients approached significance in two of the three task types, it looks as though there is a relationship between percent correct and a user’s change in difficulty, lending some support to hypotheses one and two. This result agrees with all of the theories laid out in the introduction section. In the full version of this experiment, we are hopeful that there will be significant values for these beta coefficients, which would indicate strong support of the hypothesis and provide experimental evidence for the flow theory, the formal theory of fun, and the zone of proximal development theory.

Second, there seems to be a lack of evidence for learning taking place after the beginning of the test section. Once more data is available, we will run more types of tests to be sure. These tests will include logistic as well as linear regressions, and generation of *a posteriori* logistic curves using bootstrapping.

Third, it appears that we overestimated the percentage correct that participants preferred on tasks. Few participants scored above 80% at their preferred difficulty level, and the average percent correct of 72.3% is distinctly lower than the 80% - 95% hypothesized range. Additionally, the regressions of percent correct vs. Δdiff intersected the *x*-axes at lower values than expected. The *x*-value of 39.7% for the kinesthetic task was especially noteworthy.

It was intended that the data from the practice phase of each trial would be used to attempt to construct a logistic function relating difficulty level to percentage correct for each participant using the maximum *a posteriori* method. This method uses Bayesian reasoning to establish parameters of a distribution that generate the highest likelihood of a given set of observations. For each item in phase one, a participant will have a difficulty level and either a success or failure (coded as a 1 or a 0). As described in the first part of the introduction, it is assumed that the function relating difficulty level and chance of success is logistic. Any logistic function can be completely described by a mean and a shape parameter. For each combination of mean and shape parameter, it is possible to generate a likelihood that the participant’s responses would be the given set of successes and failures. By examining this likelihood function over the set of all means and shape parameters, it might have been possible through search methods to find it and use it to predict how hard a task would actually be to a user.

However, it turns out that participants did not generate enough of the right kinds of data to pin down such a function. In order to generate this function, more practice data at more varied difficulty levels would be needed. The author is contemplating ways to coax this data from users for the full study. One such way would be to take control of the difficulty slider away from the participant until we have achieved enough data to construct a reliable curve. However, this would lengthen the time commitment of a participant and likely cause a lower completion rate among respondents.

The sample for this study leaves much to be desired. First of all, the size will need to be increased. We will seek more volunteers from more communities in the future. The IRB has given this study approval to seek volunteers from Reddit, Craigslist and the NDNU campus, so it should be easy to reach more people for the next round of data.

A second problem in the sample is the skew toward individuals who spend significant time playing video games. The most likely reason for this skew was in the presentation of the study - for instance, the name of the website included the words “video game”, and that likely caused non-gamers to believe that they were not qualified to take part. For the next round of data, the messaging will be tweaked to put less emphasis on the “video game” aspect. We hope that this will reduce the skew in the sample.

Since this study assumed that frustration is an integral aspect of fun, it would follow that an individual’s sensitivity to frustration should impact that individual’s preferred difficulty level. Specifically, according to this model those who are prone to frustration would prefer a higher percentage correct on a task. Szalma & Taylor (2011) found that the personality trait of neuroticism correlated with frustration for dealing with automation. Similarly, the author of the current paper would hypothesize that high measures on the neuroticism trait would correlate with preferences for higher percentages correct on a task. Future studies will seek to establish this correlation.

**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.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York, NY: Harper & Row.

Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life*. Basic Books.

Csikszentmihalyi, M., & LeFevre, J. (1989). Optimal experience in work and leisure. *Journal Of Personality And Social Psychology*,*56*(5), 815-822. doi:10.1037/0022-3514.56.5.815

Dunn, R. ( 1993). Learning styles of the multiculturally diverse. *Emergency Librarian*, *20*( 4), 24– 32.

Doob, A. N., & Kirshenbaum, H. M. (1973). The effects on arousal of frustration and aggressive films. *Journal of Experimental Social Psychology*, *9*(1), 57-64.

Hokanson, J. E., & Burgess, M. (1964). Effects of physiological arousal level, frustration, and task complexity on performance. *The Journal Of Abnormal And Social Psychology*, *68*(6), 698-702. doi:10.1037/h0049340

“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/>.

Koster, R. (2010). *Theory of fun for game design*. O'Reilly Media, Inc..

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

Schmidhuber, J. (2007, January). Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In *Discovery Science* (*pp*. 26-38). Springer Berlin Heidelberg.

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

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

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/

“Top 10 highest grossing video games ever”. (2012). Retrieved November 26, 2013 from <http://digitalbattle.com/2012/02/21/top-10-highest-grossing-video-games-ever/>

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433-460.

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

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.

**Figure Captions**

*Figure 1:* Flow Channel diagram

*Figure 2:* An approximation of the flow channel (from above, orange = high values, green = low values)

*Figure 3:* An approximation of the flow channel (from a 45 degree view)

*Figure 4:* An approximation of the flow channel (view from the origin into the first quadrant)

*Figure 5:* A slice of the flow channel at *y* = 3

*Figure 6:* A slice of the flow channel at *y* = 5

*Figure 7:* Logistic function with ceiling effect for difficulty values in [0, 1] and floor effect for values in [2, )

*Figure 8:* Possible composition of flow channel slice with inverse floor-ceiling function

*Figure 9:* Possible composition of flow channel slice with inverse floor-ceiling function

*Figure 10:* Possible composition of flow channel slice with inverse floor-ceiling function

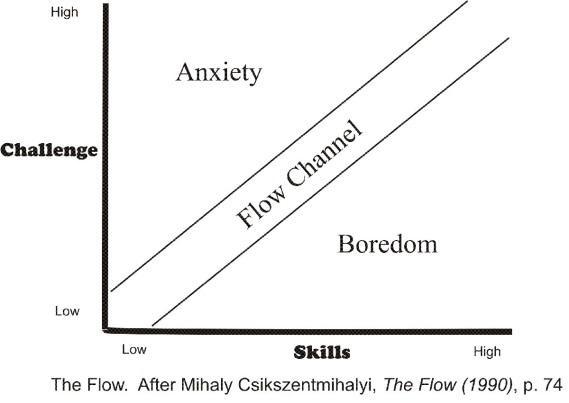
*Figure 11:* Possible composition of flow channel slice with inverse floor-ceiling function

*Figure 12:* Possible composition of flow channel slice with inverse floor-ceiling function

*Figure 13:* Yerkes-Dodson law (retrieved from http://en.wikipedia.org/wiki/File:HebbianYerkesDodson.JPG)

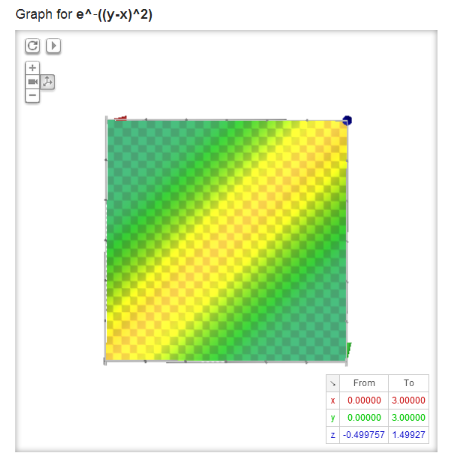
*Figure 14:* Table of percent correct by participant and task type

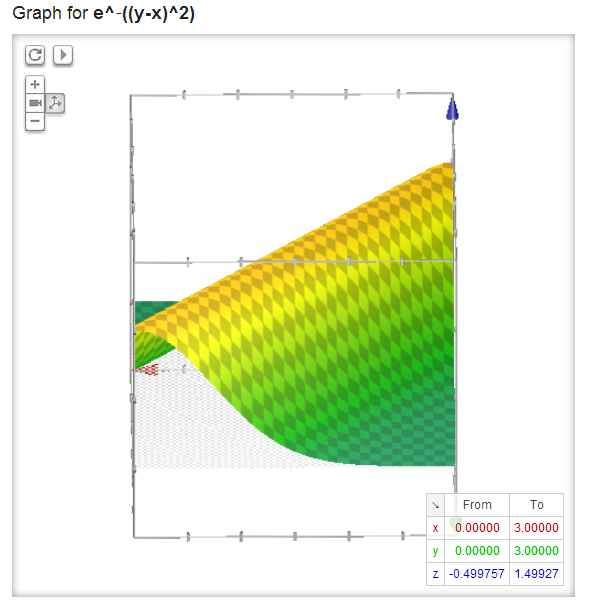
**Figure 1**

Flow Channel diagram

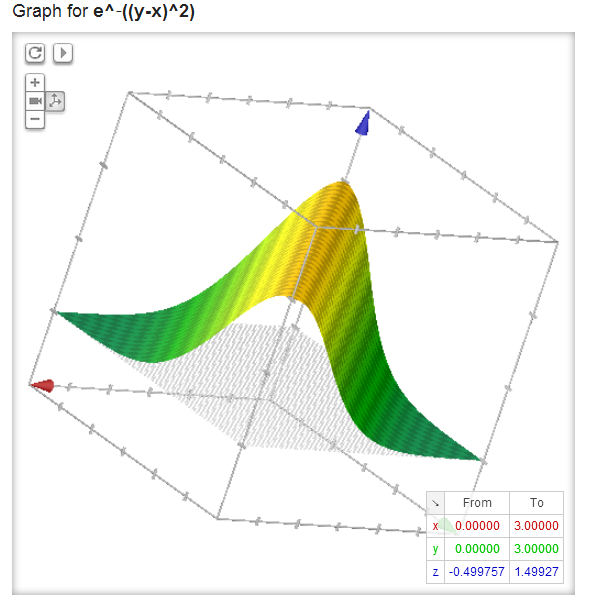
**Figure 2**

**Figure 2**

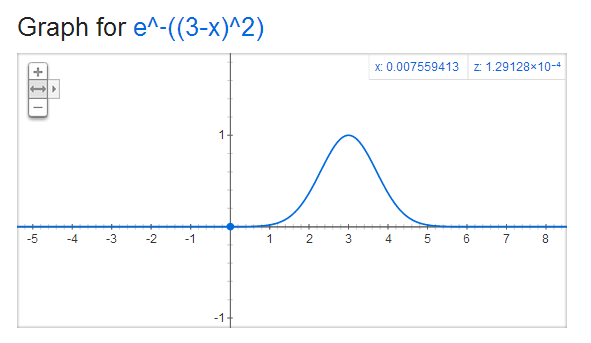
An approximation of the flow channel (from above, orange = high values, green = low values) **Figure 3**

An approximation of the flow channel (from a 45 degree view) 

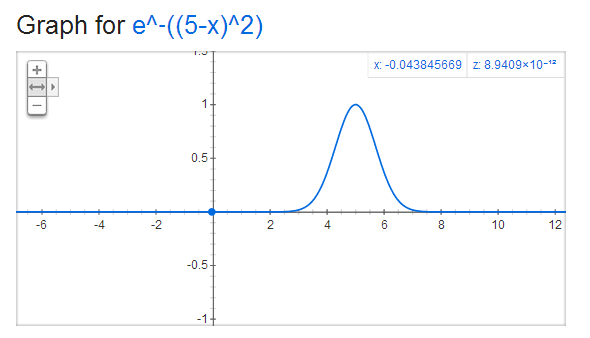
**Figure 4**

An approximation of the flow channel (view from the origin into the first quadrant) 

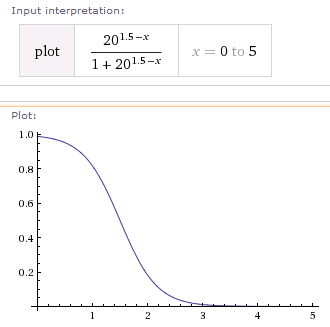
**Figure 5**

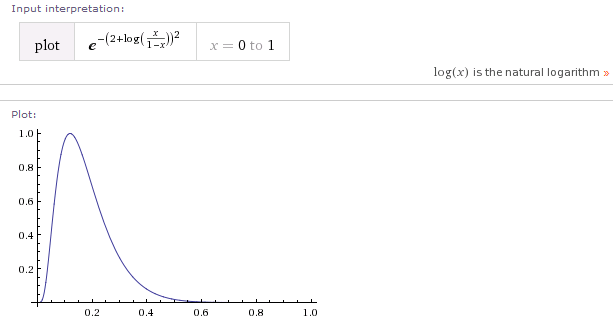
A slice of the flow channel at *y* = 3

**Figure 6**

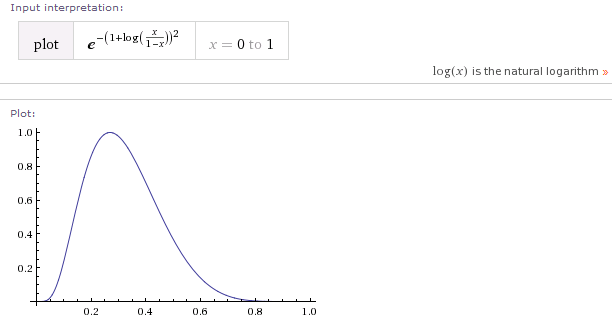
A slice of the flow channel at *y* = 5**Figure 7**

Logistic function with ceiling effect for difficulty values in [0, 1] and floor effect for values in [2, )

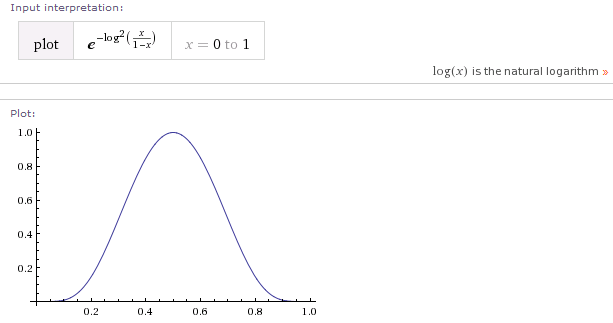


**Figure 8**

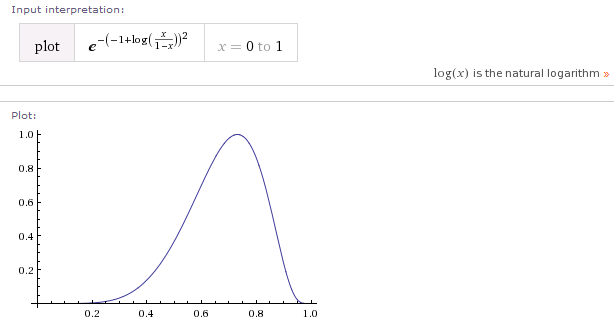
Possible composition of flow channel slice with inverse floor-ceiling function (max at .15)

**Figure 9**

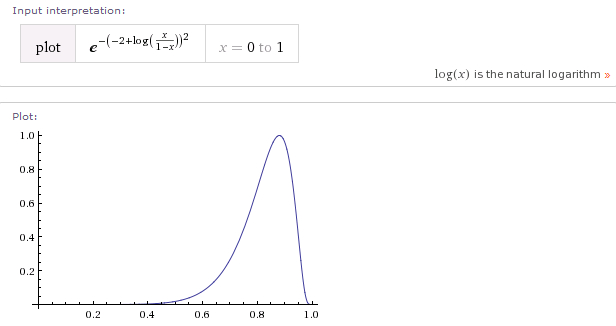
Possible composition of flow channel slice with inverse floor-ceiling function (max at .25)

**Figure 10**

Possible composition of flow channel slice with inverse floor-ceiling function (max at .5)

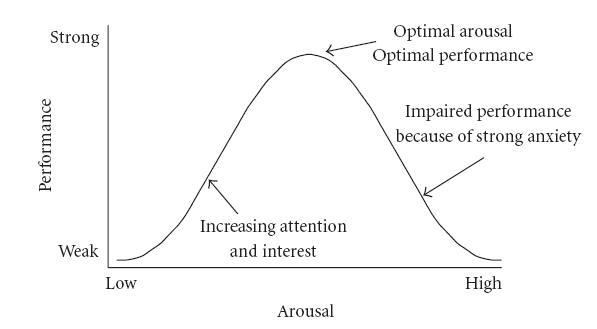
**Figure 11**

Possible composition of flow channel slice with inverse floor-ceiling function (max at .75)

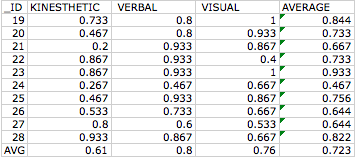
**Figure 12**

Possible composition of flow channel slice with inverse floor-ceiling function (max at .85)

**Figure 13**

Yerkes-Dodson law (retrieved from http://en.wikipedia.org/wiki/File:HebbianYerkesDodson.JPG)

**Figure 14**

Table of average percent correct by participant and task type**Appendix Captions**

*Appendix A*: Visual Task Prompt

*Appendix B*: Visual Task, Answer Revealed

*Appendix C*: Verbal Task Prompt

*Appendix D*: Verbal Task, Answers Revealed

*Appendix E*: Kinesthetic Task Prompt

*Appendix F*: Recruitment Flyer

*Appendix G*: Informed Consent Form

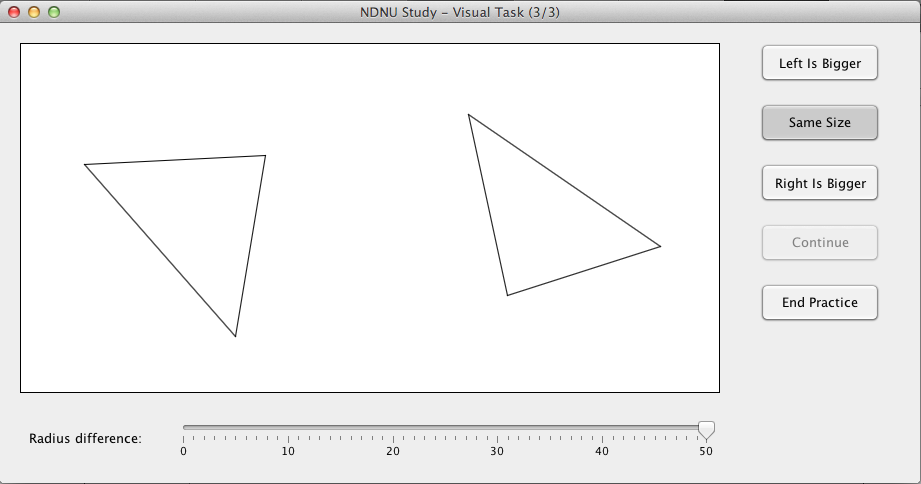
*Appendix H*: Argument from Flow and Range Effects revisited with function estimations

*Appendix I:* Debriefing Form

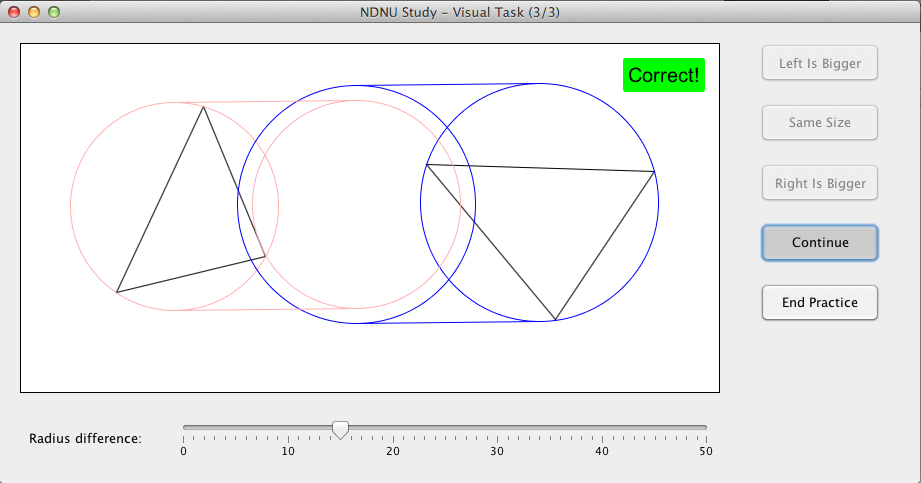
*Appendix J*: Derivation of Composition of Flow Channel Slice with Inverse Floor-Ceiling Function

**Appendix A**

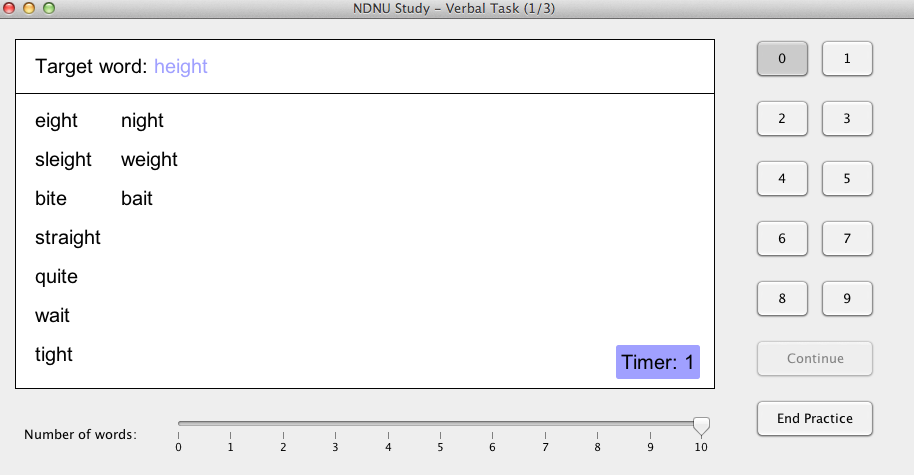
Visual Task Prompt

**Appendix B**

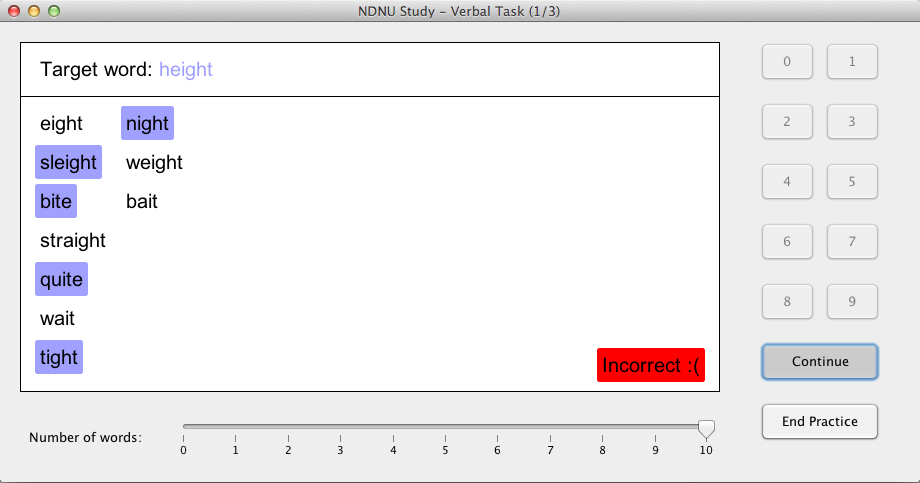
Visual Task, Answer Revealed



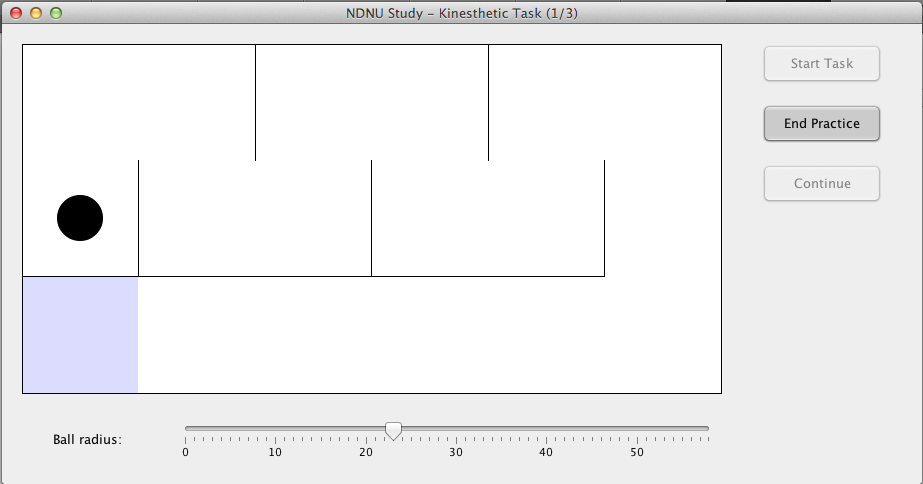
**Appendix C**

Verbal Task Prompt

**Appendix D**

Verbal Task, Answers Revealed

**Appendix E**

Kinesthetic Task Prompt

**Appendix F**

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](http://www.ndnuvideogamestudy.com)**Appendix G**

Notre Dame de Namur 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](mailto: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 H**

Argument from Flow and Range Effects revisited with function estimations

Csikszentmihalyi gives a visual description of the “flow channel” in a diagram (please see figure 1). The *x*-axis is “skills” and the *y*-axis is “challenge”, and there is a thin region along a 45-degree incline where flow is reached. Above the flow region is an area corresponding to anxiety, and below is a region representing boredom. (Csikszentmihalyi, 1990)

A model for this diagram would be to assign a function *z* = *f*(*x*, *y*) which is maximum in the flow channel and zero in the boredom and anxiety regions. In this function, z is a measure of enjoyment, x is a measure of skills, and y is a measure of difficulty of task. Preferably this would be a simple function that is also smooth. One such function which satisfies these criteria is z = *e*-(*y-x*)^2, which is graphed in figures 2 through 4. In these figures, orange (light colors) represents high values of *z* enjoyment and green (dark colors) represents low values of enjoyment.

At a given time, any individual brings a certain set of skills to a task. In terms of the mathematical model described above, this means that the value of x for that individual is fixed at the time of the task. So, for a fixed value of x, we will get a graph of z (enjoyment) vs. y (difficulty). Such a graph would resemble a bell curve, as shown in figures 5 and 6. The higher the skill, the farther to the right the bell is shifted. This z vs. y curve represents a “slice” through the flow graph of figure 1 along a vertical line.

For the second part of this model, consider a task that has a “floor effect” and a “ceiling effect”. A “Ceiling effect” is the result of a difficulty range where the participant gets nearly all tasks correct. Correspondingly, a “floor effect” is the result of a difficulty range where the participant does no better than random chance. In this model, it is assumed that the floor effect causes percentage correct to plateau at 0%. The results of this model do not depend on this assumption. However, it is easier to demonstrate these results by use of a concrete function.

Consider a task where a difficulty level between zero and one resulted in a ceiling effect, and a level of two or more resulted in a floor effect. In order to model this function with a smooth curve, one option would be to use a logistic function, as shown in figure 7. The y-axis on the graph is percent correct by a participant, and the *x*-axis is difficulty level. This logistic function approaches 100% for difficulty levels below 1 and approaches 0% for difficulty levels above 2.

The third part of this model is to combine the functions we have into a single composite function. There is a composition of these functions that maps percentage correct to enjoyment level. This function is *z* = *f*(*g*-1). Please see Appendix J for a derivation of this function. Figures 8 through 12 show different possibilities for this composite function based on different constants associated with the original functions. Though they differ in shape, all of these functions share a number of properties. They all are defined for inputs between *x* = 0% and *x* = 100%, they all have a single maximum value somewhere between *x* = 0 and *x* = 1, and all contain the points (0%, 0) and (100%, 0). This would imply that the maximum enjoyment happens at a score that is neither 0% nor 100%, and that enjoyment is minimized at these endpoints.

**Appendix I**

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](mailto: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](mailto:misterriley@gmail.com).

**Appendix J**

Derivation of Composition of Flow Channel Slice with Inverse Floor Ceiling Function

As noted above, our model for the flow channel function is *z* = *e*(*y* - *x*)^2. A more general version is to allow a shape parameter k as a coefficient on the exponent: *z* = *ek*(*y* - *x*)^2. Slicing the flow channel is equivalent to setting y to a constant c: *z* = *ek*(*c* - *x*)^2

The general form of a logistic function is *y* = 1/(1 + *ea* + *bx*). Solving for the inverse of this function yields

*a* + *bx* = *ln*((1 - *y*)/*y*)

*x* = (*ln*((1 - *y*)/*y*) - *a*)/*b*

Substituting this x into the flow channel slice:

*z* = *ek*(*c* - (*ln*((1 - *y*)/*y*) - *a*)/*b*)^2

Let us define new constants *d* = *k*/*b*2 and *f* = *cb*2. The function simplifies to

*z* = *ed*((*f - a*) - *ln*((1 - *y*)/*y*))^2

Defining *g* = *f - a*, we now have this function in two constants:

*z* = *ed*(*g* - *ln*((1 - *y*)/*y*))^2

Examples of this function are graphed in Figures 8 through 12. Each of these functions has *d* = 1. The values of *g* are integers that range from -2 to 2.