Designing Engaging Games Using Bayesian Optimization

Mohammad M. Khajah
University of Colorado
Boulder, CO
mohammad.khajah@colorado.edu

Brett D. Roads
University of Colorado
Boulder, CO
brett.roads@colorado.edu

Robert V. Lindsey
Boulder Analytics
Boulder, CO
robert@boulderanalytics.com

Yun-En Liu
University of Washington
Seattle, WA
yunliu@cs.washington.edu

Michael C. Mozer
University of Colorado
Boulder, CO
mozer@colorado.edu

ABSTRACT

We use Bayesian optimization methods to design games that maximize user engagement. Participants are paid to try a game for several minutes, at which point they can quit or continue to play voluntarily with no further compensation. Engagement is measured by both actual play duration and a projection users make about how long others will play. Using Gaussian process surrogate-based optimization, we conduct efficient experiments to identify game design characteristics that lead to maximal engagement. We study two games requiring trajectory planning, the difficulty of each is determined by a three-dimensional continuous design space. Two of the design dimensions manipulate the game in user-transparent manner (e.g., the spacing of obstacles), the third in a covert manner (subtle trajectory corrections). Converging results indicate that covert manipulations are significantly more effective in driving engagement, suggesting the critical role of a user's self-perception of competence.

Author Keywords

engagement; motivation; games; education; covert and overt difficulty manipulations; optimization

ACM Classification Keywords

I.2.1 Artificial Intelligence: Learning; H.1.2 Information Systems: User/Machine Systems; G.3 Probability and Statistics: Experimental Design

INTRODUCTION

There has been a recent surge of interest in applying game-like mechanics to enhance engagement in a variety of domains, such as personal health [11, 14], scientific discovery [16, 8], and education [9, 21, 20, 19]. The expectation is that increased engagement will improve user experiences, data collection, and outcomes. Although engagement is a broad construct [12], it can be readily quantified using electronic games. To use educational games as an example, one can measure the fraction of time students are attending to the screen [28], the rate of responses [2], the number of attempts to solve a problem, and a persistent focus on a single task [19]. All these measures of engagement relate to the user spending more time on task, which should ultimately yield better learning outcomes.

Techniques to increase engagement have produced encouraging results. For example, when students are provided software that instantly grades an essay, they obsessively tweak their text in an attempt to notch up their scores [Peter Foltz, personal communication, January 22, 2014]. Unfortunately, gamification efforts often fail when they do not increase engaagement. Following a gamified software engineering course, 2/3 of students viewed the gamification features as a waste of time that distracted from the core activity [3]. Katz et al. [15] incorporated various game-like motivational elements into cognitive training software, including real-time scoring, selection of themes and art, scaffolding via game levels, prizes and certificates. No performance gains were observed from these elements when a student's total activity time was controlled for, suggesting that the key benefit of gamification may arise from an increase in time students are willing to direct at an activity.

In gaming, 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 game design identifies the not-too-easy, not-too-hard level of play that draws players in. Indeed, we conjecture that the most addictive games provide a sense to players that even when they lose the game, they are all-so-close to success. Claw machines (Figure 1) are carefully designed to give players a near-success experience [10]. These machines consist of a large enclosed bin of plush toys and a player-controlled claw that reaches into the bin, grabs a toy, and—if the player is successful—deposits the toy

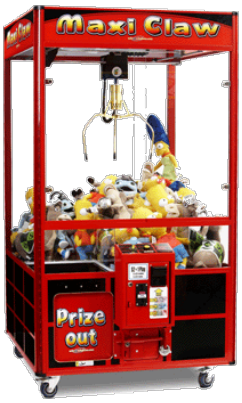

Figure 1. The claw machine which engages users via near-success rounds of play.

in a chute accessible to the player. These machines are rigged in a devious way: they are programmed to have a strong grip for only part of the claw's trajectory, making it near impossible for the claw to grip the toy until the chute is reached [10]. Because the player is not aware of this modulation of the claw's grip strength, they are left with the sense of being on the verge of success, and are convinced that surely the next round of the game will lead to a better outcome.

This aspect of gamification—manipulating difficulty to give players a sense of accomplishment when they succeed and a sense of being on the verge of success even when they fail—seems well suited to application in a range of domains. For example, in education, the influential psychologist Vygotsky [27] proposed the notion of a *zone of proximal development* to refer to challenges that are just beyond the grasp of a student. Vygotsky's claim is that a student learns most effectively when given educational challenges that are not doable based on the student's current knowledge and skills, but which ought to become tractable with a bit of scaffolding, guidance, or trial-and-error practice. From the early days of electronic tutoring systems, the promise has been *adaptivity*—the ability of the system to be sensitive to the student's abilities [5]. To the best of our knowledge, no studies have been performed to determine whether increased learning gains from appropriately chosen exercise difficulty are due to increased motivation and engagement or to a larger increment to knowledge structures.

The goal of our work is to design games that maximize engagement for a population of users via manipulation of task difficulty or challenge. Ordinarily, design decisions are made with A/B testing or with a designer's intuitions. A literature has begun to emerge that leverages the vast quantities of user data that can be collected with online software to *optimize* the design through more systematic and comprehensive experimentation. We will discuss past approaches that have been used to maximize engagement by exploration of design space, and we will present a novel approach in this domain using a technique referred to as *Bayesian optimization*. The method we present could theoretically be applied to any game after it is released in the wild, leading to automated improvement of the software with minimal intervention by designers. In the optimization framework, the role of designers is to specify a space of designs over which exploration will take place.

Recent Research on Engagement Maximization

Recent research has used an on-line educational gaming platform to search a design space to maximize engagement. 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. [21] 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. [19] 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 [20] 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.

BAYESIAN OPTIMIZATION

We propose an alternative methodology to search for engagement-maximizing designs: Bayesian optimization. Bayesian optimization can represent structure in the design space, which gives rise to two advantages over previous approaches. First, it can model ordinal and cardinal dimensions to discover functional relationships between designs and outcomes. Second, it is efficient in its use of data, leading to strong inferences with orders of magnitude less data than utilized by previously tested methods. This efficiency arises from an underlying assumption that the function relating designs to engagement is smooth, i.e., nearby points in the design space yield sim-

ilar degrees of engagement. The approach is nonparametric: the degree of smoothness is data dependent.

Bayesian optimization infers a surrogate function that characterizes the relationship between designs and a latent valuation. Here, the designs refer to parameterizations of a game, and the valuation is the degree 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 has recently been applied to the design of a shoot-'em-up game. Zook et al. [29] searched over several game 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 engagment 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 [7], two-alternative forced choice with guessing [17], and similarity judgment [24]. Here, we develop and justify a probabilistic model to predict behavioral measures of engagement from the latent index of engagement.

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. [19], 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 $\mathcal{A}$, a player will complete the next trial (and not quit). We call this the *persistence* induced by $\mathcal{A}$.

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 2a shows the empirical persistence across designs, and Figure 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. The model produces a prediction of engagment over the design space that is consistent with that obtained by the approach of Liu et al. [19], 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. Further, the result in Figure 2b replicates across regroupings of the data.

This simulation used a logistic observation model and a squared exponential kernel with ARD distance measure, yielding 6 hyperparameters which were drawn via elliptical slice sampling. This kernel effectively computes a weighted Hamming distance on the binary dimensions.

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 [13] or Weibull density [25]. Evidence about usage-duration distributions is harder to find. Miyamoto et al. [22] 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. [1] 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 better matched to empirical distributions. We evaluate these models via simulation experiments.

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)

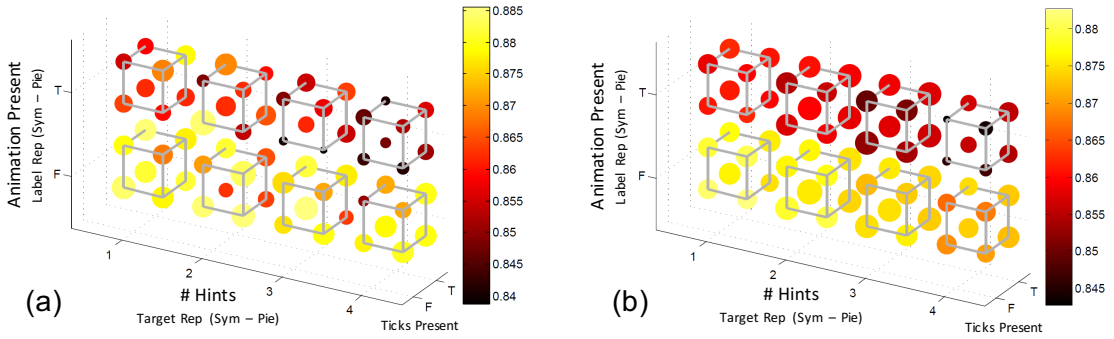

Figure 2. Persistence probability of Treefrog Treasure players over the design space: (a) empirical mean; (b) GP-posterior mean. Each disk represents a design. Color denotes persistence and diameter is inversely proportional to variance of (a) aggregated observations and (b) the GP posterior.

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 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:

$$T \sim \text{Gamma}\left(\alpha, \tfrac{\alpha}{\mathrm{e}^{f(\boldsymbol{x})}}\right)$$
$$T \sim \text{Weibull}\left(k, \frac{\mathrm{e}^{f(\boldsymbol{x})}}{\Gamma\left(1+\frac{1}{k}\right)}\right)$$
$$T \sim \ln \mathcal{N}\left(f(\boldsymbol{x}) - \tfrac{\sigma^2}{2}, \sigma^2\right)$$
$$T \sim \text{Wald}\left(\lambda, \mathrm{e}^{f(\boldsymbol{x})}\right)$$

where $\boldsymbol{x}$ is a game design and $f(\boldsymbol{x})$ 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, $\mathrm{e}^{f(\boldsymbol{x})}$, but have different higher-order moments. Note that the Gamma distribution includes the exponential as a special case. To allow a design's valuation $f(\boldsymbol{x})$ to influence the choice C as well as the play duration T, we define $\text{logit}(\pi) \equiv \beta_0 + \beta_1 f(\boldsymbol{x})$. 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, $f(\boldsymbol{x})$, that represents the engagement associated with a design $\boldsymbol{x}$. For $f(\boldsymbol{x})$, 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 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 [6], 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 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

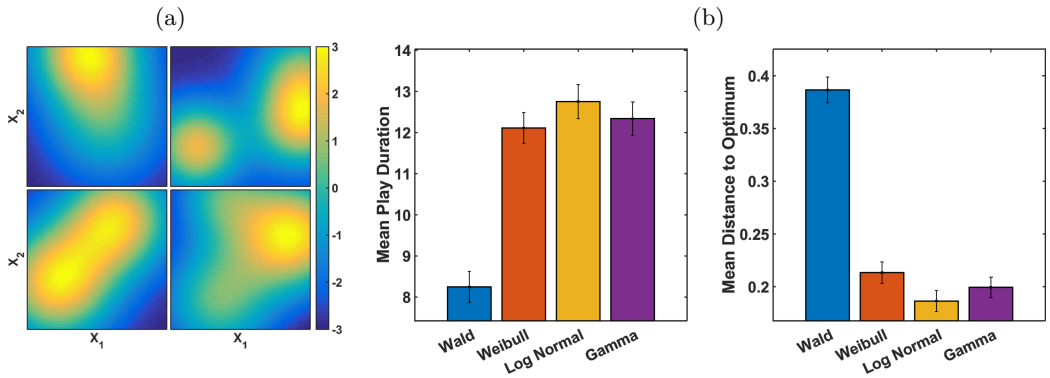

Figure 3. Simulation experiment. (a) Examples of the 2D functions used for generating synthetic data. (b) Results of synthetic experiment. The left and right plots depict the mean function value (higher is better) and the mean distance to the true optimum (lower is better) for various observation models. Results are averaged over four different generative-process models, 100 replications of each simulation, the last 10 trials per replication. Error bars indicate ± 1 standard error.

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.

EXPERIMENTS

Let us take a step back and remind the reader of our overall agenda. We wish to maximize engagement over a game design space. Engagement is quantified by play duration. We identified difficulty manipulations as a potentially effective means of influencing engagement for a population of users. 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 reanalysis 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 so that they could continue studying. In all cases, the participants' motivation was to learn, not to receive additional compensation.

Given this evidence 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.

VTA appears to be a natural and readily measured proxy for engagement. The traditional method of assessing engagement is a post-experiment survey (e.g., [23]). Recently, however, VTA-like measures have been explored. Sharek and Wiebe [26] 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 [21, 20, 19], engagement was measured by how likely a player is to switch to a different game.

Overt Versus Covert Difficulty Manipulations

In our experiments, we distinguish between *covert* and *overt* manipulations of task difficulty. To explain the distinction, consider again the claw machine (Figure 1), which extracts money from users via a perverse type of difficulty manipulation: a manipulation of which the user is unaware. The user assumes it is their own failing or misfortune that the claw prematurely releases the plush toy, when in fact the machine is designed to fail in this manner. If users understood this design characteristic, it is unlikely that they would continue to invest their coins. The covert nature of the difficulty manipulation is what makes it effective. We hypothesize that its effectiveness is due to the fact that individuals readily overestimate

their sense of agency—the amount of control they have over an outcome [18].

A specific research question we address in this article is whether covert manipulation of difficulty is more effective in engaging users than overt manipulations—those of which users are fully aware and to which they can attribute causal effects. Covert manipulations can be used to make a task more difficult than the user believes, but also to make a task easier than the user believes.

In educational software, covert manipulations might include scaffolding over a sequence of problems, e.g., first giving a student 23+45 before the problem 2.3+4.7. Another manipulation might be in the set of alternative responses in a multiple choice problem—whether or not close lures were included. Similarly, suppose students were asked to make a response by picking a point along a number line. As they ran the computer mouse over alternatives, the cursor might jump in a nonuniform manner such that the dwell time on the correct or plausible responses was longer than on incorrect or implausible responses. If the student made a close-but-incorrect selection, the software could simply switch to the near-and-correct response. Even more devious manipulations are possible. We spoke to a classroom teacher who deliberately left a solution to a related problem on his desk when he was working with a student, expecting that the student's eyes would dart around as they were trying to come up with a solution and would benefit from the example.

In gaming, the perception of accomplishment may similarly boost engagement. Players are typically aware of factors affecting game difficulty, e.g., the speed and agility of the enemy. In contrast to such overt factors, game physics might be modulated to manipulate covert difficulty We hypothesize that manipulations of which players are unaware will influence players' perception of competence which in turn will influence engagement.

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 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 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

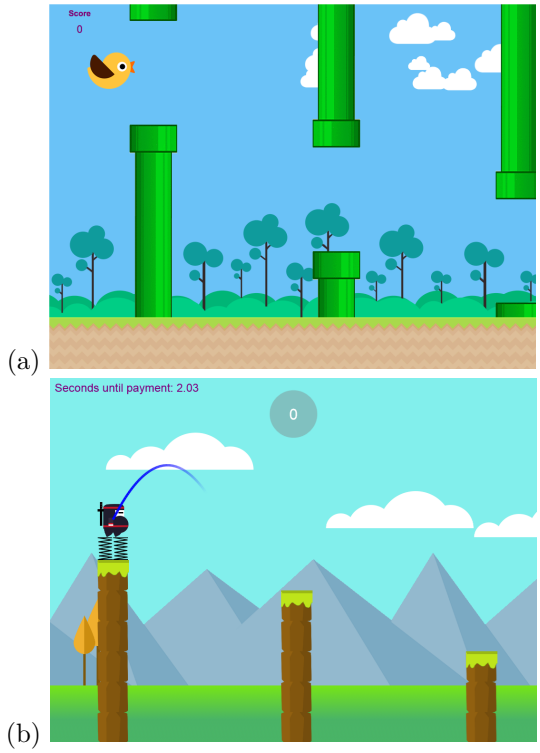

Figure 4. (a) Flappy Bird: The player flaps bird's wings to keep it aloft and to avoid hitting pipes. (b) Spring Ninja: The player jumps from one pillar to another by compressing springs in the ninja's shoes. The blue trajectory is the projected jump path for the given spring compression level.

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 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. In informal testing, players were unaware that the game dynamics were modulating to guide them along.

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). 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 Bayesian optimization. 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 are attractive because they evenly cover design space, as opposed to a sequence of purely uniform random samples. 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. 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* 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 rounds only because they had not noticed the finish button. Whilst 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

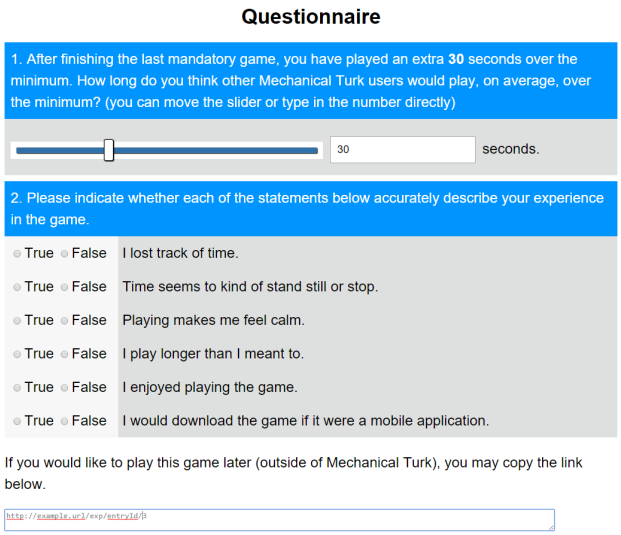

Figure 5. The post-experiment questionnaire.

incorporated a slider control that is initially anchored on their experiential VTA (see top of Figure 5).

In the active-selection study, we included a short questionnaire about the participant's experience 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 5) were taken from the Game Engagement Questionnaire [4].

Among the participants in the random-assignment study, the mean experiential VTA is 8 sec, with standard deviation (SD) 32 sec and range 0-432; only 19% of participants chose to play beyond the requirement. The mean projected VTA is 22 sec with SD 46 sec and range 0–700; 79% of participants projected that others would continue playing beyond the requirement. The Spearman correlation between experiential and projected VTA is only 0.24, likely because of the large number of zero VTA responses. In the active-selection study, the mean experiential VTA is 10 sec, with SD 42 sec and range 0–298 sec; 20% of participants chose to play beyond the requirement. The mean projected VTA is 23 sec, with standard deviation 33 sec and range 0–199 sec; 84% of participants projected that others would continue playing beyond the requirement. The Spearman correlation between experiential and projected VTA is 0.286.

Figure 6a and 6b show the model posterior mean VTA over the three dimensional design space in the random-assignment and active-selection studies, respectively. The remarkable finding is that the two independent studies yield very similar outcomes: the optimal design identified by the two studies is in almost exactly the same point in design space (the red squres in the Figures). The random-assignment study should yield reliable results due to the relatively large number of participants tested. The active-selection study is far more efficient

in its use of participants, due to intelligent selection of where to explore in design space.

In both studies, the peak design is predicted to obtain a VTA of 30 seconds—an increase of 50% of the time on task. Because Turk workers are paid by the task, this time increase reduces the pay rate by two thirds, a fairly clear indication of engagement.

The Figures indicate that engagement is sensitive to each dimension the the design space. There is not much hint of an interaction across the dimensions. Notably, with minimal covert assistance (the upper-left array in each Figure), 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 identified in Figures 6a and 6b, 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 Gaussian 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 6c 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 design optimums (denoted by red squares) lie almost exactly in the same place in 6a, 6b and 6c. 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.

Spring Ninja: Experimental Methodology and Results

We conducted a single study with Spring Ninja with 325 participants. 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.

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 finish the experiment (if the mandatory play time had elapsed). 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 5).

We measured the VTA in Spring Ninja differently than in Flappy Bird because the former is turn-based whilst 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 be 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 measured 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.

Among participants in the active selection study, the mean experiential VTA is 12 jumps with SD 27 and range 0-296; 74% of participants chose to jump voluntarily after the mandatory time has elapsed. The mean projected VTA is 10 jumps with SD 14 and range 0–118; 89% of participants projected that others would continue to jump after the mandatory time had elapsed. The Spearman correlation between the experiential and projected VTA is 0.4.

Figures 7a and 7b show model posteriors over VTA (in number of jumps) and questionnaire scores, respectively, fit in the same way as we did in the Flappy Bird study. High engagement for the mid-range of design parameter settings. The predicted optima by the two measures are

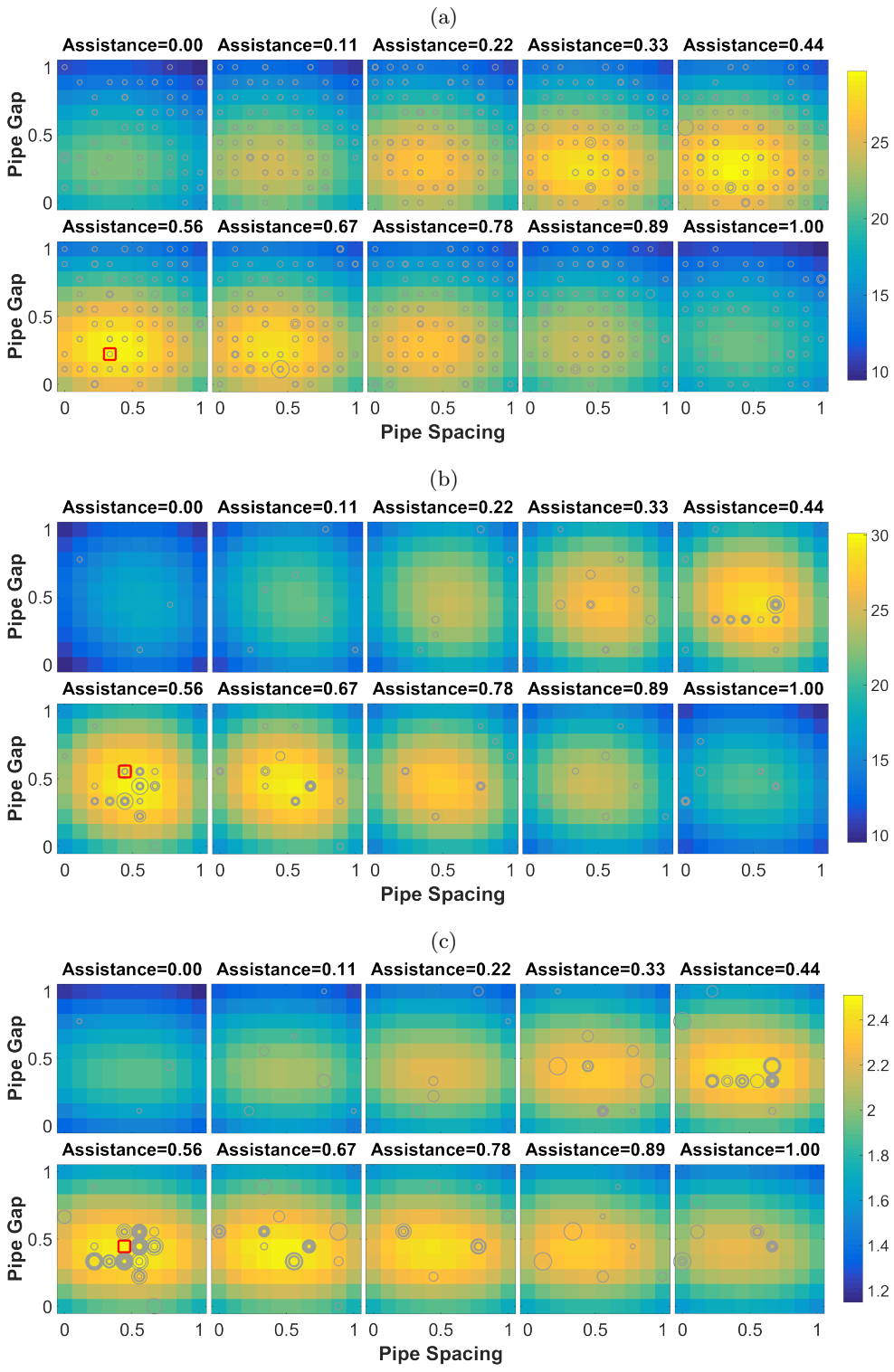

Figure 6. Bayesian model fits of VTA (in seconds) over the Flappy Bird design space for (a) the random-assignment and (b) active-selection studies. Each array corresponds to a fixed level of assistance, with the upper left array being no assistance (level 0) and the lower right array being maximal assistance (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 is a challenging game, unlikely to be played well by a novice, and 1 is readily handled by a novice. The circles correspond to observations with the radii indicating the magnitudes of the observations. At locations with multiple observations, there are co-centric circles. Red squares indicate the locations of the predicted global maximum. (c) An analogous Bayesian model fit to the questionnaire score, which indicates the number of items with an affirmative response. Higher scores indicate greater engagement.

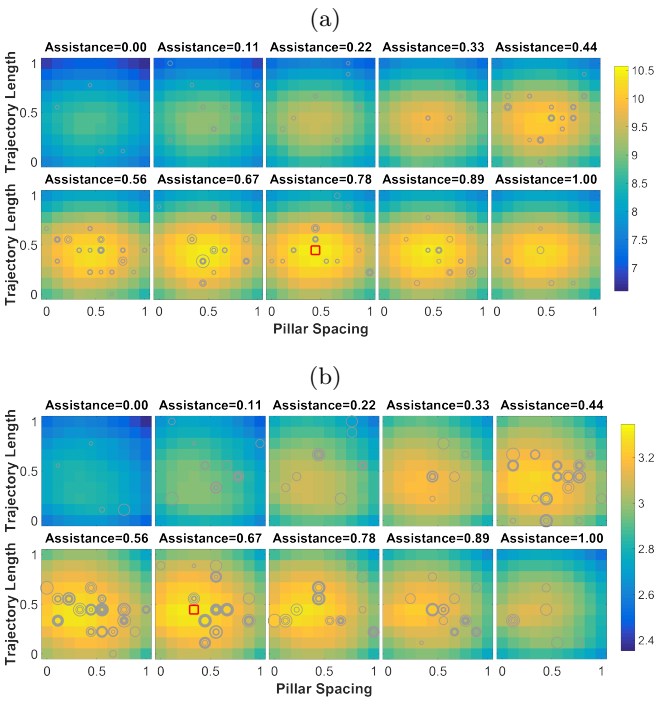

Figure 7. (a) Model predicted VTA (in ninja jumps) over the Spring Ninja design space. Each array shows VTA for a range of trajectory lengths and horizontal spacings between the pillars. The trajectory length and pillar spacing is scaled such that 0 is a challenging game, unlikely to be played well by a novice, and 1 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 radii indicating the magnitudes of the observations. At locations with multiple observations, there are co-centric circles. Red squares indicate the locations of the predicted global maximum. (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.

very close, as indicated by the red squares. The independent but converging evidence supports our conclusions concerning optimal game design. As with Flappy Bird, the Spring Ninja results indicate that the two overt difficulty manipulations have little impact on engagement when no covert assistance is provided (the upper left array), yet with moderate covert assistance, engagement significantly increases.

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 argue that predicted voluntary time may be a better measure than experiential, if the experiential time is used as an anchor to predict the usage time of other individuals. 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. While it is obvious that designing a game of the appropriate difficulty for the user population is needed to make the game engaging, 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 a covert manipulation—in which we provided assistance to the player without the player's awareness—was quite effective. We believe that our covert manipulation of game dynamics in favor of the player was interpreted by players in terms of their own competence, and the sense of competence and accomplishment 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 two 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, 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.

ACKNOWLEDGMENTS

This research was supported by NSF grants SES-1461535, SBE-0542013, and SMA-1041755.

REFERENCES

1. E Andersen, Y Liu, R Snider, R Szeto, and Z Popovic. 2011. Placing a value on aesthetics in online casual games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 1275–1278.
2. J. E. Beck. 2005. Engagement tracing: Using response times to model student disengagement. In *Proceedings of the 2005 Conference on AI in Education*. IOS Press, Amsterdam, 88–95.
3. K Berkling and C Thomas. 2013. Gamification of a software engineering course and a detailed analysis of the factors that lead to its failure. In *Intl. Conf. on Interactive Collab. Learning*. IEEE, 525–530.
4. Jeanne H Brockmyer, Christine M Fox, Kathleen A Curtiss, Evan McBroom, Kimberly M Burkhart, and Jacquelyn N Pidruzny. 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology* 45, 4 (2009), 624–634.
5. P Brusilovsky. 1998. Adaptive educational systems on the world-wide-web: A review of available technologies. In *Proceedings of Workshop on WWW-Based Tutoring at 4th International Conference on Intelligent Tutoring Systems (ITS'98)*. San Antonio, TX.
6. Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*. 2249–2257.
7. Wei Chu and Zoubin Ghahramani. 2005. Preference learning with Gaussian processes. In *In Proceedings of the 22nd International Conference on Machine Learning*. ACM, New York, 137–144.
8. Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and others. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.
9. Simone de Sousa Borges, V. H. S. Durelli, H. M. Reis, and S. Isotani. 2014. A systematic mapping on gamification applied to education. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*. ACM, New York, 216–222.
10. Phil Edwards. 2013. Claw machines are rigged—here's why it's so hard to grab that stuffed animal. `http://www.vox.com/2015/4/3/8339999/claw-machines-rigged`. (2013). Retrieved Sep 15, 2015.
11. Fitocracy. 2015. Fitocracy. (2015). `https://www.fitocracy.com/`
12. J A Fredricks, P C Blumenfeld, and A H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research* 74 (2004), 59–109.
13. R H Hohle. 1965. Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology* 69 (1965), 382–386.
14. David Jurgens, James McCorriston, and Derek Ruths. 2015. An Analysis of Exercising Behavior in Online Populations. In *Ninth International AAAI Conference on Web and Social Media*.
15. Benjamin Katz, Susanne Jaeggi, Martin Buschkuehl, Alyse Stegman, and Priti Shah. 2014. Differential effect of motivational features on training improvements in school-based cognitive training. *Frontiers in Human Neuroscience* 8 (2014), 1–10.
16. Firas Khatib, Frank DiMaio, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, and others. 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology* 18, 10 (2011), 1175–1177.
17. Robert V Lindsey, Michael C Mozer, William J Huggins, and Harold Pashler. 2013. Optimizing Instructional Policies. In *Advances in Neural Information Processing Systems 26*, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.). Curran Associates, 2778–2786.
18. K Linser and T Goschke. 2007. Unconscious modulation of the conscious experience of voluntary control. *Cognition* 104 (2007), 459–475.
19. Y.-E. Liu, T. Mandel, E. Brunskill, and Z. Popovic. 2014. Towards automatic experimentation of educational knowledge. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 3349–3358.
20. John Derek Lomas. 2014. *Optimizing motivation and learning with large-scale game design experiments*. Unpublished Doctoral Dissertation. HCI Institute, Carnegie Mellon University.
21. J. D. Lomas, K. Patel, J. L. Forlizzi, and K. R. Koedinger. 2013. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 89–98.
22. Y R Miyamoto, C A Coleman, J J Williams, J Whitehill, S O Nesterko, and J Reich. 2015. Beyond Time-on-Task: The Relationship between Spaced Study and Certification in MOOCs. http://dx.doi.org/10.2139/ssrn.2547799, *Journal of Learning Analytics* (2015). Accessed: 2015-01-09.
23. H L O'Brien and E. G. Toms. 2010. The development and evaluation of a survey to measure

user engagement. *J. of the Am. Soc. for Information Science and Technlogy* 61 (2010), 50–69.

24. B. Roads and M. C. Mozer. 2015. Improving Human-Computer Cooperative Classification Via Cognitive Theories of Similarity. (2015).

25. Jeffrey N Rouder, Jun Lu, Paul Speckman, DongChu Sun, and Yi Jiang. 2005. A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review* 12, 2 (2005), 195–223.

26. D Sharek and E Wiebe. 2015. Measuring Video Game Engagement Through the Cognitive and Affective Dimensions. *Simulation & Gaming* 45, 4-5 (Jan. 2015), 569–592.

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

28. J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing* 5 (2014), 86–98.

29. A Zook, E Fruchter, and M O Riedl. 2014. Automatic playtesting for game parameter tuning via active learning. In *Proc. of the 9th Intl. Conf. on the Foundations of Digital Games*, T Barnes and I Bogost (Eds.). Soc. for the Adv. of the Science of Digital Games, Ft. Lauderdale, FL.