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
University of Colorado
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 player persistence, projections of how long others will play, and a post-game survey. Using Gaussian process surrogate-based optimization, we conduct efficient experiments to identify game design characteristics—specifically those influencing difficulty—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 subtle and possibly covert manner (incremental trajectory corrections). Converging results indicate that overt difficulty manipulations are effective in modulating engagement only when combined with the covert manipulation, suggesting the critical role of a user's self-perception of competence.

Author Keywords

engagement; persistence; motivation; games; difficulty manipulation; optimization; Gaussian processes

ACM Classification Keywords

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

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858253

INTRODUCTION

Interest has recently surged in applying game-like mechanics to enhance engagement in a variety of domains, such as personal health [9, 14], scientific discovery [16, 6], and education [7, 25, 24, 23]. This research is based on the hypothesis that increased engagement will improve user experiences, data collection, and outcomes. Although engagement is a broad construct [10], it has been operationalized via subjective-self reports, physiological measures, player preferences, and observations of in-game player behavior [27]. In the present work we measure engagement using player persistence, which has been explored previously in game engagement research [25, 36] and in the gambling psychology literature [15], as well as using projections of other players' persistence and post-experiment subjective surveys engagement.

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 [37, 13], reduce uncertainty about an underlying cognitive model [30], and maximize success or persistence rates in an educational game [23]. In contrast, *dynamic manipulations* modulate game design parameters on the fly in reaction to player behavior or performance; examples include matching simulated players' skill [2], selecting exercises that maximize information per mistake [18], adapting AI to move into game states that maximize persistence [11], and adjusting the user interface to maximize performance [26]. Static and

dynamic difficulty manipulations can be combined to discover a dynamic difficulty adjustment (DDA) scheme that works across individuals [21]. 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 [17]. players are also more engaged when they told that the game has adaptive AI, when it actually does not [8]. In these two examples, the player's perception of difficulty was manipulated, through suspense and pregame 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.

The goal of our work is to design games that maximize engagement for a population of users via manipulation of static difficulty. 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 optimize over design spaces, and we will present a novel approach 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 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. [25] 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. [23] 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 [24] 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 (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 $\boldsymbol{\beta} \equiv \{\beta_0, \beta_1, \beta_2\}$ are unknown. If we randomly assign visitors to conditions, as in A-B testing, we will collect many noisy (x, y) observations. We can the fit the $\boldsymbol{\beta}$ 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 *uncer-*

tainty 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 $\boldsymbol{\beta}$.) 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 [35], 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.

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 [26]. 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. [37] 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 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 [5], two-alternative forced choice with guessing [21], and similarity judgment [31]. 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. [23], 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 1a shows the empirical persistence across designs, and Figure 1b 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 engagment over the design space that is consistent with that obtained by the approach of Liu et al. [23], 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

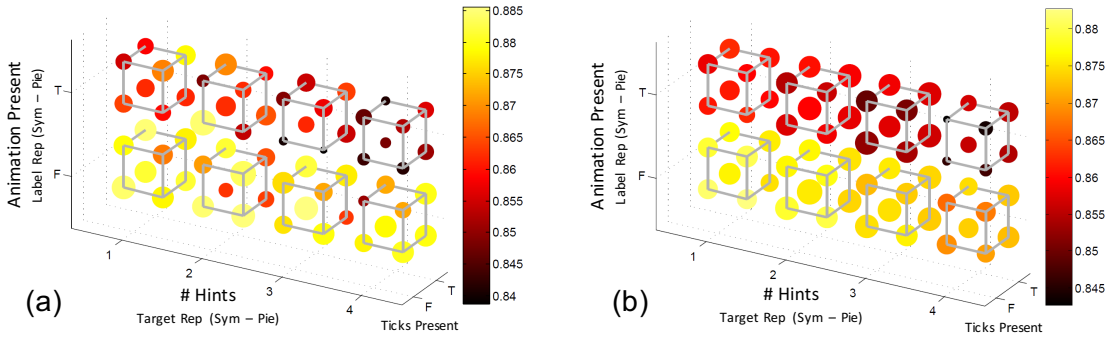

Figure 1. Persistence probability of Treefrog Treasure players over the design space: (a) empirical mean; (b) GP-posterior mean. Each disk represents a design. Color denotes persistence and diameter is inversely proportional to variance of (a) aggregated observations and (b) the GP posterior.

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 [12] or Weibull density [32]. Evidence about usage-duration distributions is harder to find. Miyamoto et al. [28] 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 well 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) 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:

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

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

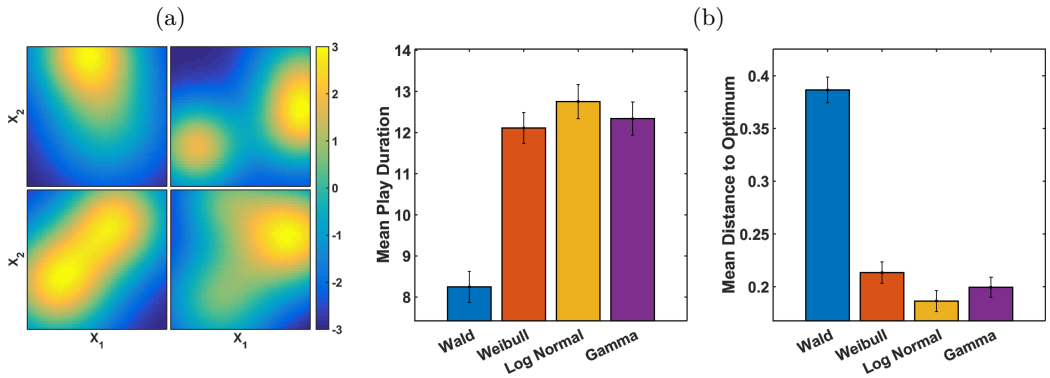

Figure 2. Simulation experiment. (a) Examples of the 2D functions used for generating synthetic data. (b) Results of synthetic experiment. The left and right plots depict the mean function value (higher is better) and the mean distance to the true optimum (lower is better) for various observation models. 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.

Gaussians with randomly drawn centers, spreads, and mixture coefficients, defined over a 2D design space. For examples, see Figure 2a. 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 [4], 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 2b. 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 [34]. 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.

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., [29, 27]). Recently, however, VTA-like measures have been explored. Sharek and Wiebe [33] 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 [25, 24, 23], 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 [15].

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* [19]. 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 [17, 8] 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 [22]. 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.

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 3a). 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 3b). 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 3b)—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, 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.

Flappy Bird: Experimental Methodology and Results

We conducted two studies with Flappy Bird. In the first study, we tested 958 participants. Each participant was

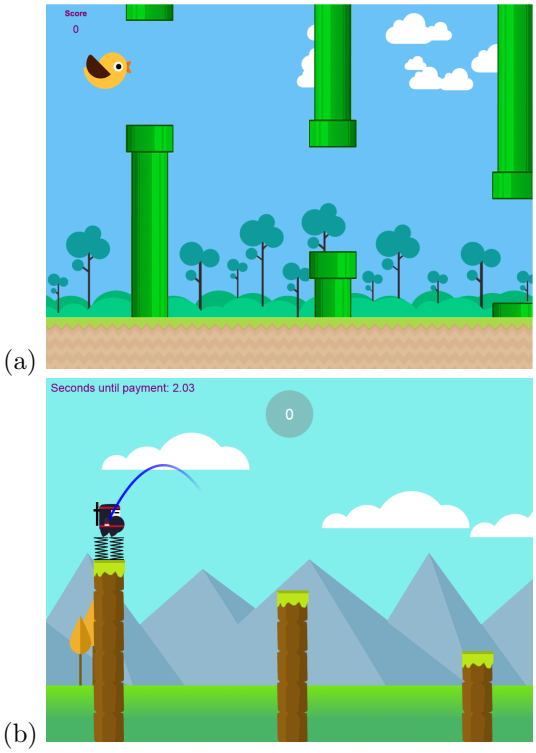

Figure 3. (a) Flappy Bird: The player flaps bird's wings to keep it aloft and to avoid hitting pipes. (b) Spring Ninja: The player jumps from one pillar to another by compressing springs in the ninja's shoes. The blue trajectory is the projected jump path for the given spring compression level.

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 1). 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 [20] 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.[1] 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 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 4). 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 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 4) were taken from the Game Engagement Questionnaire [3].

[1]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.

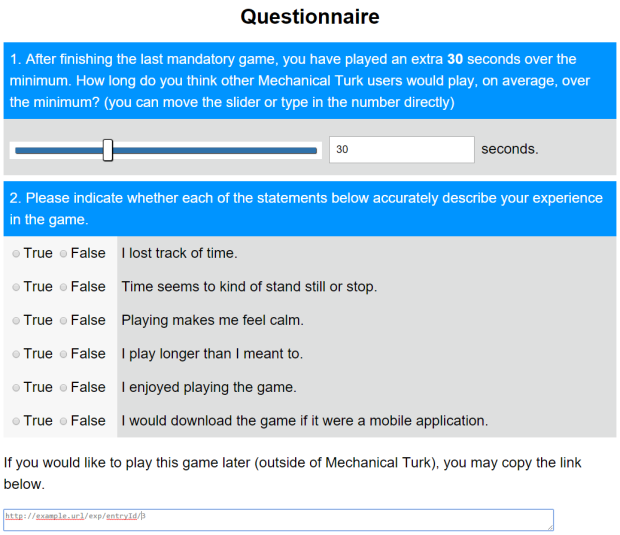

Figure 4. The post-experiment questionnaire.

Figure 5a and 5b 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.

An important question here is whether the predictions of the model are related to the observations. 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 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 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 5a and 5b, 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 5c 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 red squares—lie in almost exactly the same place in 5a, 5b and 5c. 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.

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. We did not conduct a random-assignment study with Spring Ninja due to time constraints and because, in addition

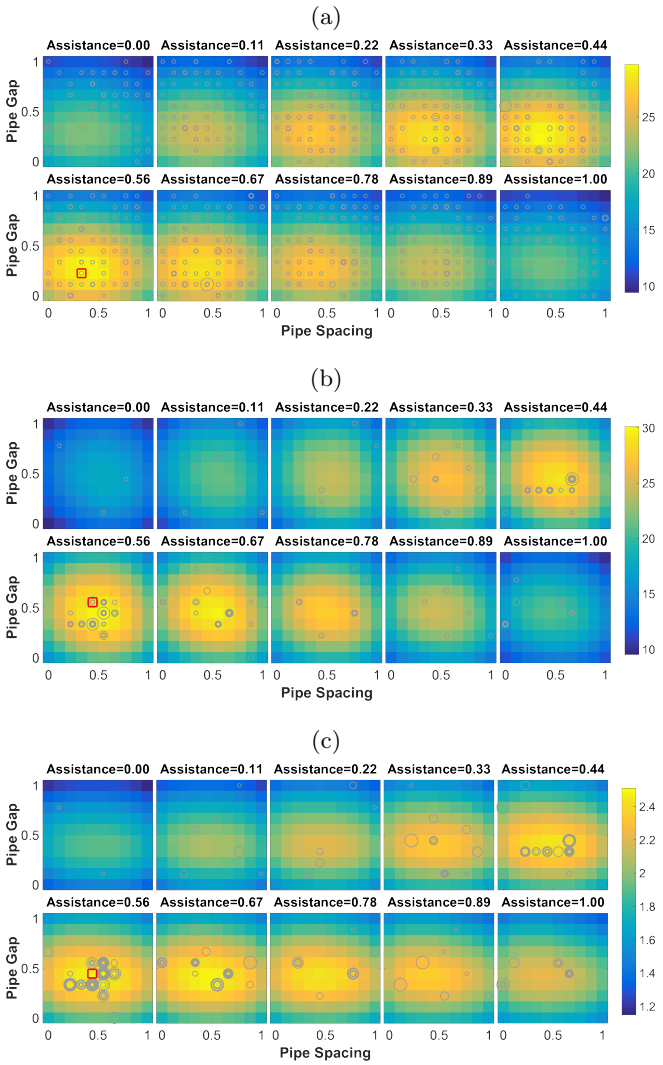

Figure 5. Bayesian model fits of VTA (in seconds) over the Flappy Bird design space for (a) the random-assignment and (b) active-selection studies. Each array corresponds to a fixed level of assistance, with the upper left array being no assistance (level 0) and the lower right array being maximal assistance (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.

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 4).

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.

Figures 6a and 6b 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 is obtained for the mid-range of design parameter settings. The predicted optima by the two measures are very close, as indicated by the red squares. Consistent with the Flappy Bird study, 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

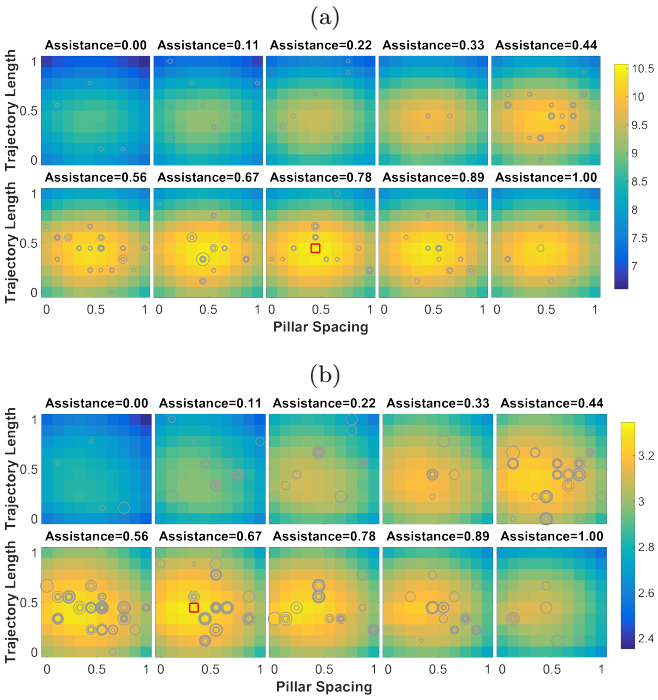

Figure 6. (a) Model predicted VTA (in ninja jumps) over the Spring Ninja design space. Each array shows VTA for a range of trajectory lengths and horizontal spacings between the pillars. The trajectory length and pillar spacing is scaled such that 0 is a challenging game, unlikely to be played well by 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.

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. 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 explicity evaluate the player's perception of task difficulty under different levels of covert assistance, rather than relying on anecdotal evidence. 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.

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. G. Andrade, G. Ramalho, H. Santana, and V. Corruble. 2005. Challenge-sensitive action selection: an application to game balancing. In *Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*. 194–200. DOI: http://dx.doi.org/10.1109/IAT.2005.52
3. 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.
4. Olivier Chapelle and Lihong Li. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*. 2249–2257.
5. 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.
6. 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.
7. 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.
8. Alena Denisova and Paul Cairns. 2015. The Placebo Effect in Digital Games: Phantom Perception of Adaptive Artificial Intelligence. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. ACM, 23–33.
9. Fitocracy. 2015. Fitocracy. (2015). https://www.fitocracy.com/
10. 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.
11. B. Harrison and D.L. Roberts. 2013. Analytics-driven dynamic game adaption for player retention in Scrabble. In *Computational Intelligence in Games (CIG), 2013 IEEE Conference on*. 1–8. DOI: http://dx.doi.org/10.1109/CIG.2013.6633632
12. R H Hohle. 1965. Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology* 69 (1965), 382–386.
13. Aaron Isaksen, Dan Gopstein, and Andy Nealen. 2015. Exploring game space using survival analysis. In *Foundations of Digital Games*.
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. Jeffrey I Kassinove and Mitchell L Schare. 2001. Effects of the" near miss" and the" big win" on persistence at slot machine gambling. *Psychology of Addictive Behaviors* 15, 2 (2001), 155.
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. Christoph Klimmt, Albert Rizzo, Peter Vorderer, Jan Koch, and Till Fischer. 2009. Experimental evidence for suspense as determinant of video game enjoyment. *CyberPsychology & Behavior* 12, 1 (2009), 29–31.
18. Janne V. Kujala, Ulla Richardson, and Heikki Lyytinen. 2010. A Bayesian-optimal principle for learner-friendly adaptation in learning games. *Journal of Mathematical Psychology* 54, 2 (2010), 247 – 255. DOI: http://dx.doi.org/10.1016/j.jmp.2009.10.001
19. Ellen J Langer. 1975. The illusion of control. *Journal of personality and social psychology* 32, 2 (1975), 311.
20. G Levy. 2002. An introduction to quasi-random numbers. (2002). http://www.nag.co.uk/IndustryArticles/introduction_to_quasi_random_numbers.pdf
21. 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.
22. K Linser and T Goschke. 2007. Unconscious modulation of the conscious experience of voluntary control. *Cognition* 104 (2007), 459–475.
23. 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.

24. Derek Lomas, Jodi Forlizzi, Nikhil Poonwala, Nirmal Patel, Sharan Shodhan, Kishan Patel, Ken Koedinger, and Emma Brunskill. 2015. Interface Design Optimization as a Multi-Armed Bandit Problem. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York.

25. 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.

26. MM Hassan Mahmud, Benjamin Rosman, Subramanian Ramamoorthy, and Pushmeet Kohli. 2014. Adapting interaction environments to diverse users through online action set selection. In *Proceedings of the AAAI 2014 Workshop on Machine Learning for Interactive Systems*. Citeseer.

27. Elisa D. Mekler, Julia Ayumi Bopp, Alexandre N. Tuch, and Klaus Opwis. 2014. A Systematic Review of Quantitative Studies on the Enjoyment of Digital Entertainment Games. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 927–936. DOI: http://dx.doi.org/10.1145/2556288.2557078

28. 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.

29. 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.

30. Anna N Rafferty, Matei Zaharia, Thomas L Griffiths, and others. 2012. Optimally designing games for cognitive science research. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

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

32. 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.

33. 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.

34. R. Sifa, C. Bauckhage, and A. Drachen. 2014. The Playtime Principle: Large-scale cross-games interest modeling. In *Computational Intelligence and Games (CIG), 2014 IEEE Conference on*. 1–8. DOI: http://dx.doi.org/10.1109/CIG.2014.6932906

35. Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*. 2951–2959.

36. Ben G. Weber, Michael John, Michael Mateas, and Arnav Jhala. 2011. Modeling Player Retention in Madden NFL 11. In *Innovative Applications of Artificial Intelligence (IAAI)*. AAAI Press, AAAI Press, San Francisco, CA.

37. 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.