Optimizing Game Engagement Via Nonparametric Models and Manipulations of Difficulty, Tension, and Perceived Performance

Mohammad Khajah

University of Colorado at Boulder, mmkhajah@gmail.com

Follow this and additional works at: https://scholar.colorado.edu/csci_gradetds

Part of the Computer Sciences Commons, Educational Methods Commons, and the Psychology Commons

Recommended Citation

https://scholar.colorado.edu/csci_gradetds/160

This Dissertation is brought to you for free and open access by Computer Science at CU Scholar. It has been accepted for inclusion in Computer Science Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
Optimizing Game Engagement via Nonparametric Models and Manipulations of Difficulty, Tension, and Perceived performance

Mohammad M. Khajah

B.Sc., Syracuse University, 2007
M.Sc., Syracuse University, 2008

A thesis submitted to the Faculty of the Graduate School of the University of Colorado in partial fulfillment of the requirement for the degree of Doctor of Philosophy

Department of Computer Science
2017
This thesis entitled:

Optimizing Game Engagement via Nonparametric Models and Manipulations of Difficulty, Tension, and Perceived Performance

written by Mohammad M. Khajah

has been approved for the Department of Computer Science

__________________________________________

Michael C. Mozer

__________________________________________

Clayton H. Lewis

Date: 25 Aug 2017

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above-mentioned discipline.

IRB Protocol #: 14-0750
Abstract

Khajah, Mohammad M. (Ph.D, Computer Science)

Optimizing Game Engagement via Nonparametric Models and Manipulations of Difficulty, Tension, and Perceived performance

Thesis directed by Professor Michael C. Mozer

I study the impact of novel game manipulations on user engagement using principled computational methods. Maximizing user engagement is important because it results in more profitable games in the commercial arena and better learning outcomes in the educational arena. It is then perhaps unsurprising that the study of user engagement is well established. Most work uses a classical A/B paradigm, in which a few, often binary (on/off), design decisions are manipulated. Recently, optimization studies have begun to explore a range of discrete or continuous levels. The majority of work in both types of studies is concerned with manipulations such as aesthetics, rewards, and difficulty. While many of these manipulations are found to increase engagement, little work has been done on utilizing theories of engagement from other domains, such as gambling and storytelling, to improve user game engagement. For instance, the tension-and-release manipulation, a common technique in storytelling and music composition for controlling event progression, is usually discussed within the gaming context only qualitatively as a way of controlling difficulty over time. The near-win effect—an increase in motivation due to almost winning a game—comes from gambling psychology. Another understudied manipulation is the perception of difficulty, where the user’s perception of the challenge is controlled independently from actual challenge or vice versa. Undoubtedly, game designers are using these manipulations—near-win, tension-release and perception of difficulty—in their games but I am not aware of work that systematically explores how different levels of these manipulations influence user engagement. In this thesis I study these manipulations systematically using Gaussian processes, neural network, and preference learning models. Results from multiple Bayesian optimization experiments show that maximum engagement occurs when the user’s perception of difficulty is manipulated moderately, suggesting the critical role of a user’s self-perception of competence. A/B and random assignment studies show that engagement in a web-based memory training game can be modulated via tension-and-release difficulty curves.
Finally, a massive study with thousands of students shows that the near-win effect significantly improves engagement of lower-performing students.
Dedication

This thesis is dedicated to the souls of my grandmother Khairulnessa’ AlMousawi, my uncle Salch Khajah, and my great-aunt Hajer Sadeq.

I also dedicate this work to my parents Mahmoud and Tahera, my aunt Hamida, my sister Fatima, and my brothers Ali and Zaid.
Acknowledgments

I would like to thank my advisor and mentor, Prof. Michael Mozer, for his belief in me and for his continuous support of my research. Prof. Mozer’s impeccable attention to detail, insistence on clarity, and firm commitment to using principled methods allowed me to grow as an academic and to become a better reader of scientific literature. Despite having a busy schedule, Prof. Mozer always found the time to give constructive feedback on experimental findings or on a piece of my writing, which helped enormously during the writing of this thesis. His vast knowledge of computational modeling literature and expertise in neural networks provided the backbone for all of my publications and of course, for this thesis. So I feel both grateful and privileged to have worked under Prof. Mozer’s supervision.

I would also like to thank the other members of my committee, Prof. Clayton Lewis, Prof. Sarel van Vuuren, Dr. Rafael Frongillo, and Dr. Nisar Ahmed, for taking the time to participate in the proposal and thesis defense despite hectic schedules and other commitments. The committee’s feedback and suggestions framed large portions of this thesis and helped make it a better contribution to the field.

Special thanks go to my friend Dr. Robert Lindsey, a recent alumnus of Prof. Mozer’s lab, for showing me the “Bayesian way” and for sharing source code of his computational models. Dr. Lindsey’s mastery of computational modeling and creative way of doing experimental manipulations is something that I always think about when doing any modeling work or when designing new experiments. I would also like to thank my lab mate and friend Dr. Brett Roads who has been in the lab since I joined and with whom I’ve shared many stimulating discussions about research and hobbies. Dr. Road’s presentations and visualizations have always been exceptional and set a high bar that I strive to match during my own presentations. In addition to Rob and Brett, my friends and fellow lab mates, Shrutih Sukumar, Dr. Shirley Quesada, Karel Ridgeway, and Camden Elliott-Williams, were a solid source of thoughtful feedback and revealing insights so I am greatly thankful to have been a part of their group.

I would like the wonderful people at Wootmath, especially Krista Marks, Sean Kelly, Brent Milne, Bill Troxel, and Adam Holt for helping me conduct a study on their online tutoring platform. A large portion of this thesis would not have materialized without their cooperation.

Finally, I would like to thank my parents, my aunt, my sister, and my brothers for their...
continuous support and encouragement and for alleviating the homesickness that accompanies the long and arduous PhD journey.
Contents

1 Background 1

1.1 Engagement Factors .................................................. 2
1.2 Player Modeling and Engagement .................................... 7
1.3 Optimization .............................................................. 12
  1.3.1 Static Optimization .................................................. 13
  1.3.2 Dynamic Optimization .............................................. 16
1.4 Limitations ............................................................... 19

2 Designing Engaging Games Using Bayesian Optimization 20

2.1 Introduction ............................................................... 20
  2.1.1 Recent Research on Game Optimization ............................. 22
2.2 Gaussian Processes and Bayesian Optimization ..................... 23
  2.2.1 An Illustration of the Bayesian Approach ............................ 26
  2.2.2 From Persistence to Total Play Duration .......................... 28
  2.2.3 Selecting an Observation Model .................................... 28
2.3 Experiments ............................................................... 31
  2.3.1 Overt Versus Covert Difficulty Manipulations ..................... 32
  2.3.2 Two Games and Three Difficulty Manipulations ................... 33
  2.3.3 Flappy Bird: Experimental Methodology and Results ............. 34
4.4 Discussion ................................................................. 96

5 Conclusion and Future Work ........................................... 99
5.1 Future Work ........................................................... 101

References ................................................................. 102
List of Tables

1.1 Summary of reviewed engagement factors. ........................................ 7
1.2 Summary of player modeling papers. ............................................. 12
1.3 Summary of the optimization papers. ............................................. 19
3.1 Summary of Experiment 3.1 results .............................................. 57
3.2 Summary of Experiment 3.3 results .............................................. 66
3.3 Summary of Experiment 3.4 results .............................................. 71
3.4 Recency vs. questionnaire ratings ................................................... 72
# List of Figures

1.1 Illustration of manipulations of the spatial n-back task ........................................ 4  
2.1 Illustration of model-free vs. model-based approaches to experimentation .................. 25  
2.2 Persistence probability of the Treefrog Treasure players over the design space ............... 27  
2.3 Simulation experiment of different observation models ............................................ 29  
2.4 Screenshots of Flappy Bird and Spring Ninja ...................................................... 34  
2.5 Post-experiment questionnaire ................................................................................ 37  
2.6 Bayesian model fits of VTA and questionnaire scores ............................................... 39  
2.7 Bayesian model predictions of VTA, questionnaire scores and difficulty in SpringNinja ... 44  
3.1 The flow channel .................................................................................................. 47  
3.2 Screenshot of the Jetset game .................................................................................. 51  
3.3 Screenshots of ShapeFactory .................................................................................. 52  
3.4 The three scroll speed functions in Experiment 3.1 ................................................. 54  
3.5 The game condition selection screen in Experiment 3.1 ........................................... 55  
3.6 The mean accuracy of each lag value in Experiment 4.1 ......................................... 56  
3.7 The experimental manipulations in Experiment 3.2 ............................................... 59  
3.8 The post-round rating screen in Experiment 3.2 ..................................................... 59  
3.9 Bayesian model fits to residual ratings in Experiment 3.2 ........................................ 61  
3.10 Subjects’ scores over time in Experiment 3.2 ......................................................... 62
4.13 Predicted replay probability as a function of the original score, excluding natural near-wins, using logistic and GP models

4.14 The interaction between near-win events and student ability

4.15 Predicted replay probability as a function of the two-dimensional design input

4.16 Predicted replay probability using the additive neural network model
Chapter 1

Background

Video game developers want to design games that attract and engage players for the obvious commercial reason of maintaining a loyal customer base. Educational game designers want to develop engaging games so that students play more and consequently, learn more. Engagement has no precise definition and researchers have operationalized it via surveys, such as the Game Engagement Questionnaire (GEQ) (Brockmyer et al., 2009), physiological measures, such as galvanic skin response, and behavioral measures, such as duration of play. The literature on player engagement can be broadly segmented into three areas: engagement factors, player modeling and optimization.

Literature on engagement factors often draws on Csikszentmihalyi’s theory of flow (Chen, 2007), which prescribes factors that can induce a desirable psychological state in which the person becomes fully absorbed in an activity, loses track of time and experiences high level of enjoyment. In the video game domain, the analogues of these factors, such as challenge, control and objectives, are A/B-tested to determine their impact on engagement.

Player modeling papers tend to be grounded in marketing literature where the objective is predict player retention (how long they will play) or churn (whether they will continue playing or not). A standard recipe in this domain is to map game or player states to feature vectors and then regress retention durations or churn labels on these features. Papers here do not typically draw inferences about design adjustments needed to maximize engagement, with a few only informally providing such
Optimization papers are split into two camps: static and dynamic optimization, both of which assume that a game can be parameterized to facilitate the application of automated search algorithms. Static game design optimization algorithms try to find game designs that are suitable for a population of individuals, in terms of engagement or some other measure. On the other hand, dynamic algorithms ensure that the player is meeting an adequate level of challenge during the game. Hence, difficulty adjustment approaches operate on a within-subject basis. The two classes of algorithms are not mutually exclusive and can work together. Skill matching papers, such as the famous TrueSkill algorithm (Herbrich, Minka, & Graepel, 2006), are not included as those are concerned with multiplayer games whereas the focus here is on single human player scenarios.

1.1 Engagement Factors

Cordova and Lepper (1996) studied three task-irrelevant manipulations to a learning game for their impact on student motivation. These included embedding the game in a fantasy context (contextualization), customizing the context with student information such as the student’s best friend’s name (personalization), and offering the student choice over in-game avatars (choice). Students competed in a race against the computer on a number line containing 50 spots where the winner was the first one to finish. The number line contained shortcuts and teleportation zones. Students moved on the line by combining game-provided numbers and arithmetic operations to determine the number of steps to advance. Students were assigned randomly to five conditions: generic-fantasy-no-choice, generic-fantasy-choice, personalized-fantasy-no-choice, personalized-fantasy-choice and a no-fantasy control condition. The primary outcome measure was the improvement in performance as measured by the difference in scores between paper-and-pencil pre- and post-tests. Additionally, students filled a post experiment survey about how intrinsically motivated they were whilst using the game. Results showed that students playing versions of the game with motivational elements experienced higher levels of intrinsic motivation, i.e., liking the game, using more complex operations, potentially using it after school, etc. Students also exhibited learning gains on post test (up to about 50% in the personalized-choice condition vs. control). The authors noted that whilst motivational manipulations can be beneficial, they should be used
judiciously because they may be detrimental to students who were highly goal-directed.

Denny (2013) evaluated virtual achievements, another orthogonal manipulation to the learning task, as a motivational tool in an online learning application. Here students authored, answered, rated and commented on multiple choice question related to material from a university course. Achievement badges were created for authoring and answering questions as well as for using the software at regular intervals. When a student earned a badge, they were notified immediately via a message on the application screen. Students were required to author at least 1 question and to answer at least 20 questions to get 1.5% bonus course credit. Students used the online learning tool for 26 days, at the end of which there was a mid-term test. Two conditions were evaluated: a control condition with no badges and an experimental condition with badges. The outcome measure was the level of participation in the online learning tool, measured by three variables: number of questions authored/answered and the number of distinct days of participation. Students also voluntarily filled a survey, at the end of the experiment, about the learning value of authoring and answering questions and the motivational value of the inclusion of badges (experimental condition only). Results of the experiment showed that students in the experimental condition answered substantially more questions, 22% more than in those in the other condition, but there was otherwise no difference in the number of questions authored and the proportion of correct responses. Students in both conditions supported the learning value of the tool and those in the experimental condition also preferred the inclusion of badges. The null difference in the number of questions authored was attributed to the soft requirement of authoring only one question to receive bonus credit and to the fact that authoring questions requires more effort than answering them. Although students in the experimental condition answered more questions, the proportion of correct responses was the same as the other condition, which indicated no loss of quality over quantity.

Katz, Jaeggi, Buschkuehl, Stegman, and Shah (2014) studied the removal of gamification features in a cognitive training task, contrary to the previous two studies. Here gamification features were removed one at a time from an n-back spatial memory training game. In this game, students were shown a stimulus in one of six locations and had to indicate if the current stimulus matched the location of the one presented n trials back. The difficulty level n was increased (more difficult) or decreased (less difficult) based on the number of errors the subject had made. Seven versions of the game were developed
with the following motivational features removed one at a time: showing points, theming, showing lives and current difficulty level, exchanging points for prizes and earning end-of-session certifications at the end of each training day (Figure 1.1). The two remaining game versions included all and none of the features. Before training, students took an 2-back assessment task to establish a baseline after which they undertook 3 days of training. Surveys measuring engagement, excitement, difficulty and effort were collected after each day and averaged over the three days to obtain four variables. Teachers also rated how engaged students looked at the end of each day and the results were similarly averaged to obtain an observer engagement score for each student. Following the third day of training, students completed the 2-back assessment task again. Results showed that students playing game versions without the persistent display of points and lives showed significant improvement, as measured by average n level, over the three days of training. No meaningful differences in training improvement were observed for the other game versions, including the one without motivational features. Collected engagement, excitement, difficulty and effort measures also did not differ by condition. Performance on the transfer 2-back task at the end of training showed no differences in improvement across the game versions. The authors explained that these findings could due to the fact that even with all motivational features removed, the training game still looked like a game. They also noted that the short 3-day training period might not have been sufficient to delineate transfer improvements between conditions. Finally, the authors argued that motivational features should be used judiciously since they could distract students from the main task, as the training improvements resulting from the removal of scoring and lives showed.

In the context of non-educational games, a series of papers by Christoph Klimmt and colleagues investigated the roles of competition (Vorderer, Hartmann, & Klimmt, 2003), difficulty (Klimmt, Blake,
Hefner, Vorderer, & Roth, 2009), effectance and control (Klimmt, Hartmann, & Frey, 2007) and suspense (Klimmt, Rizzo, Vorderer, Koch, & Fischer, 2009) on player engagement.

To evaluate the role of competition, four scenarios were verbally described to players with each scenario corresponding to a combination of factors. Two factors were studied: possibilities available to the player, e.g., having many weapons and tools, and competition, e.g., whether there were monsters or not. The experiment was conducted with Tomb Raider players with each player randomly assigned to a scenario. Engagement here was measured on a 10-item scale of 1-5. Findings indicated that the most enjoyable scenario was the one with monsters (competition) and many possibilities to act (control) whilst the least enjoyable was the one with few possibilities and no monsters.

The role of suspense was studied on a popular first person shooter, Unreal Tournament 2, with two scenarios: nonsuspenseful and suspenseful. In both scenarios the player walked in a friendly middle eastern town to buy a vase at the bazaar. In the nonsuspenseful scenario, the objective was framed as a tourist walk in town whilst in the suspenseful version the objective was framed as a top secret mission to retrieve secret intelligence and potentially fight enemies. Both scenarios were otherwise the same. Subjects were assigned randomly to the two scenarios with each subject playing the game for 15 minutes. After that, subjects were asked to answer a 10-item engagement scale. The suspenseful version was found to be more enjoyable than the nonsuspenseful version.

The roles of effectance (the responsiveness of controls) and control (the degree to which the player could influence the game) were examined on a simple two dimensional game. Here the game objective was to deflect a ball via a player-controlled racket such that it destroyed all the bricks on the screen. Three conditions were evaluated: standard, reduced effectance and reduced control. In the reduced effectance condition, the game ignored one third of the player’s inputs. In the reduced control condition, the ball moved much faster compared to the standard condition. In all three conditions the player played the standard version of the game for two minutes, answered a questionnaire then played the experimental version for another two minutes and answered a questionnaire. The questionnaire measured perceived effectance, control and game engagement. Participants were randomly assigned to the three conditions. Results showed that reduced effectance reduced engagement substantially, compared to the standard condition, whilst reduced control did not produce a meaningful reduction in engagement.
The authors hypothesized that effectance silently contributed to player engagement and only became noticeable when it was reduced. On the other hand, control was hypothesized to be a salient characteristic of the game, potentially part of its challenge, which made it still enjoyable for players to play in the reduced control condition.

The role of difficulty was examined in an experiment with expert players playing a popular first person shooter, Unreal Tournament 2. Players were randomly assigned to three difficulty settings, the easiest and hardest of which were almost impossible to lose and win at, respectively. Players played for 10 minutes after which they answered three questionnaires on engagement, satisfaction with own performance and perceived difficulty. The three difficulty settings were shown to have significant effect on the number of enemies killed, deaths of player character and perceived difficulty. In terms of player satisfaction, players were most satisfied with their own performance on the easiest of difficulty settings, despite the very low level of challenge, and they were least satisfied on the hardest of settings. Similarly, players also rated the easiest difficulty setting as most enjoyable and the hardest difficulty as least enjoyable. The authors found these results to be incompatible with common psychological theories such as flow. Unrealistic difficulty manipulation, i.e., the easiest level being actually quite challenging, was ruled out because subjects' perception of game difficulty agreed with the actual difficulty. Instead, the authors explained that subjects might have thought that the experimenter was looking for high performance in the game. The authors also noted that the short play time might have biased players towards easiest difficulty settings.

Denisova and Cairns (2015) used a manipulation that involved changing the players prior information about the game before playing. Players played an action-adventure game in which the player had to collect and build objects in order to survive. The outcome measure was the player's immersion as measured via the Immersive Experience Questionnaire (IEQ). Players were also asked to fill a questionnaire about their performance in the game. Two conditions were evaluated: a “non-adaptive-AI” and an “adaptive-AI” condition where the latter was identical to the first except that subjects were told, vaguely, that the AI would adapt to their level of performance. Two experiments were run, a within subject experiment where subjects played both conditions (counterbalanced) and a between-subjects experiment where subjects were randomly assigned to the two conditions. Results showed that in both
experiments, subjects in the adaptive-AI condition felt slightly more immersed in the game (119.86 vs. 111.90 in the within-subjects experiment and 124.15 vs. 116.00 in the between-subjects experiment) Questionnaire results showed that subjects who were familiar with the concept of adaptive AI were less immersed than those who weren’t. The paper concluded that these results demonstrated that players had prior information that needed to be taken into account, e.g., such as gameplay reviews, when play testing games and when providing descriptions of game features.

1.2 Player Modeling and Engagement

Papers in this section propose models that predict player engagement via different measures such as retention (how long players play), churn (do players quit playing or not), and preferences (a player enjoying game A over B). These models can be used to determine whether intervention is required to keep the player playing the game. The main focus here are the details of the computational model used. None of the papers used the same dataset so it is impossible to compare the performance of various models.

Borbora, Srivastava, Hsu, and Williams (2011) compared the performance of theory-driven features, such as those from the previous section, and data-driven features on churn classification. Player activity logs from a role playing game were used to generate a dataset of player features and class labels.

Table 1.1: Summary of reviewed engagement factors.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Factors</th>
<th>Main Outcome Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Engagement</td>
</tr>
<tr>
<td>(Vorderer et al., 2003)</td>
<td>Competition</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>+</td>
</tr>
<tr>
<td>(Klimmt et al., 2007)</td>
<td>Effectance</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>null</td>
</tr>
<tr>
<td>(Klimmt, Rizzo, et al., 2009)</td>
<td>Suspense</td>
<td>+</td>
</tr>
<tr>
<td>(Klimmt, Blake, et al., 2009)</td>
<td>Challenge</td>
<td>Easiest &gt; Hardest</td>
</tr>
<tr>
<td>(Denisova &amp; Cairns, 2015)</td>
<td>Prior Information</td>
<td>+</td>
</tr>
<tr>
<td>(Katz et al., 2014)</td>
<td>Progress</td>
<td>null</td>
</tr>
<tr>
<td></td>
<td>Aesthetics</td>
<td>null</td>
</tr>
<tr>
<td></td>
<td>Rewards</td>
<td>null</td>
</tr>
<tr>
<td>(Denny, 2013)</td>
<td>Badges</td>
<td>+</td>
</tr>
<tr>
<td>(Cordova &amp; Lepper, 1996)</td>
<td>Context</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Familiarity</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Choice</td>
<td>+</td>
</tr>
</tbody>
</table>
(churner vs. non-churner). Players who canceled their subscription to the game or who haven’t played for two months prior to the analysis were considered churned. Two features sets were compared: theory-driven and data-driven. The first set consisted of four achievement and socialization oriented features, e.g., quest participation and monster skills, and the second consisted of 14 data-driven features, e.g., total session length, number of deaths, etc. A decision tree classifier was used for both feature sets. Results from 10-fold cross validation showed that data-driven features achieved marginally higher performance, in terms of the F-measure, at the cost of substantially greater model complexity compared to theory-driven features. Achievement-oriented features also had greater impact than socialization features. The two feature sets were also compared in terms of marketing potential via lift curve analysis. A lift curve measures the true positive rate as a function of number of samples that are positively identified, with the assumption that samples are sorted by classification score in descending order. It is used to tackle marketing questions like: how many mails should we send out to get a certain number of replies? Lift curve analysis in this paper showed that theory-driven features had greater lift than data-drive features when considering the top 25% of samples with the highest churn probabilities. The authors argued that this made the theory-driven feature set attractive for resource-limited situations.

Castro and Tsuzuki (2015) used time series and frequency analysis of player login records in another binary churn prediction task. Login records, each record consisting of login time and session duration, of players playing three different video games were collected. An asymmetric moving window was then convolved over each player’s sequence of login records to build a dataset of input features and binary output labels indicating churn/no-churn. For every window position, input features were collected from records on the left side, i.e., from the past $\Delta t_P$ days, with the corresponding output label set to 1 if the player played at least one game session during the next $\Delta t_F$ days (the right side). Four input feature representations were compared: Recency-Frequency-Monetary value (RFM), Time Domain (TD), Frequency Domain (FD) and Time Frequency Plane Domain (TFPD). RFM summarized player behavior with three components: time since last session and the frequency and the total duration of all sessions in the past $\Delta t_P$ days. TD binned login records into $n$ equally spaced time slices with each slice containing the total number of games initiated during that slice. Power spectrum analysis of this time series produced the FD representation with $log(n) + 1$ components containing the intensities of corresponding

$^1$Whereas ROC explicitly controls the threshold via false positive rate, the lift curve implicitly controls the threshold by changing the sample cutoff point.
frequencies. Finally, TFPD balanced the temporal and frequency resolutions of the TD representation by taking the intermediate $n$-component output of the Discrete Wavelet Transform (DWT). For each of the three games, the four representations were evaluated using a $k$-nearest neighbor classifier on five randomly generated training/testing splits. In all three games, the TFPD representation outperformed other representations in terms of the area under the receiver-operating characteristic (ROC) curve.

Another time series approach was proposed in unpublished work (Hui, 2013), where the author used a player-level model that predicted churn rates on social games using two publicly available daily aggregate usage statistics, daily active users (DAU) and monthly active users (MAU). Player-level modeling was used because the DAU and MAU measures mix repeat and new users so it was difficult to get an accurate and principled estimate of retention rates. The catch here was that there were no publicly available player-level datasets, e.g., data that tracked a player’s usage over time, with which to fit player-level models. Bayesian data augmentation was used to tackle this problem. First, a three-state hidden Markov model (HMM) was created to model a player’s game usage over time in days. There three states mapped to the player never playing the game (U), the player playing the game (A) and the player churning (D). The transition matrix was setup so that once the player churned, they stayed that way, i.e., D was an absorbing state. Each observation of the HMM indicated whether the player played the game on day $t$ or not, with the restriction that players only played, probabilistically, when they were in the A state. A small sample of $R$ players were then assumed to be responsible for generating the observed aggregate statistics. Since there were no observed player-level data, the hidden state and observation sequences of each player’s HMM were treated as model parameters which were sampled via MCMC methods (this was the Bayesian data augmentation step). Finally, a Gaussian observation model explained the observed aggregate statistics: $\log(y_t) = \log(\hat{y}_t) + \epsilon_t$ where $\hat{y}_t$ was the estimated aggregate statistic, DAU or MAU, given a sample of observation sequences from the $R$ players and $\epsilon_t \sim N(0, \sigma^2)$. Unsurprisingly, results showed that the expected 1-day retention rate was higher than than 7-day rate (59% vs 10.5%) across 379 games. More interestingly, the 1-day retention rate was used to identify design choices that would lead to higher retention. Specifically, since the set of games in the dataset differed in certain design choices, such as the incentive structure, the expected 1-day retention rates were regressed on three design variables: daily virtual rewards, daily time limits and virtual punishment for not playing. The findings suggested that offering daily virtual rewards and limiting play time increased
the 1-day retention rate whereas punishments did not.

Weber, John, Mateas, and Jhala (2011) defined engagement as player retention and proposed a technique to identify the most important features for predicting retention. Data from Madden NFL 2011 video game was used to train and evaluate regression models. Each player in the dataset was represented by a vector of normalized features that were based on in-game preferences, such as the control scheme, and behavioral characteristics, such as the player’s win ratio, with the number of games played as the dependent variable. To quantify the unique impact of each feature on a regression model, a function $g_k$ produced the output of the regression model with all feature values held constant, except the $k^{th}$ feature: $g_k(v, x) = f(\{x_j: x_j(1 - \delta_j=k) + \delta_j=k v\})$. In the linear regression case, $\frac{d}{dv} g_k(v, x) \propto \beta_k$ where $\beta_k$ is the coefficient of the $k^{th}$ feature. However, since the paper sought to marginalize over arbitrary regression models, this generalized formulation was appropriate. The impact $h_k$ of feature $k$ was then defined as the mean absolute deviation: $h_k = \int_0^1 |g_k(v, x) - \mu_k| dv$ with $\mu_k = \int_0^1 g_k(v, x) dv$. In other words, the impact would be high if $f(x)$ varied significantly as the $k^{th}$ feature changed value. The overall impact of each feature was calculated by summing the impact of the feature in each regression model. In this case, ZeroR, linear regression, decision trees and additive regression models were used. Two experiments were conducted: the first experiment analyzed the impact of each feature and the second determined the optimal win ratio in four game modes. Based on findings from these experiments, several suggestions were made to improve retention: simplifying and reducing the number of playbooks, presenting controls clearly and presenting appropriate challenge to the player, as the optimal win ratio was found to be dependent on game mode. Specifically, the optimal win ratio was around 50% for single and multiplayer modes and 80% for manager modes. These findings may be interpreted within the framework of illusion of control: in single and multiplayer models, the player had more control over the actual gameplay than in manager mode so a higher win ratio in manager mode provided an illusion of control that motivated players to continue playing.

Similarly to the previous paper, Tognetti, Garbarino, Bonarini, and Matteucci (2010) identified most important features for prediction but used preference learning and physiological player features instead. Here subjects played three difficulty levels of a car racing game. Each subject played six races such that they encountered each difficulty level twice. At the end of each race, starting from
the second, subjects were asked whether they preferred the game they just played to the previous one. Five physiological signals were collected from every subject during the last 60 seconds of each race and used to construct input feature vectors. Given these input features and preference labels, linear discriminant analysis (LDA) was used to infer feature weights that were consistent with subject preferences. Specifically, if a subject preferred race \( A \) to race \( B \) on the \( i \)th comparison task, then LDA tries to find a weight matrix \( W \) such that \( F_i^A W^T > F_i^B W^T \), where \( F_i^x \) is the feature vector associated with difficulty \( x \) in the \( i \)th comparison task. This can be formulated as a linear classification problem with inputs \( x_i = F_i^A - F_i^B \) and binary outputs \( c_i = 0, 1 \). The LDA model was evaluated with one physiological signal at a time, to determine the best input signal, using leave-one-subject-out cross validation. The results demonstrated that galvanic skin response (GSR) was the best predictive single feature. The authors noted that about 40% of the subjects consistently agreed on the order of preference of game variants, which meant that the difficulty levels that were matched to their skill levels. However, the remaining subjects were inconsistent which, the authors argued, highlighted the need to use objective biological signals to measure engagement instead of game design characteristics only (difficulty in this case).

Sifa, Bauckhage, and Drachen (2014) studied the distributions of total playtimes on a massive dataset comprising six million players, three thousand games and five billion hours. The dataset came from the Steam game hosting platform and contained total playtimes per (player, game). The authors cited earlier literature that showed that the Weibull distribution was appropriate for modeling total playtime datasets like Steam’s. The objective of the current paper was to identify prototypical playtime distributions and their relationship with different categories of games. First, a Weibull distribution was fitted to the play times histogram of each game, generating values for the Weibull’s shape and scale parameters. This effectively embedded all games into a two dimensional space. Then, archetypal analysis was used to identify a number of archetypes, extreme basis vectors, that characterize the embedding. This procedure factorizes a matrix \( X \in \mathbb{R}^{m \times n} \), a \( 2 \times n \) matrix of fitted Weibull parameters in this case, using \( k \ll n \) archetype vectors \( Z = [z_1, ..., z_k] \in \mathbb{R}^{k \times n} \) such that \( X \approx ZA = XBA \) where \( A \in \mathbb{R}^{k \times n} \) and \( B \in \mathbb{R}^{n \times k} \) are stochastic column matrices. Knowing \( A \) and \( B \), individual data points could be represented as convex combinations of archetypes, \( x_i \approx \sum_{j=1}^{k} z_j a_{ij} \), which are themselves convex combinations of data points, \( z_j = \sum_{i=1}^{n} x_i b_{ij} \). Computing archetypes involves computing the Gram matrix \( X^T X \) but
Table 1.2: Summary of player modeling papers.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Problem</th>
<th>Model Type</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Borbora et al., 2011)</td>
<td>Churn prediction</td>
<td>Decision tree</td>
<td>Theory driven vs. data driven features</td>
</tr>
<tr>
<td>(Castro &amp; Tsuzuki, 2015)</td>
<td>Churn prediction</td>
<td>DWT</td>
<td>Recency-Frequency-Monetary value (RFM)</td>
</tr>
<tr>
<td>(Hui, 2013)</td>
<td>Churn prediction</td>
<td>HMM, Data Augmentation</td>
<td>Time-Frequency</td>
</tr>
<tr>
<td>(Weber et al., 2011)</td>
<td>Retention</td>
<td>Linear, Decision Tree, Additive Tree</td>
<td>Game design variables</td>
</tr>
<tr>
<td>(Tognetti et al., 2010)</td>
<td>Preferences</td>
<td>LDA</td>
<td>Game-specific features</td>
</tr>
<tr>
<td>(Sifa et al., 2014)</td>
<td>Total play time</td>
<td>Kernel Archetypal Analysis</td>
<td>Play time distributions by genre</td>
</tr>
</tbody>
</table>

the authors argued that euclidean distance metrics were not appropriate for comparing parameters of the Weibull distribution so they used the kernel trick to replace \(X^TX\) with \(K(x_i, x_j)\), where \(K\) is the Kullback divergence. Archetypal analysis was applied with \(k = 4\) archetypes to analyze (a) the four resulting prototypical Weibull distribution and (b) the membership of games to archetypes. Three of the archetypes generated distributions that peaked immediately at one hour of play time then fell away afterwards at different rates. One archetype produced a distribution that rose gradually to peak at 4 hours of play time then fell rapidly afterwards. This exceptional archetype was determined to be mostly associated with free to play strategy and role playing games. This result is somewhat counter intuitive as one would expect players to have a higher likelihood of stopping if the game is free. On the other hand, strategy games may require players to invest significant time initially to understand the dynamics of the game before the player can make an informed choice about the enjoyability of a game. To summarize, the paper found that the four different archetypes yielded different play time distributions with varying decay rates and mean values and that those archetypes could be mapped to game categories through the membership matrix \(A\) used in archetypal analysis.

1.3 Optimization

Design optimization is the problem of selecting the set of game parameters, e.g., dynamics or aesthetics, that maximizes some measure, such as player engagement. Work on game design optimization can be categorized into two classes: static and dynamic optimization. Static optimization approaches select optimal game parameters a priori; for example, the optimal game design may be determined via simulations or human trials before the game is deployed in the real world. Dynamic optimization approaches are usually framed as dynamic difficulty adjustment (DDA) algorithms which try to adjust game pa-
rameters, such as AI and procedural level design, to match the skill of the player. The two approaches are orthogonal and have been shown to work together in the context of a categorization task (Lindsey, Mozer, Huggins, & Pashler, 2013).

1.3.1 Static Optimization

Rafferty, Zaharia, Griffiths, et al. (2012) presented a Bayesian optimization approach to game design. The objective was to select game designs that maximized information about the parameters of an underlying cognitive model. Specifically, the authors wanted to select game design $\xi$ that maximized the utility

$$U(\xi) = E_{p(\theta,y)} [I(y;\theta|\xi)]$$

where $\theta$ was the set of parameters of the cognitive model, $y$ were the observations and $I(y;\theta|\xi)$ was the mutual information between $y$ and $\theta$. Player actions $a$ were assumed to be generated according to a Markov decision process (MDP) that was parameterized by $\theta$, $p(a|MDP_\theta, \xi)$. Given this likelihood model and a prior on $\theta$, expected mutual information could then be calculated for a game design $\xi$. The proposed design optimization scheme was evaluated within the context of a Boolean concept learning game. In this game, players gained points if they correctly selected symbols that belong to the true hidden category and lose points otherwise. The game design space $\xi$ consisted of reward and punishment points, energy loss rate, symbols shown and true underlying category. The objective was to estimate the relative difficulty of learning each of the six categories, so $\theta$ here specified a categorical distribution over six categories. State, action, transition and reward components of the MDP were defined for the concept learning game. The MDP was used as a proxy for human players to search for the game design with the maximum information gain. Note that the optimization process was not sequential; data points from the player playing a previous game design were not taken into account when evaluating the next candidate design. Having found an optimal game design, human subjects were assigned randomly to two conditions: the optimized game and a random game. The cognitive model was then fitted to the data from these two conditions to estimate the posterior $p(\theta|y,\xi)$. The results showed that the posteriors from the optimized game appeared more concentrated with an information gain in the optimized condition of 3.3 vs 1.62 bits in the random condition. The results also largely matched those from the authors’ earlier work on Boolean concept learning. A second human study used several game designs with various predicted information gains to confirm a positive correlation between
predicted and actual information gains. This lends support to the predictive power of this generic model and its possible use in other games and domains.

Liu, Mandel, Brunskill, and Popovic (2014) used a greedy search algorithm to identify the most influential design factors of an educational game. The objective of the algorithm was to find an optimum ordering, according to some performance measure, over a set of parameters. Given a set of fixed parameters $P$ and free parameters $E$, the algorithm computed the best setting of each parameter $p_i \in E$, marginalizing over the other parameters $p_j \in E \setminus p_i$. The algorithm then picked the parameter with the best performance, fixed its value to the best setting, so $P_{\text{new}} = P \cup \{p_{\text{best}}\}$, and recursed on the remaining free parameters. Evaluation was on a dataset from an educational number line game. Number lines test the student’s ability to estimate the value of a fraction, by pointing to the correct location on a straight line. Each number line was parameterized by four parameters with discrete values, creating a design space with $2 \times 2 \times 4 \times 4 = 64$ possible combinations. The dataset from the game consisted of 361,738 pairs of (experimental number line, test number line). The term “experimental number line” was used to refer to a particular game design or condition that was being evaluated whilst the test number line was another, independently chosen design that was used to generate the outcome measure. Two outcome measures were used in this paper: correctness, 1 if the player solved the test number line correctly on the first try, and persistence, 1 if the player answered the test number line eventually. Running the algorithm on the dataset showed two different parameter orderings and optimal settings, depending on the outcome measure, e.g., correctness was most affected by the representation of the fraction (symbolic or pie) followed by the strength of hints whereas persistence was most affected by the presence of animations followed by the representation of the fraction. Since there were only 64 parameter combinations, the authors compared the optimum found from the greedy search to the global optimum and found them to be comparable in terms correctness and persistence scores. A validation set was then used to test the hypothesis that parameters chosen by the algorithm produced better correctness/persistence than other experiment designs. Here the authors progressively set parameter values according to the order found by the algorithm and evaluated performance compared to all other designs at each step. Findings showed that only the two most influential parameters identified by the algorithm produced a meaningful increase in correctness/persistence. In addition to this positive result, the authors argued that the

\footnote{Not really independent. First and second number lines share all parameters, as do the third and fourth and so on. The authors correct for this via importance sampling.}
algorithm found an unexpectedly good design, one that used pie representation, which allowed them to evaluate it and prove its worth.

Zook, Fruchter, and Riedl (2014) used a Gaussian process-based methodology for efficiently finding optimal game-designs. The notion of optimality was defined in two ways: matching a specific difficulty level and preferring a particular game design over another. Finding a game that matches a specific difficulty and learning user preferences corresponds to regression and classification problems, respectively. The game in question placed the player in control of a battleship and required her to destroy enemy battleships whilst dodging their fire. Different sets of design variables, e.g., rate of fire, thrust sensitivity, bullet size, etc., were manipulated in the regression and classification cases where the objectives were to hit an enemy six times during a game and to learn latent function values $f(x)$ that were consistent with user preferences, respectively. A Gaussian process (GP) model was used in both cases with various acquisition functions that have different biases towards exploration/exploitation. Acquisition functions use the predictions of a trained model, the GP in this case, to select the next game design to evaluate. Depending on the nature of the function being optimized, different acquisition functions may result in faster or slower convergence towards the optimum. To evaluate the performance of these acquisition functions, a study was conducted where players played the game with uniformly-sampled values for the design variables. Each player played 10 waves where each wave was a distinct game design. In the classification case, players were asked whether they liked the current wave better than the previous one. The data from the random assignment study was then used to run simulated active selection studies. Regression results showed that acquisition functions that balance exploration and exploitation, e.g., upper confidence bound and expected improvement, achieved a smaller mean squared error on test earlier than other methods. Classification results showed that query-by-bagging, which selected the game design that maximized disagreement between randomly-trained copies of the model, achieved best results in terms of the F1 score$^3$. In both regression and classification, the acquisition functions outperformed random sampling.

$^3$The F1 score is the harmonic mean of precision and recall
### 1.3.2 Dynamic Optimization

Dynamic optimization algorithms attempt to balance the difficulty of the game to match the player’s skill. Common approaches here typically use ad-hoc algorithms to generate difficulty settings that match the skill of the player. For example, the algorithm by Harrison and Roberts (2013) forced an AI agent in scrabble to move into a game state that promoted player retention. Rather than focusing on ad-hoc approaches, the objective in this section is on papers that propose principled universal approaches to the problem of dynamic difficulty adjustment.

Two principled Bayesian criteria were proposed to adapt the difficulty of exercises, without imposing ad-hoc requirements such as maintaining a 75% success rate (Kujala, Richardson, & Lyytinen, 2010). The criteria relied on a student model that provided the likelihood of a response $r_x$ given some parameters $\theta$, $P(r_x = 1|\theta)$, where $x$ is the content associated with trial $i$, e.g., the number of choices in a multiple choice questions. The student model used in the paper was a student-level variant of the Rasch model with a student skill parameter and a set of category difficulty parameters. Based on this model, the efficiency-motivated criterion, MaxInfo, maximizes mutual information per trial, $I(R_x; \Theta|y) = H(\Theta|y) - H(\Theta|R_x, y)$, whilst the learner-friendly criterion maximizes mutual information per mistake $\frac{I(R_x; \Theta|y)}{E(R_x=0|y)}$, where $y$ corresponds to past observations. Mutual information is the expected reduction in the uncertainty associated with the model parameters $\Theta$, after observing the trial outcome $R_x$. The authors argued that mutual information naturally balanced difficulty because it avoided exercises that are too easy or too difficult, to the extent that the student model could predict those exercises with high certainty. Furthermore, they also noted that the first criterion may choose exercises that were too difficult, demotivating the students, whereas the second tried to balance trial difficulty and information content. Unfortunately, the paper did not explicitly test the hypothesis that the learner-friendly criterion resulted in greater engagement or learning outcomes. Instead, the authors (i) compared the distributions of success rates generated by simulations of MaxInfo and learner-friendly criterion and (ii) validated distributional assumptions on model priors, $p(\theta)$, by comparing simulation runs with pilot data of 28 students using a tutoring system. In all experiments, it was assumed that exercises were divided into $k$ categories, *with known difficulties*, and that exercises within the same category were equally difficult. Students practiced in sessions of about 35 exercises each. A session-level HMM
was used to account for the fact that model parameters $\theta$ may change between sessions, i.e., a transition model $p(\theta' | \theta)$ was assumed between sessions. The simulations showed that the learner-friendly criterion generated a distribution of success rates centered on 80%, a rate conventionally assumed to enable good learning, whilst the MaxInfo criterion generated a distribution centered on 65%. The pilot experiment was conducted with the learner-friendly criterion and produced a success rate distribution that was very similar to that from the learner-friendly simulations. The overall results, the authors argued, showed that the mathematically-motivated learner-friendly criterion produced a conventionally-accepted rate of success without any explicit requirements or constraints, making it generalizable to other domains.

Missura and Gärtner (2011) presented another principled universal dynamic difficulty adjustment algorithm. The authors assumed the existence of a known partially ordered finite set of difficulty settings to choose from $(K, \succ)$, where $\succ$ denotes a “more difficult than” relation. Associated with each difficulty setting was a hidden preferred difficulty level which took one of three values: too easy, just right or too difficult. The algorithm tried to pick difficulty settings which were “just right”, as often as possible. For every difficulty setting $k$, the algorithm associated a belief $w(k) \in (0, 1)$ that the difficulty setting was just right. On each round, the algorithm computed for every difficulty setting the belief that the preferred difficulty setting should be harder, $A(k)$, or easier, $B(k)$. The algorithm then picked the difficulty setting that had greatest smallest belief of $\{A(k), B(k)\}$. The algorithm then observed feedback which could be -1, 0 or +1 corresponding to too hard, just right and too easy, respectively. Based on this feedback, the learning rate parameter and the structure of difficulty settings, the algorithm then updated all beliefs $w(k)$ and started the next round. This algorithm was proven to have a bound on the number of incorrect difficulty settings chosen relative to the best static difficulty setting chosen in hindsight (BSIH).

Evaluation of the algorithm pitted it against two types of “adversaries”: a stochastic player and an evil player. Each adversary was assumed to have a set of difficulty settings which were “just right”, called the zero-zone, which either changed smoothly or non-smoothly. Results of simulations with the stochastic player that changed the zero-zone smoothly showed that the proposed algorithm outperformed a state-of-the-art algorithm and BSIH in terms of loss which was impressive, especially as BSIH assumed access to the future. On the other hand, when the zero-zone changed non-smoothly all algorithms performed poorly. For the evil adversary, the authors ran a human experiment where humans were tasked with the manipulation of the zero-zone either in a smooth or non-smooth way. Results showed that the proposed
algorithm performed similarly to BSIH in both smooth and non-smooth cases.

Another, somewhat less principled, approach based on Reinforcement learning (RL) was proposed (Andrade, Ramalho, Santana, & Corruble, 2005). RL is a well-known approach for deploying artificial agents that play optimally in game but the authors here used it to develop agents that match the skill of the player. They considered a street fighting video game where two players must attack and dodge each other until either one of them runs out of life points. The authors noted that one could construct a RL algorithm with a reward model that corresponded to the intended objective, playing at the human player’s level, but they argued that such an approach would result in an agent that was incapable of fighting immediately against experts and whose behavior was non-believable. Instead, the proposed approach was to pre-train the RL agent to master the game, using simulations against other agents, and then let the agent play against humans using a heuristic action-selection procedure. Pre-training enabled the agent to quickly adapt to the skill of a human player, regardless of skill. Heuristic action-selection facilitated difficulty adaptation by progressively selecting optimal or sub-optimal actions, depending on the estimated difficulty level of the game, which was available through a challenge function, e.g., the rate of loss of life points. For example, if there were 13 actions to choose from, the trained RL agent would start the game selecting the 6th optimal action and if the difficulty was estimated to be too easy, the agent would select the 5th optimal action and so on. The agent continued to learn how to play well during the game with the only difference being the heuristic action selection procedure. This approach was evaluated via simulations pitting different types of agents (random, state-machine, RL and adaptive RL) against each other. Performance was measured as the difference in life points at the end of each fight (each agent played 30 fights against each opponent). Results showed that the adaptive RL agent was able to match the skill of every other agent by ending fights with small differences in life points. Furthermore, the distribution of actions selected by the adaptive RL agent corresponded to the skill of the opponent with the agent selecting near-optimal actions when playing against the traditional RL agent and suboptimal actions when playing against a random agent.
<table>
<thead>
<tr>
<th>Paper</th>
<th>Problem</th>
<th>Outcome Measure</th>
<th>Model</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Rafferty et al., 2012)</td>
<td>Static</td>
<td>Model Uncertainty</td>
<td>MDP + Mutual Info</td>
<td>Optimized &gt; Random</td>
</tr>
<tr>
<td>(Liu et al., 2014)</td>
<td>Static</td>
<td>Persistence/Correctness</td>
<td>Ad-hoc</td>
<td>Optimized &gt; Random</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Optimized ≈ Global</td>
</tr>
<tr>
<td>(Zook et al., 2014)</td>
<td>Static</td>
<td>Questionnaire Scores</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Kujala et al., 2010)</td>
<td>DDA</td>
<td>Difficulty/Preferences</td>
<td>GP</td>
<td>Optimized &gt; Random</td>
</tr>
<tr>
<td>(Missura &amp; Gartner, 2011)</td>
<td>DDA</td>
<td>Success rate distrib.</td>
<td>Rasch + Mutual Info</td>
<td>Distrib. matches empirical</td>
</tr>
<tr>
<td>(Andrade et al., 2005)</td>
<td>DDA</td>
<td>Loss</td>
<td>Exponential Updating</td>
<td>Performs similarly to BSH</td>
</tr>
<tr>
<td></td>
<td>DDA</td>
<td>Skill match</td>
<td>RL + Heuristic Actions</td>
<td>Matches skill of any agent</td>
</tr>
</tbody>
</table>

Table 1.3: Summary of the optimization papers.

1.4 Limitations

There are two major shortcomings of existing work on video game engagement. First, most of the studies use an A/B methodology that does not explore the full continuum of design manipulations. Second, interesting subtle manipulations such as the perception of difficulty, near-win, and tension-and-release have not received sufficient attention. While there is a lot of work on how manipulating the overt difficulty of the video games affects engagement, little investigation has gone into manipulating the perception of difficulty. A recent study has looked at this (Denisova & Cairns, 2015) but it only considered two conditions – manipulated vs. unmanipulated – and did not consider how strongly perception should be manipulated. The tension-and-release manipulation is a well-known principle in music composition, which seeks to build and release tension in listeners via changes in pitches or layering of instruments. In video games, tension-and-release is understood qualitatively but, like with the perception of difficulty, the majority of work is focused on difficulty balancing – like DDA – rather than the experience of the player at a given difficulty level, which is what tension-and-release targets. The near-win effect from gambling psychology has been shown to maximize a behavioral measure of engagement – the voluntary time on activity. However, not much work has been done on incorporating it into video games.

In the next three chapters, I address these two shortcomings by studying each of the three manipulations using two non-parametric modeling tools: Gaussian processes and neural networks. These tools enable experimenters to (a) run few subjects in many conditions and still obtain reliable results due to nearby conditions mutually constraining the model and (b) impose only weak prior assumptions on the structure of the model, unlike linear or generalized linear models. Relevant research into video games using each manipulation is discussed in corresponding chapters.
Chapter 2

Designing Engaging Games Using Bayesian Optimization

2.1 Introduction

Interest has recently surged in applying game-like mechanics to enhance engagement in a variety of domains, such as personal health (Fitocracy, 2015; Jurgens, McCorriston, & Ruths, 2015), scientific discovery (Khatib et al., 2011; Cooper et al., 2010), and education (de Sousa Borges, Durelli, Reis, & Isotani, 2014; Lomas, Patel, Forlizzi, & Koedinger, 2013a; Lomas, 2014; Liu et al., 2014). This research is based on the hypothesis that increased engagement will improve user experiences, data collection, and outcomes. Although engagement is a broad construct (Fredricks, Blumenfeld, & Paris, 2004), it has been operationalized via subjective-self reports, physiological measures, player preferences, and observations of in-game player behavior (Mekler, Bopp, Tuch, & Opwis, 2014). In the present work we measure engagement using player persistence, which has been explored previously in game engagement research (Lomas et al., 2013a; Weber et al., 2011) and in the gambling psychology literature (Kassinove & Schare, 2001), as well as using projections of other players’ persistence and post-experiment subjective surveys engagement.

1This chapter is an expanded version of (Khajah, Roads, Lindsey, Liu, & Mozer, 2016), which was written in collaboration with Brett D. Roads, Robert V. Lindsey, Yun-En Liu, and Michael C. Mozer.
In gaming, as in related domains, a key design decision that affects engagement is how difficult to make challenges presented to users. If challenges are trivial, users become bored and lose interest; if challenges are overwhelming and utterly impossible, users quit from frustration. Successful design identifies the not-too-easy, not-too-hard challenge level that seduces users. We focus on manipulations of difficulty to modulate engagement in this work, although the methods we present are suitable for exploring any aspect of game design to achieve any measurable outcome.

Difficulty manipulations can be static or dynamic. Static manipulations modulate initial game configuration and design before game play begins based on population-level play-testing data or simulations of player behavior. Static manipulations have been used to match a particular game statistic (Zook et al., 2014; Isaksen, Gopstein, & Nealen, 2015), reduce uncertainty about an underlying cognitive model (Rafferty et al., 2012), and maximize success or persistence rates in an educational game (Liu et al., 2014). In contrast, dynamic manipulations modulate game design parameters on the fly in reaction to player behavior or performance; examples include matching simulated players’ skill (Andrade et al., 2005), selecting exercises that maximize information per mistake (Kujala et al., 2010), adapting AI to move into game states that maximize persistence (Harrison & Roberts, 2013), and adjusting the user interface to maximize performance (Mahmud, Rosman, Ramamoorthy, & Kohli, 2014). Static and dynamic difficulty manipulations can be combined to discover a dynamic difficulty adjustment (DDA) scheme that works across individuals (Lindsey et al., 2013). The dynamic difficulty manipulation is parameterized and the parameters are determined to be appropriate for a population of users. In this work we focus exclusively on static difficulty manipulations but our method applies to any parametrized measure of difficulty, including DDA.

Game difficulty parameters are typically based on the player’s physical or mental capabilities (e.g., the player’s reflexes and perceptual ability determine how small or fast enemies can be before the player is overwhelmed), but the player’s perception of difficulty is also important. For instance, players are more engaged playing a game in which suspenseful audio messages warn of potential enemies than an identical game without the messages (Klimmt, Rizzo, et al., 2009). Players are also more engaged when they told that the game has adaptive AI, when it actually does not (Denisova & Cairns, 2015). In these two examples, the player’s perception of difficulty was manipulated, through suspense and pre-game
instructions, whilst the actual game difficulty was held constant. In this work we introduce a novel twist in which we manipulate the actual difficulty whilst holding the perception of difficulty constant.

### 2.1.1 Recent Research on Game Optimization

Recent research has used an on-line educational gaming platform to search a design space to maximize player retention. The platform, called BrainPop, is a popular site used primarily in grade 4-8 classrooms. It offers multiple games, and students can switch among the games. Usage is divided into sessions, and engagement is measured by the length of a session and the number of rounds played within a session. Lomas et al. (Lomas et al., 2013a) conducted randomized controlled trials on four dimensions affecting game difficulty, the Cartesian product of which had $2 \times 8 \times 9 \times 4 = 576$ designs. Each of 69,642 anonymous user sessions were randomly assigned to a design, statistical hypothesis testing showed that less challenging designs were more engaging.

As an alternative to exhaustive search through design space, Liu et al. (Liu et al., 2014) devised a heuristic, greedy search strategy that involved selecting one dimension at a time, marginalizing over the as-yet-unselected dimensions. This strategy was used to identify the design maximizing user persistence in a five-dimensional space with 64 designs; we will return to this experiment shortly. Lomas (Lomas, 2014) used multi-armed bandits to efficiently search a design space and minimize regret—defined as games that users chose not to play. In experiments with relatively few distinct designs (5 or 6), more games are played overall with bandit assignment of designs than with random assignment.

The three search strategies just described—exhaustive, greedy, and bandit-based—deal adequately with nominal (categorical) dimensions but are not designed to exploit ordinal (ranked) or cardinal (numerical) dimensions. Further, the exhaustive and bandit strategies cannot leverage structure in the design space unless they make the strong and unreasonable assumption that choices on the dimensions are independent.
2.2 Gaussian Processes and Bayesian Optimization

We propose an alternative methodology to search for engagement-maximizing designs: Bayesian optimization (BO). To motivate this methodology, suppose one wishes choose a font size for a web site to maximize the duration that visitors stay on the site. We might posit a quadratic model to formalize the relationship between font size, denoted \( x \), and stay duration, denoted \( y \): 

\[
y(x) = \beta_0 + \beta_1 x + \beta_2 x^2,
\]

where the coefficients \( \beta \equiv \{\beta_0, \beta_1, \beta_2\} \) are unknown. If we randomly assign visitors to conditions, as in A-B testing, we will collect many noisy \((x, y)\) observations. We can fit the \( \beta \) parameters to the observations and use the resulting parameterized function, \( y(x) \) to identify the font size \( x \) that maximizes stay duration \( y \). The function serves as a surrogate for the true implicit function that describes reality.

With sufficient data, the surrogate will be a good approximation to the true implicit function. In this approach, although each data point is noisy, each data point constrains the overall shape of the function; in concert, a relatively small amount of noisy data can serve to identify the function optimum. This benefit arises because the values of \( x \) define a continuum.

BO extends this simple method in three respects. First, instead of using a parametric model—a model of fixed, prespecified form—that makes strong assumptions about the relationship between dependent and independent variables, BO uses a more powerful class of nonparametric surrogate models whose only constraint is that the function \( y(x) \) must be locally smooth. Consequently, BO can infer arbitrary sorts of structure in the design space. Second, BO is a Bayesian methodology: instead of searching for the best fitting parameter values, referred to as a maximum likelihood estimate, BO computes the posterior distribution over parameter values. The posterior effectively allows BO to assess the uncertainty in its predictions. Third, instead of choosing random \( x \) for testing, BO uses active-selection heuristics to select \( x \) in order to be data efficient. These heuristics trade off exploration and exploitation, that is, trade off testing regions of the design space where parameter values are uncertain versus those where optima are likely to be given the data previously collected. These heuristics leverage the Bayesian representation of uncertainty.

BO typically assumes a prior probability distribution over all possible smooth functions, \( f(x) \). (This is a generalization of the notion of assuming a prior distribution over the parameters \( \beta \).) This prior is known as a Gaussian process (GP) prior. The term Gaussian comes from the assumption that sets of
points on the function are jointly Gaussian. Rather than assuming a certain degree of smoothness of the function, GPs are nonparametric: the data specify the degree of smoothness. GPs can model ordinal and cardinal dimensions to discover functional relationships between designs and outcomes. GPs are also efficient in their use of data (Snoek, Larochelle, & Adams, 2012), leading to strong predictions with orders of magnitude less data than utilized by previously tested methods. This efficiency arises from the underlying assumption of smoothness, i.e., nearby points in the design space yield similar degrees of engagement. By contrast, the common multi-armed bandits approach assumes that each arm, or design, is independent which prevents the method from exploiting observations from similar designs.

Technically, a GP is a probability distribution over functions which is characterized by a mean function \( m(x) \) and a covariance function \( k(x, x') \) where \( x \in \mathbb{R}^d \) is the input. The mean function specifies a prior on the trend, the overall shape of the function, and is typically set to a constant. The covariance function or kernel specifies how a change of the function value at input \( x \) affects the function value at input \( x' \) on average. Linear regression is a special case of a GP where the kernel is linear \( k(x, x') = xx' \).

In this thesis we use the squared exponential kernel (SE), \( k(x, x') = \sigma_0^2 \exp \left( -\frac{1}{2} (x - x')^T \Sigma^{-1} (x - x') \right) \) which is a common default choice for GPs. The variance parameter, \( \sigma_0 \), specifies the amplitude of the function and the \( d \times d \) diagonal matrix \( \Sigma \) specifies length scale of each dimension in the input. The length scale characterizes how quickly the function changes so a large length scale generates functions that vary slowly while a small length scale generates functions that change rapidly.

To illustrate the power of GPs, consider our previous hypothetical example of a web designer trying to optimize font size to maximize stay duration. Suppose that we have already collected observations from several conditions – stay durations – and now we wish to analyze them. The standard approach, the model-free approach, assumes that stay durations at different font sizes are independent, making it necessary to run many subjects at each condition in order to achieve statistical reliability (left plate in Figure 2.1). An alternate model-driven approach is to analyze the results using a statistical model (middle plate in Figure 2.1). Notice how, for the same dataset, the confidence interval of the mean (the shaded area) is much smaller compared to the intervals in the leftmost plate. This is because observations at nearby conditions mutually constrain the model at those conditions; the fitting procedure uses observations at condition A to constrain the function value at condition B and vice-versa. As we
Figure 2.1: The model-free vs. model-based approach. Left plate shows a hypothetical dataset (orange circles), the sample mean (blue circles), and 95% confidence intervals (bars). The middle plate shows a linear model fit to the same dataset, with the blue curve and shaded area representing the predicted mean and 95% confidence interval of the mean, respectively. The right plate shows a Gaussian process model fit.

stated earlier, the downside of this approach is the strong set of assumptions it imposes on the generative process which, if violated, could lead to incorrect conclusions. For example, our hypothetical dataset was generated from quadratic function with a maximum around the middle of the input space but our linear model incorrectly suggests that the maximum is at the edge of input space. Of course, we can fix this by increasing the complexity of the terms in our linear model, by adding second- and third-order terms for example, but this approach is prone to over-fitting and suffers from the same problem of strong prior assumptions. In the rightmost plate of Figure 2.1 we fit a GP with a SE kernel to our hypothetical dataset. The resulting model predictions are superior to the linear model and closely match the underlying generative process we used to produce the dataset. Note how the model predictions in the middle region of the input space have greater variance due to lack of observations. In contrast, the linear model is over-confident in the middle region because it assumes a particular generative process.

BO can be used with models other than GPs. For example, BO was recently used to adaptively select control dynamics that maximize a user’s in-game performance (Mahmud et al., 2014). Here the user is assumed to behave according to a Markov decision process (MDP). This approach outperformed the traditional multi-armed bandits approach. In our case however, we want to adjust game designs statically over a population of users, which makes GPs a natural choice. As we noted earlier, it is trivial to combine static and dynamic manipulations, e.g., BO could optimize the discount parameter of the MDP.
In our context, Bayesian optimization with GPs infers a surrogate function that characterizes the relationship between designs and a latent valuation. Each design is a parameterization of a game, and the valuation is our measure of engagement. Starting with a Gaussian process (GP) prior and observations of human behavior, the optimization procedure computes a posterior over functions and uses this posterior to guide subsequent experimentation. With a suitable exploration strategy, globally optimal solutions can be obtained.

Bayesian optimization with GPs has recently been applied to the design of a shoot-'em-up game. Zook et al. (Zook et al., 2014) searched over several design parameters to achieve a gameplay objective: having the enemy hit the player exactly six times during an attack. Optimizing gameplay is different than optimizing engagement in one critical regard: the observation model required. The observation model is a probabilistic mapping from the latent valuation represented by the GP to observed behavior (called a likelihood in the general GP literature). Because engagement is a characteristic of the player’s cognitive state, the observation model is a cognitive theory of how the state of engagement induced by a given game design influences behavior. Similar probabilistic models have been developed for a variety of human responses, e.g., preference (Chu & Ghahramani, 2005a), two-alternative forced choice with guessing (Lindsey et al., 2013), and similarity judgment (Roads & Mozer, 2015). Here, we develop and justify a probabilistic model to predict behavioral measures of engagement from the latent index of engagement.

2.2.1 An Illustration of the Bayesian Approach

In this section, we re-analyze an existing data set and show the value of Bayesian methods. The data set is from Liu et al. (Liu et al., 2014), who constructed a game called Treefrog Treasure to teach fractions. In this game, the player guides a frog to jump to a series of targets which are specified as fractions on a number line. The game can be configured in one of 64 designs, specified in a discrete $2 \times 2 \times 2 \times 4$ space. The dimensions determine the representation of the target and the number line (pie chart or symbolic), presence/absence of tick marks and animations, and the number of hints provided (1-4). Over 360,000 trials were collected from 34,000 players with design changing randomly every other trial. Players could quit the game on any trial at their discretion. Engagement is quantified by the probability that, for a
trial of design $A$, a player will complete the next trial (and not quit). We call this the persistence induced by $A$.

We use the data resampling and aggregation procedure of Liu et al. to marginalize over two irrelevant aspects of the data—the design of the next trial and the specific fractions tested. Figure 2.2a shows the empirical persistence across designs, and Figure 2.2b shows the same result but smoothed via a GP classifier. The model provides a clear interpretation of which design dimensions matter, in contrast to the raw data. In support of the robustness of the model, it produces the same interpretation across regroupings of the data. Further, the model produces a prediction of engagement over the design space that is consistent with that obtained by the approach of Liu et al. (Liu et al., 2014), which they validated on a test set. For example, persistence is higher without animations (the bottom row of cubes). Animations provide a visual tutorial in dividing up number lines into fractions, and might make problems easier; however, they also take control away from the player for several seconds and could therefore be distracting. These results suggest that the distraction effect overpowers any possible learning gains, underscoring the importance of engagement in any optimization process for online games.

This simulation used a logistic observation model—yielding observations in $[0,1]$—and a squared exponential kernel with ARD distance measure, for a total of 6 hyperparameters which were drawn via elliptical slice sampling. This kernel effectively computes a weighted Hamming distance on the binary dimensions.
2.2.2 From Persistence to Total Play Duration

The logistic observation model is a natural choice to characterize persistence on a single trial. However, this model assumes that after each trial, the player flips a biased coin to decide whether to continue. Because the coin flips after each trial are independent of one another, the model predicts an exponential distribution for total play duration.

The exponential distribution is not a particularly realistic characterization of human activity times. The best studied measure of time in human behavior is the response latency, which has been characterized by positively skewed distributions in which the variance grows with the mean, e.g., an ex-Gaussian (Hohle, 1965) or Weibull density (Rouder, Lu, Speckman, Sun, & Jiang, 2005). Evidence about usage-duration distributions is harder to find. Miyamoto et al. (Miyamoto et al., 2015) performed an analysis of 20 MOOCs and found a positively skewed distribution for both the number of sessions and hours a student would engage with a course. Andersen et al. (Andersen, Liu, Snider, Szeto, & Popovic, 2011) also observed what appears to be a mixture of positively skewed distribution and an impulse near 0 representing individuals who lost interest immediately.

In order for Bayesian optimization to produce sensible results, we require an observation model that represents the mapping from latent states of engagement to a play duration. In the next section, we propose four alternative observation models that seem well matched to empirical distributions. We evaluate these models via simulation experiments.

2.2.3 Selecting an Observation Model

Our goal is to identify a model that is robust to misspecification: we would like the model to work well even if real-world data—engagement as measured by the duration of play—are not distributed according to the model’s assumptions. The observation models must have three properties to be suitable for representing play-duration distributions: (1) nonnegative support, (2) variance that increases with the mean, and (3) probability mass at zero to represent individuals who express no interest in voluntary play. To satisfy these three properties, our generative process assumes that play duration, denoted $V$, is
Figure 2.3: Simulation experiment. (a) Examples of the 2D functions used for generating synthetic data. (b) Results of synthetic experiment. The left and right plots depict the mean function value (higher is better) and the mean distance to the true optimum (lower is better) for various observation models. Results are averaged over four different generative-process models, 100 replications of each simulation, and the last 10 trials per replication. Error bars indicate ±1 standard error.

given by $V = CT$, where

$$C|\pi \sim \text{Bernoulli}(\pi)$$

is an individual’s binary choice to continue playing or not and $T$ is the duration of play if they continue.

Criterion 1 rules out the popular ex-Gaussian density because it has nonzero probability for negative values. We tested four alternative distributional assumptions for $T$:

$$T \sim \text{Gamma} \left(\alpha, \frac{\alpha}{CT+\pi} \right)$$

$$T \sim \text{Weibull} \left( k, \frac{e^{f(x)}}{T(1+\frac{1}{x})} \right)$$

$$T \sim \text{ln} \mathcal{N} \left( f(x) - \frac{\sigma^2}{2}, \sigma^2 \right)$$

$$T \sim \text{Wald} \left( \lambda, e^{f(x)} \right)$$

where $x$ is a game design and $f(x)$ is the latent valuation and has a GP prior. The first parameter of the Gamma, Weibull, and Wald distributions specify the shape, and the second parameter specifies the rate, scale, and mean, respectively. The two parameters of the log-Normal distribution specify the mean and variance, respectively. These four distributions all share the same mean, $e^{f(x)}$, but have different higher-order moments. Note that the Gamma distribution includes the exponential as a special case.

To allow a design’s valuation $f(x)$ to influence the choice $C$ as well as the play duration $T$, we define logit$(\pi) \equiv \beta_0 + \beta_1 f(x)$. This general form includes design invariance as a special case ($\beta_1 = 0$).

We performed synthetic experiments with each of these four observation models. To evaluate
robustness to misspecification, we evaluated each model using the same four models to simulate the underlying generative process (i.e., to generate synthetic data meant to represent human play durations). Synthetic data for these experiments were obtained by probing a valuation function, \( f(x) \), that represents the engagement associated with a design \( x \). For \( f(x) \), we used a mixture of two to four Gaussians with randomly drawn centers, spreads, and mixture coefficients, defined over a 2D design space. For examples, see Figure 2.3a. We generate synthetic observations by mapping the function value through the assumed generative process. The goal of Bayesian optimization is to recover the function optimum from synthetic data. We performed 100 replications of the simulated experiment, each with a different randomly drawn mixture of Gaussians and with \( \beta_0 = 0 \) and \( \beta_1 = 1 \). For the generative models, we need to assume values for the free parameters, and we used \( \alpha = 2 \), \( k = 2 \), \( \sigma^2 = 1 \) and \( \lambda = 4 \). (These parameters settings are used to generate the synthetic data and are not shared with the Bayesian optimization method; rather, the method must recover these parameters from the synthetic data.)

To perform Bayesian optimization, we require an active-selection policy that determines where in design space to probe next. The probability of improvement and expected improvement policies are popular heuristics in the Bayesian optimization literature. Both policies balance exploration and exploitation without additional tuning parameters. However, since the variance increases with the mean in our observation models, both policies tend to degenerate to pure exploitation. Instead, we chose Thompson sampling (Chapelle & Li, 2011), which is not susceptible to this degeneracy. For each replication of the simulated experiment, we ran 40 active selection rounds with 5 observations (simulated subjects) per round. The GP used the squared exponential Automatic-Relevance-Determination (ARD) kernel whose hyperparameters were inferred by slice sampling.

For each combination of the four distributions as observation model and for each combination of the four distributions as generative model, we ran the battery of 100 experiment replications each with 200 simulated subjects. The simulation results are summarized in Figure 2.3b. We collected two different measures of performance. The bar graph on the left shows, for each distribution as the observation model, the mean play duration over the 200 simulated subjects and the 400 replications of each experiment (100 replications with each of four generative models). The bar graph on the right shows the mean distance of the inferred optimum to the true optimum. Superior performance is indicated by a higher play duration
and a lower distance to the true optimum. The log-Normal distribution as observation model shows a slight advantage over the Weibull and Gamma distributions, and a large advantage over the Wald. By both measures of performance, the log-Normal distribution is most robust to incorrect assumptions about the underlying generative process. We use this observation model in the human studies that follow. This decision may seem strange in light of recent work that shows that Weibull is appropriate to model play times (Sifa et al., 2014). But that work measured total play time across multiple sessions as opposed to our present setup where we measure play time in a single session.

2.3 Experiments

Let us take a step back and remind the reader of our overall agenda. We wish to maximize retention (play duration) over a game design space. The dimensions of this space affect difficulty. We described a powerful methodology, Bayesian optimization, that can be used to efficiently search a continuous, multi-dimensional design space to identify an optimum design. Through a re-analysis of existing data and through simulation studies, we demonstrated that this methodology is promising and effective, and we developed a model that is appropriate for the dependent variable of play duration.

Finally, we can now turn to describing experiments. Our experiments were conducted using Amazon’s Mechanical Turk platform. The inspiration for using this platform came from earlier studies we conducted on Turk. In one study requiring participants to induce concepts from exemplars, we received post-experiment messages from participants asking if we could provide additional exemplars for them to use to improve their skills. In another study involving foreign language learning, participants who completed the study asked for the vocabulary list. In all cases, the participants’ motivation was to learn, not to receive additional compensation.

Given that Turk participants are willing to voluntarily commit time to activities that they find engaging, we devised a method for measuring voluntary time on activity or VTA. In each of our experiments, participants are required to play a game for sixty seconds. During the mandatory play period, a clock displaying remaining time is displayed. When the mandatory play period ends, the clock is replaced by a button that allows the participant to terminate the game and and receive full
compensation. Participants are informed that they can continue playing with no further compensation.

VTA is measured as the lag between the button appearance and the button press.

The traditional method of assessing engagement is a post-experiment survey (e.g., (O’Brien & Toms, 2010; Mekler et al., 2014)). Recently, however, VTA-like measures have been explored. Sharek and Wiebe (Sharek & Wiebe, 2015) tested several versions of a game on Turk and quantified engagement by the frequency of clicking on a game clock to reveal whether the minimum required play time had passed. Also, in work we described earlier (Lomas et al., 2013a; Lomas, 2014; Liu et al., 2014), engagement was measured by how likely a player is to switch to a different game. In gambling psychology, VTA has been extensively used to study the effect of near-misses on time spent playing slot machines (Kassinove & Schare, 2001).

2.3.1 Overt Versus Covert Difficulty Manipulations

In our experiments, we distinguish between overt and covert manipulations of game difficulty. Overt manipulations are those to which players readily attribute causal effects on difficulty, such as the speed of an enemy or the height of a wall. Overt manipulations tend to be visually salient and directly perceived from the game lay out. In contrast, covert manipulations are more subtle and involve aspects of the game to which players may not be attending or may not have an explicit theory relating these manipulations to game difficulty. An example of a covert manipulation might be the proximity that a bullet’s trajectory needs to come to an enemy in order to hit the enemy.

Although we know of no prior work in which difficulty is covertly manipulated, there is a related literature in which the appearance of difficulty is manipulated without affecting the actual difficulty. Some of this work falls under the banner of the illusion of control (Langer, 1975). In a classic study, subjects drew a card against an awkward or confident-looking confederate, winning the round if they had the highest card. Before each round, subjects placed a bet that they’d win. Subjects who played against the awkward confederate bet, on average, 47% more than those who played against a confident confederate, despite the objective probability of winning—the difficulty—being the same in both cases. In the context of video games, two examples we cited earlier (Klimmt, Rizzo, et al., 2009; Denisova & Cairns, 2015) show that players are more engaged when the perception of challenge was manipulated,
rather than the actual challenge. We hypothesize that the effectiveness of these manipulations is due to the fact that individuals readily overestimate their sense of agency—the amount of control they have over an outcome (Linser & Goschke, 2007). Unlike these previous efforts that manipulate the perception of difficulty whilst keeping the actual difficulty constant, a covert manipulation does the opposite. We may assist the player in navigating a game, thereby changing the actual difficulty, whilst giving the play an illusion that success is attributed to their own competence and skill.

In our experiments, we evaluate the effectiveness of overt versus covert difficulty manipulations on engagement.

2.3.2 Two Games and Three Difficulty Manipulations

The two games we studied are simple, popular trajectory-planning games: Flappy Bird and Spring Ninja. In Flappy Bird, the objective is to keep a bird in the air by flapping its wings to resist gravity and avoid hitting the ground, the top of the screen, or vertical pipes (Figure 2.4a). In Spring Ninja, the objective is to wind a spring to the proper tension so that the player jumps from one pillar to the next and avoids falling to the ground (Figure 2.4b). The player holds and releases a mouse button to jump. The longer the player holds, the further the ninja jumps. Both Flappy Bird and Spring Ninja involve trajectory planning, but the former requires real-time decision making whilst the latter allows players to take their time in planning the next jump.

We manipulated two overt factors affecting the difficulty of Flappy Bird—the horizontal spacing between pipes and the vertical gap between pipes—as well as one covert factor, which we refer to as the assistance. Assistance acts as a force that, when the wings are flapped, steers the bird toward the gap between the next pair of pipes. In Spring Ninja, we manipulated two overt factors—the horizontal spacing between the pillars and the visible extent of a projected trajectory (the blue curve in Figure 2.4b)—as well as the amount of covert assistance. The assistance in Spring Ninja corrects the trajectory of the player if the trajectory falls within a certain distance of the ideal trajectory. In both games, the assistance level can be adjusted to range from no assistance whatsoever to essentially a guarantee that nearly any action taken by the player will result in success. For moderate levels of assistance, the manipulation can be quite subtle. We have no experimental evidence that players were unaware of our ‘covert’ support,
Figure 2.4: (a) Flappy Bird: The player flaps bird’s wings to keep it aloft and to avoid hitting pipes. (b) Spring Ninja: The player jumps from one pillar to another by compressing springs in the ninja’s shoes. The blue trajectory is the projected jump path for the given spring compression level.

but anecdotally, players who tested our games with low-to-moderate levels of assistance were surprised when they were informed that game dynamics were modulating to guide them along. Indeed, it was shocking to realize that one could perform relatively well with eyes closed.

2.3.3 Flappy Bird: Experimental Methodology and Results

We conducted two studies with Flappy Bird. In the first study, we tested 958 participants. Each participant was assigned to a random point in the three dimensional, continuous design space. The large number of participants in this random-assignment experiment enabled us to fit an accurate model that characterizes the relationship between the game design and latent engagement, much as we fit data from the Treefrog Treasure game which was collected by random assignment (Figure 2.2). In the second study,
we ran the experiment again from scratch and tested 201 participants. Participants were assigned to
designs chosen by an active-selection policy, Thompson sampling as described earlier. Active selection
chooses a design for each participant based on the model estimated based on all previous participants.

Our pilot experiments suggested that randomly seeding Bayesian active selection is necessary,
as is often done with BO. Consequently, we assigned the first 55 participants in the active-selection
study to a Sobol-generated set of random points in design space. Sobol sequences (Levy, 2002) are
attractive because they evenly cover design space, as opposed to a sequence generated from a pseudo-
random number generator. After the seeding phase, we performed rounds of Bayesian optimization using
Thompson sampling with five subjects tested at each selected design.

The design space consisted of three dimensions: pipe spacing, pipe gap, and covert assistance.
Each dimension was quantized to 10 levels.\(^2\) Participants were given game instructions and were told that
to receive compensation (20 cents) they must play for 60 seconds, but they could continue playing without
further compensation for as long as they wished. During the mandatory-play period, a countdown timer
in the corner of screen indicated the time remaining. During the mandatory-play period, multiple rounds
of the game were played. Each round was initiated with a mouse click and ended when the bird crashed.
When the mandatory time was reached, the time-remaining display was replaced by a ‘finish’ button.
Because individuals might not notice the button mid-round, we excluded the round in play, and defined
VTA to be the time (in seconds) beginning with a mouse click to initiate the first round once the finish
button had appeared.

At any time, clicking finish took participants to a final screen that indicated how much time they
had spent beyond the mandatory time; this number could be zero if no new rounds were played following
the mandatory time. Participants were asked to enter how long they expected other mechanical turk
players to voluntarily play. The two dependent measures available then were the experiential and projected
VTA. In pilot experiments we treated both measures as independent so there were two observations per
participant. However, this led to non-smooth model fits to the data so we decided to use the projected
VTA exclusively as our measure of engagement. Projected VTA is less contaminated by confounds, e.g.,
the player would have liked to continue but had another obligation, or the player continued for several

\(^2\)This quantization may seem strange given that BO can handle continuous dimensions. However, 10 levels allows for
fine distinctions, and allows us to avoid local-optimization techniques such as hill climbing needed for continuous spaces.
rounds only because they had not noticed the finish button. Although it may seem that we are ignoring
the important behavioral signal in the experiential VTA, we are still making use of that signal because
the experiential VTA is provided as a reference when participants are asked to specify the projected
VTA. In the random-assignment study, we displayed the experiential VTA on the screen and asked
participants to enter the projected VTA. In the active-selection study, to emphasize the experiential
VTA, we incorporated a slider control that is initially anchored on their experiential VTA (see top of
Figure 2.5). The slider had a range of at least 0-100, and if the experiential VTA was greater than 100,
the top end was set to twice the experiential VTA, rounded up to the nearest multiple of 100.

In the active-selection study, we included a short questionnaire about the participant’s experi-
ence in the game. The questionnaire consisted of 6 true/false items with each item phrased such that
“true” corresponds to an engaging game. The first four phrases in the questionnaire (Figure 2.5) were
taken from the Game Engagement Questionnaire Brockmyer et al. (2009).

The top and bottom rows in Figure 2.6a show the model posterior mean VTA over the three
dimensional design space in the random-assignment and active-selection studies, respectively. The re-
markable finding is that the two independent studies yield very similar outcomes: the optimal designs
identified by the two studies are almost exactly the same (the cyan squares in the Figure). The random-
assignment study should yield reliable results due to the relatively large number of participants tested.

An important question here is whether the predictions of the model are related to the obser-
vations. Because repeated observations within the same game design are highly variable, we averaged
observations for each design tested and determined the correlation with the expected VTA predicted
by the model. We included in this analysis only designs for which we had four or more observations in
our random-assignment experiment. We obtain a Spearman correlation of 0.65. This coefficient is close
to the value obtained by fitting the model to synthetic data generated by the model itself (0.50±0.1,
10 replications). The fact that the model predicts the actual data as well as if not better than the
synthetic data suggests that the model is appropriate for the task. (We used the random-assignment
experiment for this analysis; a similar analysis for the active-selection experiment is not sensible given
the dependence among samples.)

To compare the efficiency of random-assignment vs. active selection, we randomly sampled
Questionnaire

1. After finishing the last mandatory game, you have played an extra 30 seconds over the minimum. How long do you think other Mechanical Turk users would play, on average, over the minimum? (you can move the slider or type in the number directly)

   30 seconds.

2. Please indicate whether each of the statements below accurately describe your experience in the game.

   - True  False  I lost track of time.
   - True  False  Time seems to kind of stand still or stop.
   - True  False  Playing makes me feel calm.
   - True  False  I play longer than I meant to.
   - True  False  I enjoyed playing the game.
   - True  False  I would download the game if it were a mobile application.

If you would like to play this game later (outside of Mechanical Turk), you may copy the link below.

http://example.url/webpackentry/

Figure 2.5: The post-experiment questionnaire.
200 observations from the random-assignment study and fitted our model using only those observations. We then calculated the distance between the optimum found using 200 observations and the optimum found using the full set of 958 observations. We replicated this procedure 50 times, each time sampling a different subset of observations. The mean distance over these 50 replications was 0.70 (std. error ±0.03). In contrast, the distance was 0.28 in the active selection study, using the same number of observations. This result indicates that with matched budgets for data collection, the active-selection study is more efficient than random selection in converging on the optimum.

The Figures indicate that engagement is sensitive to each dimension in the design space. There is not much hint of an interaction across the dimensions. Notably, with minimal covert assistance (the left array in each row), the other two overt difficulty dimensions have little or no impact on engagement, and are not sufficient to motivate participants to continue playing voluntarily. Thus, we conclude that covert assistance is key to engaging our participants. Consistent with the hypothesis that participants need to be unaware of the assistance, the experiments show that engagement is poor with maximum assistance (the lower-right array in each Figure). With maximum assistance, the manipulation causes the bird to appear to be pulled into the gap, and this is therefore no longer covert in nature.

To obtain further converging evidence in support of the optimum designs identified in Figure 2.6a we fitted a Gaussian process model to questionnaire scores. We defined the score as the number of "true" responses made by the participant. The higher the score, the higher the engagement because we phrased questionnaire items such that an affirmative response indicated engagement. We used Gaussian process regression with a Binomial observation model to fit the scores. (Our VTA model is appropriate for fitting play-time observations, whereas the scores lie in a fixed range of 0-6.) Figure 2.6b shows the model posterior mean score over the three dimensional design space. The notable result here is that the posterior mean score looks similar to the posteriors from the random-assignment and active-selection studies. More importantly, the predicted optima—marked by cyan squares—lie in almost exactly the same place in 2.6a and 2.6b. Whereas the posterior inferred from the questionnaire scores looks different, e.g., for assistance=1, we remind the reader that the objective of BO is to find the maximum of a function, rather than map out the full design space. The consistency across studies and across response measures provides converging evidence that increase our confidence in the experiment outcomes, and also provide
support for the appropriateness of using VTA as measure of engagement in place of a more traditional questionnaire.

Figure 2.6: Bayesian model fits of VTA (in seconds) over the Flappy Bird design space for (a) the random-assignment and active-selection studies. Each array corresponds to a fixed level of assistance, with the left array being no assistance (level 0) and the right array being maximal assistance (level 1). For each fixed level of assistance, the corresponding array depicts model-fit VTA across the range of horizontal spacings between pipes (x axis) and vertical gaps (y axis). The pipe gap and pipe spacing is calibrated such that a level of 0 (bottom left) is a challenging game, unlikely to be played well by a novice, and 1 (top right) is readily handled by a novice. The circles correspond to observations with the areas indicating the magnitudes of the observations. Cyan squares indicate the locations of the top 5% of predictions. (b) An analogous Bayesian model fit to the questionnaire score, which indicates the number of items with an affirmative response. Higher scores indicate greater engagement.

2.3.4 Spring Ninja: Experimental Methodology and Results

We conducted two studies – A and B – with Spring Ninja with 325 and 328 participants, respectively. As in the active-selection Flappy Bird study, we seeded the optimization procedure with participants evaluated with designs generated from a Sobol sequence, 54 in total. The remaining participants were tested in groups of five with a game design chosen from an active-selection policy, Thompson sampling. We did not conduct a random-assignment study with Spring Ninja due to time constraints and because, in addition to cited literature on the efficiency of Bayesian optimization, we have already established the effectiveness of our active selection method in Flappy Bird, which is a similar three-parameter game.

The design space of Spring Ninja consisted of three dimensions: the spacing between pillars, the visible extent of the projected trajectory and the covert assistance. Each dimension was quantized into
10 levels in the range 0–1 with 0 and 1 corresponding to difficult and easy game settings, respectively. The optimization procedure sought to maximize the VTA, defined for this game as the number of jumps a player would make after the appearance of the finish button.

As in the Flappy Bird studies, Spring Ninja participants were required to play for a minimum of 60 seconds in order to receive compensation (20 cents). A countdown timer was shown in the corner of the screen and replaced with a finish button when the timer reached zero. The timer counted down only from the time at which the participant began compressing the spring and stopped after the ninja landed on a pillar or fell off the screen. When the player falls off the screen, a game-over screen is shown offering the player to start a new game or—if the mandatory play time had elapsed—finish the experiment. When the finish button is clicked, participants are redirected to a post-experiment screen in which they specify their projection of others’ VTA and respond to the same questionnaire as in the Flappy Bird studies (Figure 2.5).

We measure the VTA in Spring Ninja differently than in Flappy Bird because the former is turn-based whereas the latter is continuous. Specifically, Spring Ninja players are likely to notice between jumps when the countdown timer hits zero and the finish button appears because they are not under time pressure. We could define VTA as the time after the finish button appears but this poses a problem when we ask participants for the projected VTA since there is a mismatch between the game’s sense of time—time advances only when the Ninja is flying or about to fly—and real world time. So a participant would be perplexed if they found out that they have played for only 20 seconds extra when they have actually played for one more minute. Indeed, we received several emails from pilot participants complaining about this issue. To avoid this problem, we instead measure the number of jumps after the finish button appears. The number of jumps is agnostic to the way the game measures time and is a non-negative quantity that is directly proportional to VTA so we can still use our VTA model. We shall continue to refer to the number of voluntary jumps as the VTA.

To assess the coveryness of assistance, we added two complimentary questions about the difficulty of the game to the post-experiment screen in study B. The two questions ask subjects to rate the difficulty of the game and their own performance on a scale from 1-5. We combine both measures of difficulty by flipping the performance rating and taking the average of the two measures, \( \frac{\text{difficulty} + (6 - \text{performance})}{2} \).
yielding one measure of difficulty.

Figure 2.7a shows model posteriors over VTA (in number of jumps) for both active selection studies (top and bottom rows), fit in the same way as we did in the Flappy Bird study. Both studies predict optimal engagement at medium to high covert assistance values. Study B predicts that the easiest designs are the most engaging (top right corner), while study A predicts that optimal engagement is closer the middle of the design space. Predictions of questionnaire scores in Figure 2.7b indicate that subjects prefer designs with moderate covert assistance, but there is overlap among the top 5% of predictions of VTA and questionnaire scores, despite the fact that active selection did not optimize for questionnaire scores. In terms of reported difficulty, predictions in the top row of Figure 2.7c show that the most difficult designs are in the lower left corner of the design space. Covert assistance values smaller than 0.6 reduce the reported difficulty in the lower left corner but values greater than 0.6 do not have much of an impact on reported difficulty. Yet, subjects are scoring the highest during the mandatory period when all design dimensions $\geq 0.9$, suggesting a dissociation between the perception of difficulty (reported difficulty) and the actual difficulty (highest mandatory score). Of course, the lack of difference in reported difficulty for covert assistance $> 0.6$ could also be due to a floor effect where subjects couldn’t rate the difficulty any lower. Finally, we pooled the observations from both studies and plotted model predictions of VTA and questionnaire scores in the top and bottom rows of Figure 2.7d, respectively. VTA predictions indicate that subjects play designs with moderate to high covert assistance the longest, but they subjectively rate designs with moderate assistance the highest. This could be due a ceiling effect where the VTA for assistance values $> 0.6$ saturates so subjects rate designs that promote greater self-attribution of success (moderate covert assistance) higher than ones that do not (high covert assistance). Nonetheless, pooled VTA and questionnaire results both predict that the easiest designs are the most engaging which is consistent with findings in (Lomas, Patel, Forlizzi, & Koedinger, 2013b). Consistent with the Flappy Bird study, both un-pooled and pooled results indicate that the two overt difficulty manipulations have little impact on engagement when no covert assistance is provided (the left array), yet with moderate to high covert assistance, engagement significantly increases.
2.4 Discussion

In this article, we’ve applied an increasingly popular tool from the machine learning literature, Bayesian optimization, to a problem of intense interest in the fields of gaming and gamification: How do you design software to engage users? In contrast to traditional A/B testing, Bayesian optimization allows us to search a continuous multi-dimensional design space for a maximally engaging game design. Bayesian optimization is data efficient in that it draws strong inferences from noisy observations. Consequently, experimentation with users on suboptimal designs can be minimized. When placed in a live context, Bayesian optimization can be used to continually improve the choice of designs for new users.

Bayesian optimization is a collection of three components: (1) Gaussian process regression to model design spaces, (2) a probabilistic, generative theory of how observations (voluntary usage times) are produced, and (3) an active-selection policy that specifies what design to explore next. A key component of the research described in this article is our exploration of candidate generative theories, and a contribution of our work is the specification of a theory that is robust to misspecification, i.e., robust to the possibility that humans behave differently than the theory suggests.

We collected multiple measures of engagement, including experiential and predicted voluntary time on activity and a post-usage survey with questions indicative of engagement. We also showed that usage time and the survey yield highly consistent predictions of maximally engaging designs. The converging evidence from these two very different measures gives us confidence in our interpretations of the data.

Beyond our methodological contributions, we explored a fundamental question regarding engagement and game difficulty. Moving beyond the well-trodden notion that game difficulty can affect engagement, we compared covert versus overt manipulations of difficulty. We found that overt manipulations on their own were relatively ineffective in modulating engagement (at least over the range of designs we tested), yet they became quite effective when coupled with a covert manipulation in which we provided assistance in a subtle manner, possibly skirting the player’s awareness. We believe that players attributed the improved performance resulting from our covert manipulation of game dynamics to their own competence. Their boost in perceived competence led to increased engagement. We envision that
this covert-assistance trick could be used to draw players into a game and then be gradually removed as the player’s true skill increases.

In future research, we plan to address three limitations of the present work. First, we would like to conduct longer-term usage studies to show that the effects we observe on engagement scale up with longer use of software. Second, we would like to optimize simultaneously for both VTA and questionnaire scores so that the BO procedure exploits all the data provided by each subject. Third, rather than optimize design parameters for a user population as a whole, the same methodology could be applied to optimize for a specific user, conditioned on their play history. For such a task, the data efficiency of Bayesian optimization is critical.
Figure 2.7: (a) Model predicted VTA (in ninja jumps) over the Spring Ninja design space in the two active selection studies (top and bottom rows). Each array shows VTA for a range of trajectory lengths and horizontal spacings between the pillars. The trajectory length and pillar spacing is scaled such that 0 (bottom left) is a challenging game, unlikely to be played well by a novice, and 1 (top right) is readily handled by a novice. Each of the 10 arrays represents a fixed level of assistance, with 0 being none and 1 being maximal. Each cell in an array corresponds to a setting of the trajectory length and the horizontal spacing between pillars. The circles correspond to observations with the areas indicating the magnitudes of the observations. Cyan squares indicate the locations of the top 5% of predictions. (b) An analogous Bayesian model fit to the questionnaire score, which indicates the number of items with an affirmative response. Higher scores indicate greater engagement. (c) A Bayesian model fit to the subjects’ reported game difficulty (top row) and the maximum score during the mandatory period (bottom row) in study B. Higher scores indicate greater difficulty and mandatory scores. (d) VTA and questionnaire model fits to the combined dataset from both active selection studies.
Chapter 3

Tension and Release

Tension-and-release is a common technique for structuring events which aims to induce a pleasurable feeling in the subject as he or she prepares for an *expected* event. For example, a musician can progressively enrich the texture (number of instruments) - creating tension in the listener - before resolving to a single or few instruments - creating release. In ancient and Shakespearean dramatic structure, drama is divided into acts that are collectively known as an arc: exposition, rising action, climax, falling action and denouement (Wikipedia, 2016). Contemporary Hollywood films exhibit tension-and-release too: in a typical super hero movie, viewers are *certain* that the hero will win at the end, but the hero’s struggle to get there increasingly builds up the audience’s anticipation or tension, making the final positive outcome extremely rewarding. Video games make extensive use of tension-and-release (Rose, 2016), usually by manipulating game difficulty over time: first-person shooter (FPS) games typically include periods of intense firefights followed by lull exploration phases. These previous examples suggest that tension-and-release relies on abrupt change to accentuate enjoyment: high to low music texture, a movie hero going from a state of hopelessness to victory, a game player getting chased by tank then commandeering a tank herself. The anticipation of this abrupt change has been found to be itself rewarding, according to a Neuroscience study: “*Notably, the anticipation of an abstract reward can result in dopamine release in an anatomical pathway distinct from that associated with the peak pleasure itself*” (Salimpoor, Benovoy, Larcher, Dagher, & Zatorre, 2011). This raises another question: If you are listening to a piece of music for the first time, then surprise can account for enjoyment, but why do you still enjoy a piece of music
after listening to it multiple times, i.e., after surprise has vanished? D. Huron and Hellmuth Margulis (1993, p. 580) cites an article that argues that listeners of music maintain “schematic expectations” that characterize music in general or a specific genre and “veridical expectations” that characterize a specific musical piece. So while a piece of music may not be surprising after listening to it multiple times, it may still be surprising under the schematic expectations. To borrow natural language processing terminology, listeners maintain two language models, one general and the other specific, and the tension arises during the buildup to a rewarding event that is unexpected under one of the two language models.

Tension-and-release is not the same as notion of surprise. The latter occurs after an unexpected event occurs, whereas the former occurs in the buildup to an expected event onset. Surprise can break tension by introducing an unexpected event during the buildup of tension (D. B. Huron, 2006), effectively neutralizing the tension before it reaches the peak. One can see this in some action movies or animations where a tense fight scene is interrupted by an unrelated cutaway. In these instances, tension is reduced and engagement may even be negatively affected. Experiments in this chapter operationalize tension-and-release by manipulating difficulty over time – the difficulty curve – but surprise is also addressed in the Experiments 3.3 and 3.4 by shuffling tension-and-release difficulty curves.

One may confuse tension-and-release with Dynamic Difficulty Adjustment (DDA), which has been studied extensively in literature. The two techniques are similar in that they both manipulate difficulty over time but they operate on different levels. DDA tries to maximize player engagement by selecting the difficulty setting that best matches the current skill level of the player. This objective is grounded in the theory of flow (Chen, 2007) which postulates that maximum engagement occurs when the player is neither too bored nor too frustrated. Examples of DDA include AI players that adapt to player skill (Andrade et al., 2005; Andrade, Ramalho, Gomes, & Corruble, 2006), adaptive spawning of health and ammunition kits in FPS games (Hunicke, 2005), and adaptive procedural level design (Jennings-Teats, Smith, & Wardrip-Fruin, 2010). On the other hand, tension-and-release adjusts difficulty over time – the difficulty curve – assuming that the current difficulty is matched to the player’s skills. In order words, tension-and-release introduces structured variability into the difficulty curve. A qualitative example of how this works is illustrated in Figure 3.1. Here you can imagine a DDA algorithm controlling the progression along the flow channel, while a tension-release algorithm introduces periodicity, given a
difficulty level. Of course, it is possible that tension-and-release is generated as a side effect of a DDA policy. For example, a simple policy that decreases difficulty if the player dies \( x \) number of times will produce tension as the policy will not kick in before the threshold \( x \) is crossed so as the player dies more often, he or she will predict with increasing certainty that the game will intervene to make it easier, thereby increasing tension up to the moment of intervention.

### 3.1 Related Research

Qin, Rau, and Salvendy (2010) studied the effects of different difficulty curves on player immersion. The authors considered combinations of difficulty curves, difficulty change rates, and peak difficulty levels in a 2D fighting game. Difficulty changes were carried out over four “scenes” of the game, with the difficulty of each scene being measured by game elements such as the number of enemies, their strength, and durability. Three difficulty curves were evaluated: up-down (easy \( \rightarrow \) hard \( \rightarrow \) easy), down-up (hard \( \rightarrow \) easy \( \rightarrow \) hard), and continuous (easy \( \rightarrow \) hard). Difficulty change rates—slow, medium, and fast—refer to how quickly difficulty changes over game scenes, e.g., in the slow setting, the difficulty changes slowly from scene one to two to three but then change abruptly to meet the peak difficulty at scene 4. Two peak difficulty levels were considered: easy and hard. Subjects played all 3 difficulty curves and 2 peak difficulty conditions over 6 rounds of the game, where an immersion questionnaire was administered after each round. A round consisted of 4 scenes in the game over which difficulty was manipulated. Subjects were assigned to one of the three difficulty change rates. Player immersion was higher in the up-down condition than the other two conditions. But, when the difficulty change rate was slow, immersion was found to be higher in the down-up condition than in the up-down condition. As the change rate
increased, the up-down condition was found to be more immersive than the down-up condition. Medium change rates were found to be optimal in terms of immersion, with no reliable differences between slow and fast rates. The advantage of the up-down condition was attributed to players feeling challenged as difficulty increased, then feeling greater skill as it decreased.

Three difficulty curves were investigated during the evaluation of a procedural content generator (PCG) (Adrian & Luisa, 2013), a program that generates game levels automatically using a set of rules and constraints. The PCG used genetic algorithms to create game levels whose difficulty matched a given difficulty curve. The game used for analysis was an obstacle avoidance course where the player slides through a tunnel while collecting items and avoiding holes and obstacles. Difficulty was characterized by the number of obstacles and empty tiles within segments of the tunnel. Two versions of the PCG – unconstrained and constrained – generated levels that matched constant, random, and tension-and-release difficulty curves. Subjects (n = 22) then played 6 rounds of the game, one for each combination of difficulty curve and PCG version. After completing each round, subjects rated the round on difficulty, level design, and fun. The authors argued that the least and most fun difficulty curves were the constant (mean rating=7.59) and tension-and-release (mean rating=7.77) curves, respectively. But the reliability of the results is unknown due to lack of significance testing and error bars in the plots.

Gutwin, Rooke, Cockburn, Mandryk, and Lafreniere (2016) looked at how peak-end manipulations affect, among other things, player enjoyment. Peak-end manipulations assume that subjects evaluate an experience based on its peak and end moments. For example, the paper cited a study where subjects immersed their arms in painfully cold water for short and long durations. In the short condition, subjects removed their arms after 60 seconds, while in the long condition, they removed their arms after 90 seconds but during the last 30 seconds the water was gradually warmed up by one degree Celsius. The study found that most subjects chose to repeat the long condition over the short one, even though the former has greater overall pain. To evaluate the peak-end effect within the context of video games, the authors ran a study to see whether subject’s enjoyment was affected by peak-end manipulations of game difficulty. The study manipulated the temporal sequence of game difficulties with three different conditions: positive (easy peak in the middle and end), neutral (constant difficulty throughout) and negative (difficult peak in the middle and end). All conditions had the same average difficulty. Participants
played twelve games and provided pairwise comparisons between each consecutive pair. Two different
game applications were evaluated in this study. In the first application, about 80% of the participants
\(n = 24\) preferred the positive condition over the negative one and 75% indicated that they would
repeat the positive condition over the negative one (both of these results were significant). While there
were trends showing players preferring positive over neutral and neutral over negative, none of those
were significant. In the second application, no significant differences were found in pairwise preference
judgments among the three conditions. The authors hypothesized that success in the first game is more
rewarding than the other. The results by Gutwin et al. (2016) appear to suggest that it might be better
to use difficult-easy-difficult sequences which is contrary to the traditional notion of tension-and-release.
However, the conditions evaluated in the paper confounded the peak and end difficulty; the authors only
evaluated conditions with both peaks and ends and did not evaluate conditions where one was present
but not the other. It is possible for example, that players were merely remembering the last experience
they had in the game, regardless of whether there was a peak or not.

This chapter addresses the issues in previous research by analyzing the full spectrum of condi-
tions in four tension-and-release manipulations using Gaussian processes and preference learning models.
Multiple measures of engagement are used for analysis and robust baseline conditions, constant and ran-
dom, are evaluated. The next section discusses the design of the game I developed and used in my
experiments.

3.2 The Shapes Game

3.2.1 Motivation

Psychologists who study human memory use different instruments or tasks to measure \textit{working memory}
(WM), a cognitive component responsible for short-term storage and processing of information. The
\textit{N-Back Task} (Kirchner, 1958) is particularly popular because it is simple to administer and has a single
knob, \(N\), that controls the WM load. In this task, subjects are presented with a sequence of items and
are instructed to hit a key if the current item matches the item presented \(N\) items back. The original
form of the task is spatio-temporal, where the subject sees a grid with one cell highlighted and he or
she are expected to click on the cell that was highlighted $N$ items back. As $N$ increases the difficulty increases because the subject has to maintain a larger set of items to recognize in working memory.

Despite the attractiveness of N-back as a WM measure, performance on N-back has been shown to have weak correlation to more complex WM tasks (Jaeggi, Buschkuehl, Perrig, & Meier, 2010). Here, the authors correlated N-back with the reading span task (RST) and the self ordered pointing task (SOPT) on accuracy and response time (RT). In RST, subjects are presented with a set of sentences sequentially, all syntactically correct but not necessarily semantically meaningful, and they are cued to recall the final word of each sentence—in order—at the end of presentation. In SOPT subjects see the same set of stimuli on each trial (the locations of stimuli are randomized) and are required to select a stimulus that hasn’t been selected before. No correlation between performance on RST and N-back, neither on RT nor on accuracy, was found. On SOPT, only modest correlation was found with the N-back task on accuracy and RT. The authors hypothesized that the weak correlations between N-back and RST and SOPT are due to latter two enabling subjects to develop strategies, such as how to remember last words in the RST task.

I address the shortcomings of the N-back task as a WM measure by designing a game, ShapeFactory, that combines the strengths of N-back and other complex WM measures. The game is easy to interact with: subjects only need to click or press a key to provide responses. The WM memory load can easily be manipulated without having to create new content (as opposed to RST, where additional sentences are required to increase WM load). Finally, the load on executive function is easily manipulated independent of subject performance (compared to SOPT, where past performance dictates future rules). Good games are engaging so in this chapter I focus on studying engagement in ShapeFactory as function of several tension-and-release manipulations. Once an engaging version of ShapeFactory has been identified, future studies can look at the validity of the game as a WM measure and how performance transfers to other measures.

### 3.2.2 Game Design

ShapeFactory is inspired by an iPhone game called Jetset (Bogost, 2009), where the player assumes the role of a screener at an airport security checkpoint. When passengers arrive at the checkpoint, the player
sees the contents of their suitcases and what they are wearing. The player has to remove all prohibited items that passengers may be carrying or wearing but there is a catch: the list of prohibited items, which is visible at all times, changes frequently. The player is allowed to make up to 5 mistakes, removing an allowed item or ignoring a prohibited item, before losing the game and he or she gets awarded for quick processing of passengers. A screenshot of the game is shown in Figure 3.2.

In place of passengers, ShapeFactory has *objects* which have shape (triangle, square, pentagon, and circle), color (green, blue, purple, and red) and numeric value (1-4). A set of partial templates that tells the player what to select, e.g., blue objects or pentagons or circles with the number 1. Objects move from right to left in a line on the player’s screen so the player has a limited time to make a decision to select or ignore an object. The game proceeds in rounds with each round containing a fixed number of objects. Before each round, the player sees the rules (Figure 3.3a) but once the round starts, the player has a limited number of opportunities to see the rules, putting pressure on the player’s WM as he or she must remember the set of rules in order to achieve a good score (Figure 3.3b). The scoring system is designed such that the player achieves a minimum score if they choose to select or ignore all objects so the player is not incentivized to behave either way. Different sound effects give the player feedback on correct/incorrect selections and incorrect misses (the first two effects are loud and the last one is subtle). Animations are also used to update the score bar on top of the line. Game difficulty can be manipulated with scroll speed, inter-object arrival times, or by the number and complexity of the rules, e.g., introducing disjunctions of conjunctions of rules.

Figure 3.2: Jetset game (Bogost, 2009)
3.3 Experiments

I conducted four experiments exploring different tension-and-release manipulations in ShapeFactory, all of them manipulating the number of objects per second via scroll speed or inter-item arrival times. All experiments require subjects to apply a disjunction of two singular rules that target different attributes. So a subject may be required to select blue objects or circles, but he or she will never be required to select blue objects or orange objects, blue pentagons or circles with 1, and so on. Experiments were conducted on Amazon’s Mechanical Turk platform, where subjects complete web-based experiments, called hits to receive compensation. To enroll, subjects were required to have 100+ approved hits at at least 99% approval rate. Enrollment was limited to populations in the United States and Canada. A brief orientation was administered after subjects had accepted the hit to ensure that they understood how to play the game and apply the rules. In experiments 3 and 4, a post-experiment survey was administered.
to gather additional measures of engagement.

### 3.3.1 Experiment 3.1: Tension and Release via Scroll Speed

In the first experiment, I evaluate a few conditions in a classic A/B design to see if tension-and-release can modulate engagement. Three conditions which manipulate the scroll speed over time are tested: easy-hard-easy or *EHE*, corresponding to slow-fast-slow scroll speeds, constant or *CONST*, corresponding to a constant scroll speed, and hard-easy-hard or *HEH*, corresponding to fast-slow-fast scroll speeds, all having the same average speed of 2 objects/sec (Figure 3.4). The maximum number of objects subjects see at any given point in time is 3.3 objects. CONST is a simple baseline condition and EHE is what is traditionally recognized as tension-and-release but it is possible that release-and-tension works just as well, hence the inclusion of HEH.

Subjects are randomly assigned to pairs of conditions (EHE vs. CONST, EHE vs. HEH, and CONST vs. HEH); the order of conditions within each pair is randomized, so there is a total of 6 possible pairs. The game lasts for 10 rounds and after playing both conditions in the first two rounds, subjects are free to pick the condition to play before each of the remaining 8 rounds. The condition selection screens shows screenshots of both conditions along with labels, “switch to” and “repeat”, to help subjects remember which condition they’ve just played (Figure 3.5). Screenshots of the conditions are always presented in same order as the assigned pair so the order of the labels changes depending on the most recently played condition. Subjects are told beforehand that the conditions differ in scroll speed but they are not told what the conditions are. Different background colors (light blue or pink) and container shapes (crate or book) are assigned to each condition on a per-subject basis to help subjects distinguish conditions. There is no time limit in the selection screen and the next round only starts after a selection is made.

Rounds last 35 seconds, including 1.5 seconds of setup time to help subjects prepare. The total number of objects is $4 \times 4 \times 4 = 64$, with two rules specifying what to select (color or number, color or shape, and shape or number) so the total number of matches is $2 \times (4 \times 4) - 4 = 28$. Two rules are assigned randomly to each subject, but attribute values change randomly on each round, e.g., blue or circle, then orange or square, etc., with the constraint that all values of each rule are used for the 10
rounds. During a round, subjects have 3 opportunities to see the rules – 1 second per opportunity – but objects keep scrolling during these periods. Subjects select an object by left clicking on it with the mouse, triggering a “robotic” arm painted in black and yellow to come up from the bottom edge and pull the object out of the line. Objects on the line are randomly arranged in blocks of 8 objects, with the constraint that 4 blocks contain exactly 4 matches per block and the remaining 4 blocks contain exactly 3 matches per block.

Engagement is measured by how often subjects pick one condition over the other. Since conditions are matched for average difficulty, subjects are assumed to select the condition that is more engaging rather than least difficult. The first two rounds are discarded so subjects can select a condition a maximum of 8 times. The independent variable is the pair of conditions (ignoring the order of conditions in the pair) and the dependent variable is how often conditions in the pair are played. Subjects were paid 80 cents to play the game for about 8 minutes on Mechanical Turk.

3.3.1.1 Results

A total of 106 subjects completed the experiment but records from 8 were discarded from the analysis because logs showed rounds where the number of items processed by the player (selected or ignored) was less than 64, indicating possible software malfunction. Subjects were assigned to conditions randomly.

Before analyzing engagement, we look at whether our manipulation of scroll speed corresponds to difficulty. After all, if difficulty is not affected by the scroll speed, then the manipulation is not really modulating tension-and-release. The lag leading up to an item, the inter-item arrival time, changes with

Figure 3.4: The three scroll speed functions in Experiment 3.1.
the scroll speed. So for each value of the lag we compute the mean accuracy across all conditions and subjects on items with that lag value. I discard the first item because it is preceded by a long setup time (1.5 seconds). Figure 3.6 plots mean accuracy for each inter-arrival time or lag. The arrival times were collected empirically so there is some noise in the value of the arrival time, e.g., the dense group of observations at 500ms corresponds observations from the constant scroll speed condition. This noise makes it possible to assign different colors to observations that come from different conditions, making it easier to visualize the interaction between inter-item arrival times and condition. Overall, the accuracy increases almost linearly with lag in the region from 300ms to 500ms and climbs slowly before saturating at 95% accuracy afterwards. Spearman’s correlation between accuracy and inter-arrival time is very high at 0.93 and the absolute percentage difference between the minimum and maximum accuracy is large at 20%. Looking at the observations from EHE and HEH, the latter has marginally lower mean accuracy at small inter-item arrival times and both conditions have very similar mean accuracy in the 500ms region. So the objective difficulty of the game, characterized by accuracy, is modulated by the inter-item arrival times and to a smaller extent, the scroll speed condition. Accuracy may reflect engagement – people could focus more and score higher on engaging games – so the small difference in mean accuracy at small inter-item arrival times between the two conditions could be due to EHE being more engaging than HEH.

This brings us to the engagement question. Statistics and linear model fits of 98 subjects are shown in Table 3.1. Results of a linear model that predicts play count difference as a function of the
Figure 3.6: The mean accuracy of each lag value in Experiment 4.1. Blue, orange, and black points correspond to observations from the EHE, HEH, and CONST conditions, respectively.

assigned pair are shown in the third column of Table 3.1. The model is significant \( F(3, 95) = 10.44, p < .001, \text{ Adj. } R^2 = 0.18 \) and shows that subjects reliably play 2-3 rounds more of CONST over EHE or HEH, and about 3 rounds more of EHE over HEH. Results of a linear model that predicts difference in mandatory scores – scores from the first two rounds – as a function of the assigned pair are reported in the fourth column of Table 3.1. The model is significant \( F(3, 95) = 8.875, p < .001 \), Adj. \( R^2 = 0.20 \) and indicates significant differences between CONST and the other two conditions, with players scoring 4-5 percentage points greater on CONST vs. other conditions, but there is no difference in scores between EHE and HEH. Note that the mandatory score difference is a much coarser measurement than the measurements in Figure 3.6 and it only limited to the mandatory rounds.

It is possible that subjects are picking the CONST condition because they can score higher, as differences in mandatory scores suggest. It is also plausible that subjects are just picking whichever condition happens to be on the left side of the screen (recall that screenshots of conditions are always presented according to their order in the assigned pair). To address these two possibilities, I fitted a linear model that predicts play count difference as a function of assigned pairs, difference in mandatory scores, and whether the first condition in the pair is presented on the left side. The model is significant \( F(5, 93) = 7.389, p < .001 \), Adj. \( R^2 = 0.21 \) and the corresponding ANOVA shows a significant effect of
<table>
<thead>
<tr>
<th>Pair</th>
<th>Subjects</th>
<th>Mean Play Count Diff.</th>
<th>Mean Mandatory Score Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONST - EHE</td>
<td>32</td>
<td>1.88 ($p = .022$) (0.278, 3.472)</td>
<td>0.052 ($p &lt; .001$) (0.026, 0.079)</td>
</tr>
<tr>
<td>EHE - HEH</td>
<td>32</td>
<td>2.75 ($p = 0.001$) (1.153, 4.347)</td>
<td>0.00 ($p = .971$) (-0.027, 0.026)</td>
</tr>
<tr>
<td>CONST - HEH</td>
<td>34</td>
<td>2.94 ($p &lt; 0.001$) (1.392, 4.491)</td>
<td>0.044 ($p = 0.001$) (0.018, 0.069)</td>
</tr>
</tbody>
</table>

Table 3.1: Summary of Experiment 3.1 results and linear model fits. A greater play count or score difference indicates that the first condition in the pair has a greater play count or score. Bold face fonts indicate differences that are significant at the $\alpha = 0.05$ level. Number pairs in parenthesis are 95% confidence intervals.

The condition pairs ($F(3, 93) = 4.754, p = 0.004$) and of mandatory score difference ($F(1, 93) = 4.579, p = 0.034$, effect size = 13.224) but the effect of presentation order is unreliable ($F(1, 93) = 1.811, p = 0.181$).

The model predicts an unreliable difference of 1.731 rounds between CONST and EHE ($p = 0.062$), an almost unreliable difference of 1.847 between CONST and HEH ($p = 0.048$), and a reliable difference of 2.208 rounds between EHE and HEH ($p = 0.015$).

### 3.3.1.2 Discussion

Differences in play counts between CONST and other two conditions suggest that CONST is most engaging. However, when controlling for differences in mandatory scores and presentation order, CONST no longer has an reliable advantage over EHE, although the trend is still in its favor. In both analyses, EHE has reliably greater play counts compared to HEH, which is consistent with existing work that found greater immersion in subjects in the “up-down” condition vs. the “down-up” condition (Qin et al., 2010). One explanation is that the subjects’ engagement is affected by their latest experience which is easy or positive in the EHE case and difficult or negative in the HEH case. This consistent with existing evidence that found that players prefer a difficulty curve that ends positively over one that ends negatively (Gutwin et al., 2016). In Experiment 3.4, I explore this recency effect in detail. Gutwin et al. (2016) also found no significant difference in player preferences between positive and “neutral” or constant conditions, which explains the lack of reliable difference between CONST and EHE when
controlling for differences in mandatory scores and presentation order. The difference in mandatory scores between CONST and the other two conditions suggests that performance depends on the peak difficulty, rather than the mean. Given that the difference in mandatory scores is a significant and positive predictor of play count differences, subjects are more likely to replay a condition if they achieve a greater score. This is supported by findings by Lomas et al. (2013a), where subjects found easier games to be more engaging than harder ones.

In Experiment 3.2, I use a different tension-and-release manipulation that does not affect difficulty as profoundly as the scroll speed does but still provides release. I also use a different paradigm than A/B testing where I consider the full design space of the manipulation, rather than just a few conditions.

3.3.2 Experiment 3.2: Tension and Release via Hard Intermissions

Rather than gradually changing difficulty over time via scroll speed, the manipulation in the second experiment introduces a sudden intermission period where objects stop appearing on the screen for 5 seconds. During this intermission period, the line is cleared and the rules are shown to the subject. Unlike the first experiment, subjects do not see the rules any other time during the round and the score is displayed as a progress bar during the round. The speed of the line is constant but the inter-object spacing decreases over time so difficulty is always increasing. Examples of designs with different intermission locations are shown in the left plate of Figure 3.7. Here the inter-object arrival times decrease linearly as a function of object number but not as a function of time because points on the abscissa get closer as inter-object arrival time decreases. All curves start at 1.5 seconds, which corresponds to the setup time before object appear on the screen. A simple step-wise controller that adjusts next round duration based on performance on the previous round switches between the five difficulty curves in the right plate of Figure 3.7 to ensure that the subject always scores in 70% to 80% range. If the score is less than 70%, the next easiest difficulty level is chosen and if the score is greater than 80%, the next hardest level is chosen. Rounds in the the hardest and easiest difficulty levels last for 28 and 45 seconds, respectively. The independent variable is the intermission location which takes values from 1 to 63 (after item 1 and after item 63).

As in Experiment 3.1, subjects get assigned pairs of conditions – intermission locations – but
Figure 3.7: The experimental manipulations in Experiment 3.2. The left plate shows, for intermediate difficulty, the log inter-object arrival times for a set of intermission locations. The large jumps correspond to the intermission period. The right plate shows the 5 difficulty levels in the game for a given intermission location. The curves are colored from easy (blue) to hard (red).

Figure 3.8: The post-round rating screen in Experiment 3.2.

they no longer choose which condition to play; instead, they play conditions in alternating order over 10 rounds. At the end of each round, subjects rate the round on a (0-10) scale that has a marker on the rating from the previous round (Figure 3.8). The rating screen also shows the subject’s score in the round he or she just played. Subjects are told during the pre-experiment orientation that they will play 10 “versions” of the game that differ in the pace of items on the line but they are not told that they will be comparing two conditions. A different background color is used for each round to give subjects the illusion of several versions in the experiment.

Engagement in this experiment is measured by the absolute and relative ratings – the differences in ratings of consecutive conditions. To model engagement, two Gaussian process (GP) models are fitted
to absolute and relative ratings (recall that each subject provides 10 absolute and 9 relative ratings). Absolute ratings are assumed to be normally distributed \( r(x) \sim N(f(x), \sigma^2) \) where \( f(x) \sim \text{GP}(c, \Sigma_{SE}) \) is the latent engagement associated with input \( x \), \( c \) is a constant mean prior, and \( \Sigma_{SE} \) is the squared exponential kernel. The preference-learning Gaussian process (Chu & Ghahramani, 2005b) is used to infer a latent engagement \( f(x) \) that is consistent with relative ratings, i.e., if subjects prefer condition \( x_a \) over \( x_b \), then \( f(x_a) > f(x_b) \). Rating differences are assumed to be normally distributed \( d(x_a, x_b) \sim N(f(x_a) - f(x_b), \sigma^2) \), where \( f \) is drawn from a zero-mean GP and the observation noise \( \sigma \) accounts for inconsistencies in the relative ratings. Markov chain Monte Carlo (MCMC) was used to fit the GPs with 10 chains, each running 1000 iterations with 500 burn-in. A Gamma(3, 1) prior is placed on the variance of the GP kernel and a Gamma(8, 1) prior is placed on the length scale of intermission location, reflecting the belief that small differences in intermission locations should not cause large differences in engagement (the expected value of Gamma(8, 1) is 8.0). Since the previous rating and score are both displayed on the rating screen, GP models with these two inputs are also used, with a length scale prior of Gamma(3, 1) placed on both input dimensions.

Subjects were randomly assigned – via Sobol sampling – to pairs of conditions, where a condition corresponds to an intermission location \( x \in (1, 63) \). The sampler was restricted so that conditions in any given pair are at least 8 positions apart because it does not make sense to compare nearly identical intermission locations. Subjects were required to play for about 10 minutes for a payment for 100 cents.

### 3.3.2.1 Results

A total of 98 subjects completed the experiment, providing ratings for 980 rounds. Mean rating is 3.90 (±0.089, range 0 – 10) and mean score is 0.71 (±0.003, range 0.00 – 0.97). Spearman’s correlations between (rating, intermission location), (rating, previous rating), and (rating, score) were 0.05, 0.72, 0.26, respectively.

Figure 3.9 plots model predictions of the absolute residual ratings as functions of intermission location, score, and previous rating. For a specific input dimension, residual ratings are computed by subtracting the predictions of a model of the other two inputs from the original absolute ratings. One dimensional GP models are then fitted to the residuals. So this procedure examines the variance explained.
Figure 3.9: GP model fits to the residuals in Experiment 3.2. The three plates show rating residuals as a function of intermission location, score, and previous rating. Residuals in each plate are computed by fitting a model of the other two inputs and subtracting its predictions from the original ratings. Orange circles and their color intensity correspond to observations and their counts, respectively. Blue lines are the predictions of the models and the shaded areas are the 95% confidence intervals of the mean.

by one input dimension, taking into account the variance explained by the other two dimensions. Only ratings from rounds 2 to 10 were considered (as previous rating is undefined on the first round), forming a dataset of 866 observations (the original number is 882 but ratings are ignored if score < 0.5). There is no effect of intermission location on the ratings. The current round’s score has a moderate effect as subjects are more likely to provide higher ratings as the score increases. The previous rating has a strong effect on the rating with an almost 1:1 relationship. These results agree with the Spearman’s correlation numbers between the absolute rating and each of the three inputs: no monotonic relationship between the intermission location and the rating ($\rho = 0.05$), a weak monotonic relationship between the score and the rating ($\rho = 0.26$), and a strong monotonic relationship between the previous rating and rating ($\rho = 0.72$).

The weak relationship between the score and the rating is unsurprising, given the adaptive round-by-round difficulty controller. Recall that the controller changes the next round’s duration – the difficulty – depending on the current round’s score. If a difficulty controller is operating perfectly, it will place subjects in the correct score range – 70% to 80% in this case – all time. So it is less likely that subjects will use the score as the differentiating factor when rating rounds. Figure 3.10 shows that the controller takes about 3 rounds to adapt to the subject’s skill, as the median line sits at about 75% from round 3 onwards, confirming that the controller is operating as intended.

Finally, a preference-learning GP model was fitted to the subjects’ relative ratings, where
observations from each subject were pooled into a dataset of 866 observations. MCMC settings from the previous experiments were used for training the model. Observations and predictions of the model are plotted in left and right plates of Figure 3.11, respectively. Pairs involving the same conditions, \((a, b)\) and \((b, a)\), are moved so that all observations lie in the lower triangular region, with blue and red colors indicating positive and negative rating differences, respectively. The plot fails to show a pattern in the distribution of relative ratings, a finding that is supported by the predictions of the preference model which lie in a region of \(0\) to \(-0.25\), which is small compared to the maximum possible region of \(-4.5\) to \(+4.5\). The bump at the high intermission locations might reflect a peak-end effect where subjects rate the condition higher because of the recency of the intermission period. However, a definitive conclusion cannot be made due to the high predicted variance.

### 3.3.2.2 Discussion

Analysis of absolute and relative ratings shows no effect of intermission location on engagement. The dominating factors on a subject’s ratings are the previous round’s rating and the current round’s score. I included the score in the post-round screen so subjects take the game seriously, while the rating anchor helps them judge relative to a reference point. But it is possible that hiding these pieces of information could yield more informative ratings. Also, the intermission manipulation might not be salient enough
to allow subjects to distinguish between different intermission locations reliably. For example, a longer round duration could be required so that there is a sufficient build up of tension before release. Another possibility is that the intermission location is not a manipulation of difficulty so all conditions are essentially the same.

### 3.3.3 Experiment 3.3: Tension and Release via Mean-Controlled Inter-Item Spacing

I revisit manipulating the rate of objects per second by adjusting inter-object arrival times, rather than the scroll speed. Both manipulations effectively change the rate of objects per second, but the latter handles discontinuities gracefully; sudden changes in scroll speed require smooth transitions to avoid choppy animation, whereas inter-object arrival times can have sharp transitions without affecting the animation. Two continuous dimensions are manipulated: the amplitude $\in (-1, 1)$ controls the template of the difficulty curve – the inter-object arrival times and their order – and the entropy $\in (0, 1)$ controls the randomness of the order of inter-object arrival times. Here, positive amplitudes correspond to tension-and-release curves. This two dimensional design space includes the baseline constant (amplitude $= 0$) and random conditions ($|\text{amplitude}| = 1$ and entropy $= 1$). Samples from this space are shown.
Figure 3.12: Design space of tension-and-release in Experiment 3.3. Blue circles are items on the line.

in Figure 3.12 which plots inter-object arrival times – the difficulty curves – as a function of time for 20 objects. In the game, a difficulty curve is repeated three times or cycles so rounds consist of 60 items and last for about 60 seconds. When entropy > 0, the template is resampled on each cycle. By resampling the set of inter-object arrival times for a given amplitude instead of randomly generating them, a direct comparison could be made between structured variability – tension-and-release – and unstructured variability. Positive and negative amplitudes differ in how they transition between the easy and hard phases: positive amplitudes gradually transition to the hardest difficulty but instantly jump to the easiest, while negative amplitudes gradually transition to the easiest difficulty but instantly jump to the hardest (when the cycle repeats). The numeric value of the amplitude controls how far the maximum and minimum inter-object arrival times are from the mean, which is the same in all designs – 750 ms.

Unlike Experiments 3.1 and 3.2, the three measures of engagement from Chapter 2 are used: persistence, projections of how long others will play, and a post-experiment engagement questionnaire. Recall that in Chapter 2, subjects play for a mandatory duration of time after which they are free to continue or quit. Persistence is characterized by the Voluntary Time on Activity (VTA), which is defined as the duration of time played after the mandatory time had elapsed and a new round had begun (e.g., if the mandatory time is reached in the middle of the round, persistence is 0 until the subject chooses to
play a new round). After completing the experiment, subjects are told how long they’ve play voluntarily, the VTA, and are asked to project how long others on Mechanical Turk would play on average, thereby providing the *projected VTA*. Subjects also fill-out the same post-experiment engagement questionnaire as in Chapter 2, which contains six true/false questions formulated such that true indicates greater engagement. Finally, the post-experiment screen also asks subjects to rate their performance and the difficulty of the game on a 5-point scale.

The engagement measures from Chapter 2 require subjects to play one condition only, as opposed to two, thereby shortening the experiment time and allowing collection of data from more subjects. If the mandatory time does elapse in the middle of a round, a “Finish HIT” button appears in the bottom right corner for subjects to click on. Instead of scores, subjects can make up to 3 mistakes before the round is over, at which point they can start a new round or finish the hit if the mandatory time had elapsed. The remaining number of mistakes that can be made is displayed at the bottom of the line using gold circles. Before the first round, subjects play a practice round, where mistakes aren’t penalized, for about 25 seconds; during the round, the number of mistakes is displayed turns red if it exceeds 3. The practice round is sufficiently long to contain one cycle of the difficulty curve so subjects have an idea of the manipulation (without a practice round, subjects could repeatedly fail rounds before experiencing a full cycle of the difficulty curve).

Gaussian processes with non-conjugate observation models are used to model engagement as a function of the two dimensional design – as in Chapter 2. VTA and Projected VTA at design $x$ are modeled as $V(x) = C(x)T(x)$ where $C(x) \sim \text{Bernoulli}(p_{\text{continue}})$ indicates whether the subject continues playing beyond the mandatory time, $T(x) \sim \text{LogN}(f(x), \sigma^2)$ is the voluntary time played, and $f(x) \sim \text{GP}(c, \Sigma_{SE})$ is the latent engagement associated with design $x$. The probability of continuing also depends on the latent engagement as $\text{logit}(p_{\text{continue}}) = \beta_0 + \beta_1 f(x)$, where $\beta_0$ and $\beta_1$ are free parameters. Questionnaire scores are modeled with a Binomial distribution, $Q(x) \sim \text{Binomial}(N = 6, p_{\text{true}})$, where $\text{logit}(p_{\text{true}}) = f(x)$. MCMC is used to fit the models with 10 chains, each running 1000 iterations, and 500 burn-in. All three parameters of the covariance kernel, length scales and variance, have a Gamma$(3, 1)$ prior.

Subjects were randomly assigned via Sobol sampling to points in the two dimensional design.
<table>
<thead>
<tr>
<th>Measure</th>
<th>Median (std)</th>
<th>Range</th>
<th>Zeros (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full VTA (sec)</td>
<td>20.12 (78.37)</td>
<td>0.04 - 643.32</td>
<td>0.00</td>
</tr>
<tr>
<td>VTA (sec)</td>
<td>0.00 (75.67)</td>
<td>0.00 - 628.57</td>
<td>63.93</td>
</tr>
<tr>
<td>Projected VTA (sec)</td>
<td>20.00 (49.85)</td>
<td>0.00 - 405.00</td>
<td>16.43</td>
</tr>
<tr>
<td>Questionnaire</td>
<td>2.00 (1.67)</td>
<td>0.00 - 6.00</td>
<td>13.93</td>
</tr>
<tr>
<td>Difficulty</td>
<td>3.00 (0.78)</td>
<td>1.00 - 5.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Practice Mistakes (%)</td>
<td>11.36 (13.15)</td>
<td>0.00 - 50.00</td>
<td>10.36</td>
</tr>
<tr>
<td>Mean Completion %</td>
<td>66.67 (23.60)</td>
<td>16.81 - 100.00</td>
<td>0.00</td>
</tr>
<tr>
<td># Voluntary Rounds</td>
<td>0.00 (3.22)</td>
<td>0.00 - 32.00</td>
<td>63.93</td>
</tr>
</tbody>
</table>

Table 3.2: Descriptive statistics of Experiment 3.3. Full VTA is the voluntary time, measured immediately after the mandatory time had elapsed.

space. They were paid 20 cents if they play a minimum of 85-95 seconds. The required play duration was randomized for each subject between 85 to 95 seconds to eliminate unwanted interaction with the rounds, otherwise the mandatory time could always pass at a specific point in the round.

3.3.3.1 Results

A total of 361 subjects completed experiment but a severe Firefox browser animation stuttering bug required the elimination of observations from 78 subjects plus 3 additional subjects due to data corruption so observations from 280 subjects are used for analysis. Table 3.2 shows descriptive statistics of engagement and other interesting measures. Some subjects play the game for 11 minutes – a massive increase over the mandatory 90 seconds – indicating that they found the game engaging enough to forgo working on other hits. Mean completion % – how far into the round the subject makes it – of 66.67% indicates that subjects are going through two cycles of the difficulty curve on average before failing the round. Spearman’s correlations among the three measures of engagement are moderate: $\rho = 0.58$ between VTA and projected VTA, $\rho = 0.33$ between VTA and the questionnaire score, and $\rho = 0.36$ between the projected VTA and the questionnaire score. Reported difficulty also correlates moderately with the number of practice mistakes ($\rho = 0.46$), mean completion % ($\rho = -0.51$), and number of rounds played ($\rho = 0.47$). But there is no correlation between reported difficulty and the three measures of engagement.

Figure 3.13 plots model predictions of engagement and difficulty in the top and bottom rows, respectively. Here color indicates the magnitude of the prediction: yellow and dark purple regions correspond to high and low predictions, respectively. The circles correspond to observations and their
radii indicate the magnitudes. Crosses are observations with zero value. Since it is hard to visualize prediction variance in these two dimensional spaces, whiteness is used to convey uncertainty with white regions indicating maximum uncertainty. Specifically, the color \( c(x) \) at a design \( x \) is computed with \[ c(x) = (1-U(x))c'(x) + U(x)[1, 1, 1], \] where \( U(x) \) is the normalized uncertainty and \( c'(x) \) is the original unmanipulated color. The normalized uncertainty is defined as \[ U(x) = \min \left( 1, \frac{\sigma(x)}{\max(\hat{f}(x)) - \min(\hat{f}(x))} \right), \] where \( \hat{f}(x) \) and \( \sigma(x) \) are the model’s prediction and the prediction variance at \( x \), respectively. The predictive standard deviation is normalized with respect to the overall range of the predictions so there is a scale against which the standard deviation is measured. If the model is not confident of its’ predictions \( U(x) \) will increase, i.e., whiteness will increase.

VTA predictions in Figure 3.13 indicate an optimal region closer to the middle of the space, without much of an effect of entropy, where the optimum design is predicted to be played 48% longer than the small predicted regions. No reliable differences exist over the design space in the predictions of projected VTA and questionnaire scores. Two difficulty measures—practice mistakes % and reported difficulty—indicate that the constant designs (amplitude between -0.2 and 0.2) are the easiest. Finally, there are no reliable differences in the mean completion rate across the design space.

Designs in high-entropy regions should be symmetrical with respect to amplitude (Figure 3.12) but predictions in the top row of Figure 3.13 do not reflect this, possibly due to lack of data. Folding the design space by considering the absolute value of the amplitude is one way to (a) increase the amount of observations available at non-zero amplitudes and (b) eliminate the symmetry constraint. Figure 3.14 plots the predictions of models fitted using the absolute value of the amplitude, with the left half of the space being a flipped image of the right to enable easier comparison to Figure 3.13. VTA predictions indicate an optimum close to the middle of the space with not much of an effect of entropy. Questionnaire predictions indicate that difficulty curves with medium to high absolute amplitudes are most engaging, regardless of entropy. There is stronger evidence that subjects make more mistakes during practice on high absolute amplitudes, possibly due to the greater peak difficulty. Reported difficulty predictions are less reliable but they show that subjects perceive constant designs to be the easiest. Finally, there are no reliable differences the actual difficulty (the mean completion %).
Figure 3.13: Model predictions for each of the three engagement (top row) and difficulty (bottom row) measures in Experiment 3.3. The colors correspond to model predictions. Circles and crosses correspond to non-zero and zero observations, respectively. Circle areas are proportional to the magnitudes of observations. The lightness of the color at any cell corresponds to the predicted standard deviation at that cell, divided by the overall range of predictions. Lighter regions indicate higher uncertainty.

3.3.3.2 Discussion

VTA predictions using signed and absolute amplitudes indicate little to no effect of entropy on engagement and that the optimal difficulty curves have small absolute amplitudes. Taken together with the predictions of two difficulty measures (practice mistakes % and reported difficulty) that show that the small absolute amplitude designs are the easiest, this suggests that subjects find the easiest conditions the most engaging. In other words, subjects may be playing near-constant designs the longest because these designs have a lower peak difficulty (minimum inter-object arrival time). However, questionnaire scores fitted using the absolute amplitudes show an opposite trend with subjects rating medium to high absolute amplitude designs the highest. So the behavioral measurement shows preference for easy designs while the subjective measurement shows preference for challenging designs. To see whether this is always the case, in the next experiment I extend the template duration and control for peak difficulty instead of average difficulty, making the constant designs the hardest.
Figure 3.14: Model predictions for each of the three engagement (top row) and difficulty (bottom row) measures in Experiment 3.3, with symmetry enforced by using the absolute value of the amplitude. The colors correspond to model predictions. Circles and crosses correspond to non-zero and zero observations, respectively. Circle areas are proportional to the magnitudes of observations. The lightness of the color at any cell corresponds to the predicted standard deviation at that cell, divided by the overall range of predictions. Lighter regions indicate higher uncertainty.

3.3.4 Experiment 3.4: Tension and Release via Peak-Controlled Inter-Item Spacing

I modified the design space from Experiment 3.3 to control for peak difficulty (Figure 3.15). Round duration is kept constant so the number of objects is allowed to vary. The constant condition is most difficult, with 100 objects in a single template. As the absolute value of the amplitude increases, the number of objects decreases, but the peak difficulty – the minimum inter-object arrival time – stays the same. Since a template lasts for 30 seconds, a game round contains 2 cycles of the difficulty curve, instead of 3. This decision was made to ensure that the build-up in tension is not too short. Subjects no longer see the number of mistakes they have remaining, in order to keep them focused and to increase their uncertainty about how many mistakes they can make. All other aspects of the experiment are identical to Experiment 3.3.
3.3.4.1 Results

A total of 535 subjects completed the experiment but observations from 40 subjects were discarded due to data corruption and the Firefox animation bug that affected Experiment 3.3, making the effective number of subjects 495. The Firefox animation bug was discovered and corrected during the experiment so Firefox users who participated after the fix are included in the experiment. Descriptive statistics are shown in Table 3.3. Despite having a high peak difficulty of 200 ms, some subjects play the game for 12 minutes. The median VTA and projected VTA are both lower than in experiment 3.3 but the reported difficulty is slightly greater. The mean completion rate is much lower than in Experiment 3.3, consistent with high peak difficulty. Spearman’s correlations among the three engagement measures are moderate: $\rho = 0.54$ between VTA and projected VTA, $\rho = 0.47$ between VTA and questionnaire scores, and $\rho = 0.43$ between projected VTA and questionnaire scores. Reported difficulty also correlates moderately with the number of practice mistakes ($\rho = 0.34$), mean completion rate ($\rho = -0.55$), and number of rounds played ($\rho = 0.44$).

Model predictions on each of the three engagement and difficulty measures are shown in Figure 3.16. There are no reliable differences across the design space in the predictions of VTA and projected

Figure 3.15: Design space of tension-and-release in Experiment 3.4. Blue circles are items on the line.
<table>
<thead>
<tr>
<th>Measure</th>
<th>Median (std)</th>
<th>Range</th>
<th>Zeros (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full VTA (sec)</td>
<td>13.74 (62.28)</td>
<td>0.03 - 716.94</td>
<td>0.00</td>
</tr>
<tr>
<td>VTA (sec)</td>
<td>0.00 (61.43)</td>
<td>0.00 - 712.51</td>
<td>57.98</td>
</tr>
<tr>
<td>Projected VTA (sec)</td>
<td>15.00 (55.47)</td>
<td>0.00 - 600.00</td>
<td>16.57</td>
</tr>
<tr>
<td>Questionnaire</td>
<td>2.00 (1.69)</td>
<td>0.00 - 6.00</td>
<td>17.78</td>
</tr>
<tr>
<td>Difficulty</td>
<td>3.50 (0.80)</td>
<td>1.00 - 5.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Practice Mistakes (%)</td>
<td>16.25 (13.39)</td>
<td>0.00 - 51.85</td>
<td>1.41</td>
</tr>
<tr>
<td>Mean Completion Rate (%)</td>
<td>36.23 (23.37)</td>
<td>8.37 - 100.00</td>
<td>0.00</td>
</tr>
<tr>
<td># Voluntary Rounds</td>
<td>0.00 (3.48)</td>
<td>0.00 - 41.00</td>
<td>57.98</td>
</tr>
</tbody>
</table>

Table 3.3: Descriptive statistics of Experiment 3.4.

VTA. Subjects rate low-amplitude, low-entropy designs the highest, an increase of about 33% in questionnaire score over the least engaging regions. Predictions of the three difficulty measures all agree that the constant difficulty curves are most difficult, with predictions on the mean completion % showing the greatest advantage. However, those predictions only serve to validate the inference procedure because it is expected that the mean completion % will be higher on designs with fewer objects. The predictions on the other two difficulty measures still show that both practice mistakes % and reported difficulty are higher in the near-constant designs.

The preference for negative amplitudes is rather surprising as one would think that positive (tension-and-release) amplitude designs would be preferred over negative amplitude designs. Both positive and negative amplitude difficulty curves end with an easy period of about 8 seconds to provide release. However, the negative amplitude difficulty curve progressively eases into the final release period whereas the positive curve abruptly switches from hard to easy, so it is possible that subjects’ ratings of engagement are primarily influenced by recency effects. To explore this possibility, an exponential weighted average (EWA) of inter-item arrival times in the last round played is calculated for each subject. The EWA calculates an average of inter-arrival times by weighing recent arrival times higher than older ones. Specifically,

$$EWA = \sum_{i=1}^{N} \omega_i L_i$$

$$\omega_i \propto \alpha(1 - \alpha)^{N-i}$$

where $N$ is the number of items processed in the round, $\alpha \in (0, 1)$ is the decay factor, $L_i$ is the lag leading up to the $i$th item, and $\omega_i$ is the weight associated with the $i$th item. As $\alpha$ increases, more
Table 3.4: Spearman’s correlation between the exponential weighted average (EWA) of inter-arrival times in the last round and questionnaire ratings. The EWA is calculated for several values of decay $\alpha$ and the median EWA across all subjects is shown in the second column. The correlation is shown in the third column.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Median EWA (ms)</th>
<th>Questionnaire corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>543</td>
<td>-0.02</td>
</tr>
<tr>
<td>0.20</td>
<td>490</td>
<td>-0.03</td>
</tr>
<tr>
<td>0.35</td>
<td>465</td>
<td>-0.02</td>
</tr>
<tr>
<td>0.50</td>
<td>449</td>
<td>-0.01</td>
</tr>
<tr>
<td>0.80</td>
<td>428</td>
<td>-0.01</td>
</tr>
<tr>
<td>0.95</td>
<td>419</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

weight will be placed on recent items. The number of items processed $N$ varies depending on how far the subject got into the round. The weights are normalized to sum to 1, but the findings are similar in the unnormalized case. Under this scheme, the EWA for $amp = -1$ and $amp = 1$, when entropy = 0 and $\alpha = 0.1$, is 824ms and 684ms, respectively. Table 3.4 shows the median EWA and correlation with questionnaire scores for several values of $\alpha$. Regardless of the value of $\alpha$, no correlation exists between the EWA and questionnaire scores. It is therefore unlikely that subjects are rating conditions based on recent difficulty.

Engagement and difficulty as functions of the absolute values of amplitudes are plotted in Figure 3.17. VTA and questionnaire predictions both agree that high absolute amplitudes are optimal, with the former showing less engagement in deterministic designs (low-entropy) than in high-entropy designs. In terms of questionnaire scores, entropy does not have much of an effect although the optimum is shifted towards low-entropy designs. Interestingly, questionnaire predictions along the spine or the middle of the design space are not identical, as one would expect when the amplitude is 0. This could be due to a high level of noise in the observations. No reliable differences exist in the predictions of projected VTA across the design space. Predictions of the three difficulty measures are consistent with the signed amplitude analysis (where no absolute values were taken), with subjects finding the lowest absolute amplitudes to be the most difficult.

Given how subjects are rating high absolute amplitudes (the easiest conditions) to be the most engaging and taking into account evidence from Lomas et al. (2013a), where subjects found easier designs to be most engaging, one might ask how difficulty affects engagement. To answer this, the latent predictions of a GP model of questionnaire scores as a function of reported difficulty are shown in Figure
Figure 3.16: Model predictions for each of the three engagement (top row) and difficulty (bottom row) measures in Experiment 3.4. The colors correspond to model predictions. Circles and crosses correspond to non-zero and zero observations, respectively. Circle areas are proportional to the magnitudes of observations. The lightness of the color at any cell corresponds to the predicted standard deviation at that cell, divided by the overall range of predictions. Lighter regions indicate higher uncertainty.

3.18. A reliable downward trend starts from a reported difficulty of 3.0, indicating that subjects find easier conditions more engaging on the questionnaire. Recall that the GP model of questionnaire scores uses a Binomial distribution to characterize the observations, with the probability of success directly proportional to the latent engagement, $\text{logit}(p) = f(x)$. So at the maximum and minimum points, the curve predicts questionnaire scores of 1.29 and 2.84, respectively. Only 18% of the observations have a reported difficulty $\leq 3$ which explains the greater uncertainty in that region and the fact that reported difficulty predictions in Figure 3.16 starts at around 3.0.

3.3.5 Discussion

Signed amplitude analysis shows no reliable differences in the predictions of VTA and projected VTA while questionnaire predictions indicate that subjects prefer low-entropy, low-amplitude designs. One explanation for the preference for low-entropy designs is the increase in the duration of a difficulty curve template from 20 to 30 seconds. Surprisingly, subjects rate negative amplitude designs, the opposite of tension-and-release, higher than positive amplitude designs. This finding is not explained by the
difference in recent difficulty of the two conditions. One possible explanation is that subjects want to get through the hardest phase early in the game, thus favoring negative amplitude curves. If this is true, it suggests that subjects view the game as more of a chore—an unpleasant but necessary task\(^1\)—than a fun activity. The preference for negative amplitudes may appear to contradict evidence from Experiment 3.1, where the easy-hard-easy condition was found to be more engaging than hard-easy-hard. The two experiments differ on several fronts: round duration, the difficulty curves, reward structure, etc., which may produce differences among the two experiments. But, I think the most importance difference is the engagement measurement: in Experiment 3.1, subjects have the opportunity to choose the condition they wish to play, which they cannot do in Experiment 3.4. So subjects in Experiment 3.1 can compare their experiences in the two conditions; a subject may think “this condition is bad but the other one is a lot worse!”\(^1\). It is therefore plausible that in the absence of an “anchor” to compare against, the release-and-tension difficulty curves come out superior to tension-and-release curves. In the conclusion of this chapter, I suggest an approach which may alleviate the problem of which engagement measure to

\(^1\)Dictionary definition from Google
Both VTA and questionnaire predictions agree that high absolute amplitudes are most engaging. Given the robust evidence that constant designs are the hardest—as one would expect—the VTA predictions are consistent with Experiment 3.3, where subjects played the easiest conditions the longest. However, questionnaire predictions from the absolute amplitude analyses in Experiments 3.3 and 3.4 consistently show that subjects prefer high absolute amplitude designs, despite the fact that such designs were the hardest in the former and the easiest in the latter. Of course, it could also be that subjects rate the easiest designs the highest, as illustrated in Figure 3.18, but that does not explain why they rated the hardest designs the highest in Experiment 3.3. A more plausible explanation is that subjects prefer the one thing that medium to high amplitude conditions in both experiments share: variability.

3.4 Conclusion

In this chapter I developed a custom web-based game that aims to address the shortcomings of existing working memory tasks. In order to make the game engaging, I studied four tension-and-release manipulations systematically. Whereas existing literature only treats tension-and-release in video games qualitatively or quantitatively evaluates a few conditions, I use Bayesian non-parametric modeling tools to study a continuous range of designs quantitatively. The first study proved that engagement can be modulated by adjusting the difficulty curve over time, while keeping mean difficulty constant. The second study inserted hard release periods in the midst of a constantly increasing difficulty. The locations
of these release periods had no effect on engagement which could be due to the short round duration in ShapeFactory. The third and fourth studies compared designs in a continuous design space that modulated the amplitude and randomness of difficulty curves, with the former experiment controlling for mean difficulty and the latter controlling for the peak. Both experiments showed no reliable differences in the predictions of projected VTA across the design space. In the third experiment, absolute amplitude analysis indicated that subjects played the easiest designs—the near-constant difficulty curves—the longest and they subjectively rated the most challenging designs—high-variability curves—the highest. The same analysis in the fourth experiment showed that subjects played the easiest designs—the high-variability curves—the longest, but they also rated the same designs the highest. Taken together, results from both experiments suggest that subjects rate difficulty curves that have variability the highest, but play the easiest conditions the longest. Signed amplitude analysis of the fourth experiment showed that subjects preferred low-entropy negative amplitude designs. The conventional design wisdom is to use tension-and-release to increase engagement, so a result favoring the opposite difficulty curve is unexpected. Possible explanations include inadequate control for recency effects and the game not being fun enough. Consistent with prior research, when peak difficulty was controlled, subjects found the easiest games to be most engaging.

For future studies, I would like to (a) make ShapeFactory intrinsically more engaging by introducing extra variability beyond simple randomization of rules and objects, (b) extend the round duration to enable proper buildup of tension, and (c) use a hybrid design that combines multiple measures of engagement; for example, VTA and preference-learning setups could be combined so that subjects play two conditions for a mandatory period of time after which they would be free to select and play either condition.
Chapter 4

The Near-Win Effect

4.1 Introduction

The near-win effect is a manipulation that has been studied extensively in gambling addiction literature. This effect occurs when a player almost wins a game, e.g., getting two cherries and a lemon in slot machine (Figure 4.1). Reid (1986) cites several other examples of this effect in real-life: near-win number sequences in lottery tickets, close finishes in horse racing, and customers thinking they are half-way to success when completing half the winning sequence in a promotional sales campaign. Reid also cites a study that found that players who experienced an early loss were less likely to continue playing voluntarily than those who experienced a near win (Reid’s own study also showed a trend in that direction). Reid argues that near-win events are useful in skill-based games, such as darts, because they provide feedback that winning is close. Even though such feedback is useless in games of pure chance, Reid notes that gamblers still think they can influence in the outcome with behaviors such as whispering to the dice, choosing lottery numbers carefully, or consulting books of lucky numbers.

A complimentary manipulation that often gets studied alongside the near-win effect is anticipation. Studies on the near-win effect usually find differences in engagement between losing early and nearly-winning. The study cited by Reid (1986) is one example where subjects preferred an near-win over an early loss. Video game developers also use anticipation, usually in scenarios with “close calls”.
For example, a scene from Half-Life 2 (Figure 4.2), a first-person shooter, shows a large chimney slowly falling down to block the path of the player. The scene is gripping because all the player can do is press full throttle – to escape pursuers – and hope that she can clear the falling chimney. Ultimately, the chimney falls before the player gets to it so she has to lift off the throttle and maneuver around the rubble. If the scene was different, say with the chimney already in ruins, it would not be as captivating.

In this chapter I explore engagement as a function of the two-dimensional design space of near-win probability and anticipation. I implement this design space in a large-scale math-learning software used by thousands of students.
4.1.1 Related Research

Engagement as a function of the frequency of near-win events was investigated in the context of slot machine gambling (Kassinove & Schare, 2001). Engagement was defined as the number of trials subjects voluntarily played, after completing a mandatory minimum number of trials. A near-win was defined as “the occurrence of three out of four of the same number with the last number being different”. Subjects were assigned to one of 2x3 conditions corresponding to (early big win x near win frequency). In the big win condition, subjects received a big payoff on trial 8. Three near-win frequencies were evaluated: 15%, 30% and 45%. Subjects played for 50 mandatory trials after which they could continue playing for as long as they wanted. The winning rate during the mandatory trials phase was held constant at 10%. There were no wins, near-wins or payoffs during the voluntary play phase. Results showed no effect of the early big win on engagement. However, subjects who had 30% near wins played almost twice as long as those in the 15% and 45% near-win conditions. These results show that carefully choosing the near-win rate is crucial as high rates lead to desensitization, i.e., crying wolf too many times, and low rates do not have enough impact.

Lomas (2014) studied the effect of close/far losses/wins on engagement in an educational fraction learning game used by thousands of students. The game asks players to estimate the position of a fraction or number by clicking on a number line representing the range of values. On the left and right edges of the screen there are posts indicating the minimum and maximum values of the number line, respectively. Students played levels of the game – sets of exercises – of increasing difficulty and after each level they were presented with a scorecard. A large factorial design was used to evaluate the effect of several manipulations on engagement, defined as the number of exercises completed by students. A “goals present” manipulation controlled whether students see the target percentage of exercises required to win a level during the game. When the manipulation was active, students also saw the percentage of exercises they solved correctly on the scorecard, accompanied by a message telling them whether they had won or lost. To control for student skill, another manipulation was used to specify the target criterion. Ordinarily, the game has a menu that allows students to select a difficulty level to play. One manipulation controlled whether harder levels were locked, if the goals manipulation was active, so players could only progress to harder levels if they had won easier ones first. Subjects attempted 43% more exercises over a
control condition, which did not show scorecards or allow selection of difficulty levels, when the goals were on and the criteria for success was 80%. Players who had goals, locked levels, and achievements shown were the most engaged, with high ability players particularly showing the most engagement. Analysis of the effect of close winning or losing was done post-hoc by looking at the number of items played after the first two levels (limited to players who had locked levels and goals on). Players who had a close loss on the first two levels (which are the easiest) played double the number of exercises than those who did not have a close loss and players who had a close win also played more exercises than those who had blow-out wins, but the effect was smaller – about 3 exercises more. It is unclear how "closeness" was measured in these results, but other analysis did show that the number of additional exercises played also increased as the absolute difference between the target criterion and the actual score decreased, even when controlling for player score, providing the famous inverted U shape. The author argued that these results indicated that the inverted U was a result of the drama of nearly winning or losing, rather than a moderate amount of challenge, as is often hypothesized by theories of flow.

The near-win manipulation in this chapter modulates the rate of near-wins (almost) independently of student performance, enabling assessment of the optimal rate of near-wins. Findings by Lomas (2014) suggest that showing students the target criteria created anticipation and subsequently led to greater engagement. I investigate this by using animation as the means of controlling how much anticipation occurs over time. In other words, rather than exploring how early losing influences engagement – like gambling studies do – I look at engagement as a function of the distribution of anticipation over time.

4.2 Application and Manipulations Overview

I collaborated with the developers of Wootmath (Wootmath, 2017), an interactive web-based fraction learning software used by thousands of students, to implement the near-win and anticipation manipulations. Students practice a series of lessons where each lesson consists of a set of exercises that are chosen dynamically depending on the student’s performance (Figure 4.3). Examples of these exercises include comparing fractions, placing decimals on the number line, adding and removing fractions, multiplication, etc. After every lesson, a scorecard is shown with a performance bar indicating the score and 3 goalposts.
corresponding to the number of stars the student can earn (Figure 4.4). An animation function is used to fill the performance bar and the three stars below it up to the student’s score. Replay and continue buttons on the scorecard allow the student to retry or go back to the lesson selection screen where he or she can select another lesson or quit. Not all lessons are available to the student immediately as a minimum score of 1 star is required to clear a lesson and move on to the next.

The near-win and anticipation manipulations are both implemented in the scorecard screen. The near-win manipulation has a single parameter \( \nu \in (0.1, 0.9) \), which characterizes the probability of a near-win event. Near-win events are instances where the score is within 0.1 of the next goalpost, e.g., a score of 1.95 stars is within (1.9, 2.0) stars so it is classified a near-win. Rather than relying on the natural frequency of students getting to within 0.1 of the next goalpost, the near-win manipulation artificially induces a near-win event according to the probability \( \nu \). When an artificial near-win is induced, the student’s score is increased to within 0.1 of the next goalpost, so the displayed score \( S = U(\lceil R \rceil - 0.1, \lceil R \rceil) \), subject to the original score not being within 0.1 of any goalpost; so that students who just made it over one goalpost don’t get suddenly boosted all the way to just below to the next goalpost. If the student is already within 0.1 of the next goalpost, the near-win manipulation is excluded. In other words, under no circumstance does the score get boosted beyond the original number of stars. If the student scores over 3 stars by solving bonus exercises, no near-win events are induced. It is important to note that, contrary to gambling literature, the rate of wins is not controlled as it is determined by the student’s skill.

Anticipation is manipulated by changing the ease-out magnitude \( \eta \in (0, 1) \) which controls
the deceleration of the animation as it approaches the target score. Figure 4.5 plots the score in the performance bar as function of time for different ease-out magnitudes and target scores. All curves are generated using $y(t) = Sb(t/T)$ where $S$ is the target score, $T = 0.3 + S$ is the animation duration, and $b = \text{Bezier}(0, 0, \eta, 1, 1, 1)$ is a unit cubic Bezier function. Animation duration depends on the score because otherwise the mean speed would change, enabling the student to estimate the target score early on. Small $\eta$ values produce a linear animation which stops at the target score suddenly so the student is maximally uncertain about when animation will stop. On the other hand, large $\eta$ values produce an animation which quickly jumps to near the target score and then slowly approaches the target score. I hypothesize that $\eta = 1$ will produce more anticipation than $\eta = 0$ because in the case of near-wins, the first creates more anticipation as the animation slows down while the latter distributes anticipation evenly over multiple goalposts. To see this in action, consider a simple hypothetical model of the student where anticipation linearly increases if the score in the performance bar is within 0.25 units of the next goalpost, otherwise the anticipation is set to zero (Figure 4.6). Here, the target score is set to 2.90 and the figure plots how anticipation changes over time for $\eta = 0$ (left) and $\eta = 1$ (right). Given the same score, both conditions will animate over the same duration, but there are clear differences in how anticipation builds-up. When $\eta = 1$, animation clears the first two goal posts quickly, causing very little anticipation, but then it slows down and spends of most the time within the 0.25 region of the next
Figure 4.5: The anticipation manipulation. Each plate plots the score in the performance bar as a function of time. Different curves correspond to different target scores.

Figure 4.6: Simulation of a simple model of anticipation using different animation functions. The magenta lines indicate when the progress in the performance bar clears a goalpost. The area under the curve is shown for both cases.

goalpost, causing a large build-up in anticipation.

The engagement measure in this experiment is whether the student replays the lesson, if he or she hadn’t scored perfectly. The experiment was conducted over a period of 3 months from March 2017 to June 2017 on the Wootmath platform. Students were randomly assigned to designs \((\nu, \eta)\) in the two-dimensional design space with no student receiving two different designs. Every time a student completed a lesson, a log entry was added to the dataset with information including the design, the original score, whether a near-win was triggered, and whether the student replayed, continued, or quit afterwards. I limit the analysis to entries in the dataset where the score was less than 3.0. The analyses in the next section will apply additional filters on the dataset, which are discussed in the corresponding sections.
Students typically practice over multiple sessions but the analyses in this chapter do not take
the session into account. In other words, trials from multiple sessions are concatenated to form one
sequence. The analyses does not assume however, that the student will do the concatenated sequence
all at once. In other words, I still keep track of when the student quits at the end of a session.

4.3 Results

The original dataset consists of 129,214 entries from 7,976 students. Median number of lessons attempted
by the students is 9.00 lessons (std. 22.00, range 1-311), over 3.00 sessions (std. 4.83, range 1-51), with
a lesson duration of 2.8 minutes (std. 11.00, range 0.15-1,441.60). After excluding perfect and zero
scores and instances where the student moved on before the progress bar animation had completed ¹,
the dataset reduces to 29,470 entries from 5953 students. The median number of lessons drops to 3.00
lessons (std. 6.54, range 1-124), over 2.00 sessions (std. 3.20, range 1-39), with a lesson duration of 3.8
minutes (std. 9.64, range 0.31-408.88).

4.3.1 Does the Displayed Score Affect Engagement?

The first and most basic question that can be asked is whether the score shown to the student—the
displayed score— affects engagement. Entries where the student’s original score happens to be a near-
win are included in the analyses because this section studies the relationship between the displayed
score and engagement. Because it is unclear which threshold constitutes a near-win from the player’s
perspective, Figure 4.7 plots the replay probability as a function of displayed score using two thresholds,
0.75 and 0.9. For example, a threshold of 0.75 indicates that all events within 0.25 of the next goalpost
are near-wins (orange bars). Bars heights correspond to the mean replay probability of observations that
fall within each bin. Both binning schemes show a clear downward trend in mean replay probability as
the score increases, with lower-performing students (those scoring in the 0.75-1.00 and 0.90-1.00 bins)
more likely to replay when they experience a near-win event. So Figure 4.7 supports the hypothesis

¹This information is not explicitly provided so I infer it by measuring how long the student stayed before continuing
or quitting. If the duration is less than the animation duration, the entry is excluded. The dataset does not explicitly
provide an animation duration field so it is computed, as described earlier in the Chapter, via \( T = 0.3 + S \), where \( S \) is the
displayed score. It turns out that this filter also eliminates all entries where the student quit the program afterwards. In
other words, students who quit do so immediately, without waiting for the animation to finish.
that the displayed score modulates engagement and also indicates that near-win events increase the engagement of lower-performing students.

4.3.2 Does a Near-Win Increase Engagement for a Given Score?

Since near-wins are induced (almost) independently of performance, the data in Figure 4.7 can be binned according to the original unmanipulated score and split into near-win and no-near-win events. In other words, each bar in Figure 4.7 is dissected into near-win and no-near-win events. A direct comparison of engagement can then be made between near-win and no-near-win events at a given score. Figure 4.8 plots the probability of replay given the original score bin and the occurrence of a near-win. Only one binning threshold, 0.75, is considered here because a threshold of 0.9 would show empty bars for no-near-win events in bins within 0.1 of the next goalpost (since near-wins are defined by our manipulation to be within 0.1 of the next goalpost). Similarly to the previous analysis, when students score in the 0.75-1.00 bin, they are more likely to replay if they experience a near-win event.
Figure 4.8: Analysis of replay rates as a function of the original score and the occurrence of the near-win. Each bar corresponds to a score range \([a, b)\), with \(a\) included and \(b\) excluded. Bar heights indicate the mean replay probability of observations in the corresponding score range. Blue and orange bars correspond to no-near-win and near-win events, respectively. Error bars correspond to the 95% confidence interval for a binomial proportion, using Wilson’s score interval.

The coarse level of analysis in Figure 4.8 may obscure finer-level trends in the data so a regression model was selected and fitted to the dataset to predict whether the student replays the lesson or not. This eliminates the issue of choosing a specific bin size as the regression model will use all the available data to infer the relationship between input features and whether the student replays a lesson or not.

Three features were considered: the score bracket \(B\), the original distance to the next goalpost \(O\), and the square of the distance \(O^2\). The model predicts whether the student replays a lesson or not, so it is a logistic regression model with

\[
\text{logit}(\rho_{\text{replay}}) = f_{\text{model}}
\]

where \(f_{\text{model}}\) is the regression equation. Since logistic regression is a parametric model, a specific functional form of the model is required to do the regression, i.e., relationships among different features have to be specified explicitly. There is no strong prior belief about these relationships so I considered models that include all permutations of the three features mentioned as well as a feature indicating whether the player saw a near-win or not (here, near-wins are defined to be scores within 0.1 of the next goalpost, which is the manipulation’s assumption).

Each model contains all main and interaction effects of the included features, e.g., if a model has 3 features, it would have \(\binom{3}{1} + \binom{3}{2} + \binom{3}{3} = 7\) terms in the regression equation plus the bias term. To select
a regression model, a total of 8 models were evaluated via 10 replications of 5-fold cross-validation on the dataset. The model that includes all the features \((B, O, O^2)\) performs the best in terms of test log-likelihood as indicated by Figure 4.9, which shows the negative test log-likelihood of each model (lower is better) along with the standard error bars. The star bracket \(B\) is the most significant feature and including it in any model brings out the predictive power of the distance \(O\) to the goalpost and its’ square \(O^2\), implying an interaction among those three features.

Figure 4.10 plots the predictions of the best model found via cross-validation, \((B, O, O^2)\). Orange and blue curves correspond to near-win and no-near-win predictions, respectively. If the near-win manipulation is working perfectly, we expect to see complete dissociation between the original score and the probability of replay in case of near-win events. In other words, if students are not aware of their own performance then the near-win predictions would be flat, indicating no relationship to the original score. Figure 4.10 does not support that, with the near-win curve closely following the no-near-win curve. Near-win events appear to consistently improve engagement when students score under one star. The upward kinks in the near-win predictions relative to no-near-win predictions at the end of each of the

---

2Another way of selecting models is to fit the biggest model, the one with all features included, and eliminate feature coefficients that are not reliably different from zero. However, this may introduce too much bias into the selected model as the eliminated coefficients may still have predictive benefits. Cross-validation measures the predictive power of the models, which is what we ultimately care about.
three score brackets are expected because by definition, no near-wins exist within 0.1 of a goalpost. An interesting phenomenon in Figure 4.10 are the bumps in the middle of the score bracket. These suggest that the students may treat the middle of the score bracket as some kind of an invisible goalpost to get over.

To obtain corroborating evidence of the conclusions from the logistic regression analysis, a Gaussian process (GP) model was fitted to the dataset. The model accepts three features as inputs: the star level, the original distance to next goalpost, whether a near-win had occurred. Since a GP requires the specification of a covariance kernel of input features—the degree to which changes in function value at one location affect the function value at another location—Gamma priors with means 0.5 and 0.05 were placed on the star level and distance to goalpost dimensions, respectively. The length scales for the near-win indicator was fixed at 0.01—specifying that almost no relationship between near-win=0 and near-win=1 exist—because this mimics how the logistic model specifies a different regression equation depending on the values of the two features. Similarly to logistic regression, the GP model has a logistic observation model, \( \text{logit}(p_{\text{replay}}) = f(x) \) where \( f \sim \text{GP} \). The model was trained via MCMC with 10 chains, 1000 iterations/chain, and 500 burnin iterations in each chain. Predictions of the GP in Figure 4.11 are smoother than the logistic regression model predictions, possibly because the GP has fewer parameters than the best logistic regression model (the GP has two kernel parameters, one mean parameter, and one latent noise parameter while the logistic model has 8 parameters). The GP model does appear to better capture the lack of difference in engagement at higher scoring brackets, which is observed in binning analysis in Figure 4.8. When the original score is under one star, the absolute replay probability increases by about 0.1 when the student experiences a near-win. Similarly to the logistic
model, the near-win curve is not flat in the one star score bracket, indicating that students are aware of their own performance even when a near-win is induced. Nonetheless, both logistic regression and the GP agree that near-wins increase engagement when the student scores in the lowest score bracket.

Given that there is evidence that students are aware of their own performance, the previous analyses omit an important factor: whether a near-win is *induced* or *natural*. Induced near-win events are events where the score is artificially moved close to the next goalpost, while natural near-wins occur due to the student’s actual performance on the lesson. The next section repeats the previous analyses while taking into account the distinction between natural and induced near-wins.

### 4.3.2.1 Excluding Natural Near-Wins

The procedure used to generate Figure 4.8 was repeated except now all natural near-win events were excluded from the analysis. Figure 4.12 plots the replay probability as a function of the original score and whether a near-win was induced or not. There is no advantage of induced near-win events over no-near-win events. This indicates that the boost in engagement observed in the 0.75-1.00 score bracket in Figure 4.8 was due to natural near-win events.

Similarly to the earlier analysis, logistic and GP models were fitted to the dataset. The best logistic model selected by cross-validation is the one that includes all features: the star level $B$, the original distance to the next goalpost $O$ and its’ square $O^2$. Figure 4.13 plots the logistic regression (top plate) and the GP model results (bottom plate). The engagement advantage of near-win events near the ends of score brackets is gone, similarly to Figure 4.12. But, both logistic and GP models predict higher engagement from near-win events in the lower half of the one-star score bracket.
Figure 4.12: Analysis of replay rates as a function of the original score and the occurrence of the near-win, excluding natural near-win events. Each bar corresponds to a score range \([a, b]\), with \(a\) included and \(b\) excluded. Bar heights indicate the mean replay probability of observations in the corresponding score range. Blue and orange bars correspond to no-near-win and near-win events, respectively. Error bars correspond to the 95% confidence interval for a binomial proportion, using Wilson’s score interval.

The diminished effect of near-wins on engagement when excluding natural near-wins indicates that the students decision to replay is not solely based on the displayed score. Rather, students may be aware of their own performance and/or other factors such as lesson difficulty. I explore this question by comparing the predictive power of logistic regression models using the displayed vs. the original score in the next section.

### 4.3.3 Displayed Score vs. Original Score

If the students are deciding to replay based purely on the displayed score, then a model using the displayed score should be superior to one using the original score. To see whether this is the case, I compare a logistic regression model that includes all main effects and interactions among the star level, the original distance to the next goalpost and its’ square, to a similar logistic model that replaces the original distance to the next goalpost with the displayed distance to the next goalpost. Natural near-wins were excluded from the analysis because a natural near-win may match the student’s underlying
expectations, and in this analysis I am interested whether the displayed score can predict engagement reliably better than the original score. Mean test log-likelihoods were gathered from 10 replications of 5-fold cross validation on the filtered dataset.

Results of the cross-validation indicate that the original score model has a greater mean log-likelihood than the displayed score model (paired t-test statistic=25.61, \( p < 0.001, \mu_O = -0.652 > \mu_D = -0.657 \)), where \( \mu_O \) and \( \mu_D \) are the mean test log-likelihoods using the model with the original and displayed scores, respectively. One possible explanation is that the original score encodes extra information about the trial, such as the skill of the player or difficulty of the lesson, which is partially lost when a near-win event is triggered.

A valid question then is what happens if both the original score and displayed score are used in the model? does the displayed score add additional information beyond what is encoded in the original score? To answer this, two models were tested: a logistic regression model using the original score with a binary near-win indicator (model A) and a model that does not have the near-win indicator (model B). Cross-validation results indicate that the model with the near-win indicator feature has reliably greater mean log-likelihood than the model without the indicator feature, although the increase in likelihood is
small (paired t-test statistic: $-2.80, p = 0.01, \mu_A = -0.6517 > \mu_B = -0.6519$).

These results suggest that the original score is the main predictor of engagement and that the inability of the displayed score to predict as well indicates that students are basing their decision to replay a lesson on more than just the information given to them.

### 4.3.4 Interaction Between Near-Wins and Student Ability

Overall, the regression analyses in section 4.3.2 suggest that near-win events are most beneficial for low-skill students. To examine this directly, students were ranked and grouped according to their median score on all practiced lessons. For each group, or percentile range, the difference in mean replay probability of near-win and no-near-win events was computed. This was done when including all near-win events or when including only induced near-win events. Figure 4.14 plots the results of this analysis. Bar heights correspond to the differences in mean replay probabilities, where greater differences indicate that students are more likely to replay after near-win events. Looking at all near-wins (blue bars), low-skill students are more likely to replay after experiencing a near-win events. However, the effect is greatly diminished when looking at the induced near-wins (orange bars), with a maximum difference of about 0.02 in replay probabilities. This finding suggests that the artificial induction of near-wins is not as effective as natural near-wins in promoting engagement.

One possible explanation is related to how much effort the student has put in vs. the reward. If the student puts in small effort and almost gets the reward, due to an induced near-win, then they might not value the expended effort high enough to justify replaying the lesson. On the other hand, if the student puts in a lot of effort and marginally fails to clear the goalpost, due to a natural near-win, then he or she may feel compelled to replay in order to justify the expended effort.

### 4.3.5 Design Space Analysis

The ultimate objective of this experiment is to characterize engagement as a function of the two-dimensional design space of near-win probability and ease-out magnitude. To evaluate the effect of the near-win probability, the student must do enough lessons to get a sense of the frequency of near-win
Figure 4.14: The difference in mean replay probability between near-win and no-near-win cases as a function of student ability (percentile rank of the student). The difference in replay probability is shown for different percentile ranges. A higher difference indicates that lessons are played more frequently after near-win events. Orange and blue bars correspond to the differences in replay probabilities when considering induced and all near-win events, respectively.

So all students who did not do at least 10 lessons were filtered out. Entries where the student moved on before the animation had finished or where the near-win was excluded were filtered out. This reduces the dataset to 8,330 entries from 503 students. The median number of lessons attempted is 14 lessons (std. 9.18, range 10-83), over 7 sessions (std. 3.66, range 1-34), with a lesson completion time of 3.34 minutes (std. 11.00, range 0.54-409).

Logistic regression and neural network models were trained to predict whether the student replays a lesson, given the design inputs \((\nu, \eta)\). The logistic regression model is specified with

\[
\text{logit}(p_{\text{replay}}) = \beta_0 + \beta_1 \nu + \beta_2 \nu^2 + \beta_3 \eta + \beta_4 \eta^2 + \beta_5 \nu \eta
\] (4.1)

The neural network model has one layer of 20 hidden TanH units and 50% dropout rate between the hidden units to the output. An ensemble of 10 models were trained and their predictions were weighted by the training loss – binary cross-entropy. Learning rate was adapted using RmsProp and all models were trained for 2000 epochs with a mini-batch size of 250 observations. Significant effects were found of the near-win probability \((\beta_1 = 3.48, p < 0.0001)\), near-win probability squared \((\beta_2 = -2.37, p < 0.0001)\), ease-out \((\beta_3 = 1.33, p < 0.0001)\), and the interaction between the near-win probability and ease-out.
Figure 4.15: Predictions of the engagement as a function of the two dimensional design space. The left plate shows the predictions of the logistic regression model in equation 4.1 and the right plate shows the predictions of the weighted ensemble of neural network models. Red squared indicates the maximum.

(β₄ = -2.01, p < 0.0001). Predictions of the logistic regression model and the weighted ensemble of neural network models are plotted in Figure 4.15. Both models predict that staying along the anti-diagonal leads to maximum engagement which suggests that at high near-win probabilities students expect not to win so they don’t want to anticipate something that won’t happen. The optimal region in logistic regression has 125% greater engagement relative to the minimum region, while the neural net’s predicted optimum is 21% better than the minimum. The predictions of the weighted ensemble of neural networks look similar to those of a single neural network as one can see the individual hyperplanes so I re-trained the models several times and found that it is often the case that one of the models dominates the others in terms of weight. Setting the weights to be all equal yields the very similar predictions to those in the right plate of Figure 4.15. Out of sample testing of both models using 10 replications of 5-fold cross-validation indicates that the neural network is reliably better than the regression model in terms of mean log-likelihood (paired t-test statistic = -4.72, p < 0.0001, \( \mu_{NN} = -0.675 > \mu_{LR} = -0.676 \)).

The previous analysis does not control for score or completion time. To factor out these effects using logistic regression would require a plausible theory of how all factors in the model interact among each other. Crafting such a theory or model by hand can be dangerous as one risks overfitting or “fishing” different subsets of the data to get the desired effect. Instead, I use an additive neural network architecture to study the main effects of several factors, while eliminating variance due to interactions.
Specifically, the network computes:

\[
\logit(p_{\text{replay}}) = f_1(\nu) + f_2(\eta) + f_3(S_o) + f_4(\tau) + f_5(\nu, \eta) + g(\nu, \eta, S_o, \tau)
\]  

(4.2)

where \(\nu\) is the near-win probability, \(\eta\) is the ease-out, \(S_o\) is the original score, and \(\tau\) is the time taken to complete the lesson. This specification is inspired by the approach taken by Duvenaud, Nickisch, and Rasmussen (2011), where the authors used an additive Gaussian process kernel that included first-order and full interaction terms. Specifying the model in this manner allows the network to decompose the predictions into first-order terms, if such decomposition is possible; otherwise, the full interaction term, \(g\), would enable the network to predict just as well as the non-additive network. In other words, the main advantage of the additive network is interpretability rather than predictive power. For \(g\), two layers of TanH units (20 and 10 units, respectively) were used with 50% dropout between the layers and between the last layer and the output. Each function \(f_i\) is represented as one layer of 10 TanH units with 50% dropout between the hidden layer and the output. As with the previous analysis, an ensemble of 10 models were trained and weighted by the likelihood of the data. Figure 4.16 plots the log-odds contributions of each \(f_i\); the y-axes of the one-dimensional terms have identical ranges to make it easier to compare effect sizes. The original score and the lesson time are the dominant contributors to the probability of replay. The probability decreases as lesson time increases which indicates that the student’s decision to replay or continue depends on the effort expended. Neither the near-win probability nor the ease-out magnitude have a main effect. However, the interaction among the two does have a significant contribution to the log-odds of replay with a range from -1.3–1.3, or about 26% of the range of log-odds contributions. The predictions of the interaction are similar to those in the right plate of Figure 4.15, which is reassuring.

The analyses up to this point have focused on predicting whether a student replays a lesson, given design information and other factors. This is appropriate for the ease-out manipulation but the near-win probability manipulation by definition requires exposure to multiple lessons. I tried to control for that by restricting the dataset to students who completed a minimum of 10 trials. But the ideal way of studying the effect of the near-win probability is similar to that of Kassinove and Schare (2001) – where the player would do \(N\) trials followed by a period of voluntary play – although it is unlikely that
Analyses of displayed scores indicate an increase in engagement when low-performing students experience a near-win event, with engagement increasing as the distance to the 1-star goalpost decreases. However, engagement appeared to drop half-way through the higher score brackets, which is inconsistent with the findings of Lomas (2014). Also, whereas Lomas studied near-win events that occur naturally due to player skill, the first contribution of this chapter is the ability to manipulate scores (almost) independently of skill. Because of this, I was able to directly estimate how much more engaging a near-win is for a given
original score. Multiple analyses show that a near-win event will increase engagement when scoring in
the 1-star bracket. However, these analyses also show that engagement is significantly influenced by the
original unmanipulated score rather than just the displayed score, indicating that students may be aware
of their own performance. It appears that this is why, when excluding natural near-wins, the boost in
engagement from near-win events disappears.

Because of the partial dissociation between the original and displayed scores, the second contribu-
tion of this chapter is the comparison of predictive power between models of the original and displayed
scores. A logistic regression model using the original score enjoys a reliable advantage in terms of test
log-loss over a model that uses the displayed score. The original score could be encoding extra bits of
information—such as student skill—that get lost when a near-win event is triggered.

The third contribution of this chapter is the independent control of the frequency of near-win
events in educational software, similarly to what studies in gambling psychology traditionally do. A
related contribution is the continuous control of anticipation over time. Whereas near-win experiments
often manipulate anticipation—by inducing an early-loss for example—I control the distribution of antic-
ipation over time via a “knob” that modulates how much deceleration or ease-out occurs as the animated
performance bar approaches the target display score. Multiple analyses of the design space show an in-
verse relationship between the near-win probability and ease-out magnitude, suggesting that students
prefer not have a long ease-out phase when the near-win probability is high, i.e., students would rather
not have false hope if they know that they will never win. An additive neural network model suggests
that the effect of the two-dimensional design space is small compared to the original score and the time
taken to complete the lesson.

A likely explanation for the relatively small effect of ease-out on engagement is the short max-
imum animation duration of 3.6 seconds. While a longer duration increases the buildup of anticipation,
it may come at the cost of less engagement. My informal testing indicates that a minimum of about
7-8 seconds is required to buildup anticipation, which could be too long given that the median duration
students stay on the scorecard screen is 8.3 seconds. One way to solve this problem is to accentuate the
ease-out manipulation by adding sounds – such as the ticking sound of a roulette wheel – to increase the
impact of the effect.
Students often score 3 stars – potentially due to the adaptive nature of the program – and attempt a small number of lessons, making it hard to analyze the effect of near-win probability on engagement. Given the high exclusion rate of the near-win manipulation, a redesign is necessary so that the manipulation is independent from student performance. One possibility is to define “winning” as getting a bonus virtual pen color in Wootmath or getting to play a round of a simple game like Flappy Birds or SpringNinja. The new manipulation would then control how often students win, nearly-win, or lose. Another possibility is to show the student a fake list ranking the student’s performance relative to students in other schools (so that students in the same class don’t notice the manipulation). The list would contain targets or goalposts and the location of the target or goalpost changes depending on whether a near-win event is triggered or not.

The marginal effect of the design space on engagement suggests two modifications for future studies. First, the near-win manipulation should be decoupled completely from student performance so that the student can get a better sense of the rate of near-wins. Second, the ease-out manipulation should be complimented with additional effects, such as sound or other animations, to accentuate the buildup of anticipation. Additionally, one possible extension of this work is to include an adaptive near-win controller that manipulates the score based on a student model of engagement. For example, if the model predicts that the student will not replay if they experience a near-win, then maybe it is a good idea to reduce the score so that the student is more likely to play.
Chapter 5

Conclusion and Future Work

In this thesis I reviewed literature on video game engagement which mostly discusses manipulations of aesthetics, rewards, progress, and choice. Despite the existence of rich literature in other domains on novel manipulations, such as tension-and-release and the near-win effect, the benefits of such manipulations are often discussed qualitatively in the context of video games. Even in instances where these manipulations are analyzed quantitatively, only few conditions within the manipulation space – the design space – are explored. I believe that the manipulations I discussed in this thesis go to the heart of what makes a video game engaging; while superficial elements such as aesthetics or competition are important, they represent the lower building blocks of an engaging game. Manipulations such as tension-and-release and near-win control the narrative of the game – higher-level manipulations of engagement.

I considered a manipulation of difficulty that covertly helps the player, so that the player attributes success to his or her own efforts. Rather than exploring a few levels of this manipulation, as is commonly done in A/B testing, I explored the full continuum of covert difficulty settings. To do this, Bayesian optimization with a Gaussian process model that is motivated by a cognitively plausible theory of engagement was used. Different distributional assumptions of this theory were evaluated in a synthetic optimization experiment and the best one was picked for real-world experiments. In addition to the behavioral measurement being optimized (Voluntary Time on Task), a subjective measure of engagement was elicited from subjects via a post-experiment questionnaire. Across two games and four
studies, evidence points to moderate covert assistance providing optimal levels of engagement, which is consistent with self-attribution theory as high levels of assistance tend to be overt so the player cannot attribute success to his or her own effort.

For tension-and-release, randomized studies were conducted to map out the impact of the full space of game designs on engagement. To this end, I built a custom web-based memory training game that aims to solve the issues with existing working memory tasks. Within this game, a tension-and-release difficulty curve was shown to modulate engagement, controlling for average difficulty. A two-dimensional space of tension-and-release manipulations was designed, incorporating baseline and random conditions. Random conditions were made such that direct inference can be made about whether the structured variability of tension-and-release is necessary for engagement vs. random variability. Non-parametric analysis using Gaussian process regression indicated no difference between structured vs. random designs, but when the tension-and-release cycle duration was increased, subjects’ preferences shifted towards structured designs.

The near-win effect was operationalized in a large study involving thousands of students using a fraction learning software. Unlike other work on the near-win in education, I attempted to study the effect of the rate of near-wins on engagement. The manipulation was subtle; it merely manipulated the final score the student sees – as point on a performance bar – after completing a series of exercises. Animation of the score within the performance bar was also controlled by manipulating the amount of ease-out or deceleration. The near-win rate and ease-out manipulations thus formed a two-dimensional design space. Near-win events significantly increased engagement for lower performing students and manipulated scores have been found to increase engagement relative to un-manipulated scores. Analysis of the design space using parametric and non-parametric methods found a marginal effect of the near-win probability and ease-out on engagement. The two dimensions were found to be inversely proportional: if the near-win probability is high, the ease-out amount should be low to maintain optimum engagement.
5.1 Future Work

A common limitation to the experiments of Chapters 2 and 3 is the lack of within-subject adaptation. Both chapters sought to optimize or map out game design spaces for a population of users. Experiment 3.2 used a rudimentary difficulty controller that adapted to the subject’s skill and successfully kept the median score of the subject population within the target range. Rather than limit adaptation to dynamic difficulty adjustment, I would like to use Bayesian optimization on a subject-by-subject basis to find game designs that are optimal for a specific subject.

Experiments in Chapter 3 used random assignment to map out the design space, which is feasible for a few dimensions but quickly becomes impractical as the number of dimensions increases. It is therefore necessary to use Bayesian optimization but unlike Chapter 2, the objective is to map the design space rather than to find the global maximum. A principled information-theoretic approach to this problem has been proposed in (Houlsby, Huszár, Ghahramani, & Lengyel, 2011). This approach works for regression and preference-learning models alike and in the latter case, pairs of conditions are jointly selected without relying on an ad-hoc selection procedure. Future studies will use this approach to enable efficient exploration of high dimensional game design spaces.

The efficacy of near-win events in a large-scale educational software has been confirmed in Chapter 4. However, there were two major issues: (i) most of the time, students did not get exposed to near-win events because they were scoring perfectly, thereby reducing the influence of the near-win rate on engagement and (ii) the duration of animation was too short for anticipation to buildup. For the first problem, I plan on implementing a manipulation that is completely decoupled from student performance, such as spinning a wheel to get to play a round of simple game. For the second problem, it is not possible to increase animation duration due to concerns about student engagement so I plan to couple the animation with sound effects and possibly other effects to accentuate the buildup of anticipation.
References


Kassinove, J. I., & Schare, M. L. (2001). Effects of the” near miss” and the” big win” on persistence at slot machine gambling. Psychology of addictive behaviors, 15(2), 155.


from https://www.wootmath.com/
