Player Modeling via Multi-Armed Bandits

Robert C. Gray
Drexel University
robert.c.gray@drexel.edu

Jichen Zhu
Drexel University
jichen.zhu@gmail.com

Danielle Arigo
Rowan University
arigo@rowan.edu

Evan Forman
Drexel University
emf27@drexel.edu

Santiago Ontañón
Drexel University
so367@drexel.edu

ABSTRACT

This paper focuses on building personalized player models solely from player behavior in the context of adaptive games. We present two main contributions: The first is a novel approach to player modeling based on *multi-armed bandits* (MABs). This approach addresses, at the same time and in a principled way, both the problem of collecting data to model the characteristics of interest for the current player and the problem of adapting the interactive experience based on this model. Second, we present an approach to evaluating and fine-tuning these algorithms prior to generating data in a user study. This is an important problem, because conducting user studies is an expensive and labor-intensive process; therefore, an ability to evaluate the algorithms beforehand can save a significant amount of resources. We evaluate our approach in the context of modeling players' *social comparison orientation* (SCO) and present empirical results from both simulations and real players.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing** → **User models**.

KEYWORDS

Multi-armed Bandits, Player Modeling, Experience Management, Social Comparison

ACM Reference Format:
Robert C. Gray, Jichen Zhu, Danielle Arigo, Evan Forman, and Santiago Ontañón. 2020. Player Modeling via Multi-Armed Bandits. In *Proceedings of FDG '20, September 15–18, 2020, Bugibba, Malta.* ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION

Player modeling focuses on modeling and predicting player characteristics of interest, such as preferences, skill level, or behavior [31]. One of the reasons player modeling is interesting is because it plays a key role in the creation of adaptive games. In this paper, we present two main contributions to player modeling: (1) a novel player modeling approach based on *multi-armed bandits* (MABs), and (2) an approach for evaluating and fine-tuning these algorithms before having access to real player data in a user study.

A common approach to player modeling is the use of machine learning (ML) [9, 26]; however, ML algorithms typically require large amounts of training data. Our proposed approach to player modeling based on MABs [2] solves both (1) this problem of training data acquisition as well as (2) the problem of how to use the player model to adapt the game.

We address the challenge of adapting a game to achieve a desired effect on the player by periodically choosing one among many potential ways to adapt the game. After observing the player's behavior in response to the bandit's choice, a reward value is generated based on the efficacy of that choice, which the bandit observes. We assume a lack of any prior training data before the user starts interacting with the system (though pre-existing data can be exploited). MAB strategies naturally solve this problem by balancing *exploration* (i.e., trying new ways to adapt the game to improve its understanding of the player) and *exploitation* (i.e., adapting the game in ways that have proved to work well in the past).

Moreover, a second problem needs to be solved to effectively deploy this strategy, which constitutes our second contribution. Consider the common problem of choosing the appropriate AI approach before performing an adaptive game user study. How can we gain insights into which AI approaches would be best suited for our user study before engaging in the resource-intensive activity of actually carrying out the study? Additionally, how do we design the parameters of the user study (e.g., participants, duration) without knowing how the AI will perform? To solve this problem, we leverage publicly available data to create simulated players that exhibit statistical behavior patterns close to actual humans. Through the use case of modeling social comparison orientation (SCO) to maximize motivation toward physical activity, we show the promise of our approach as well as the effectiveness of our simulated players to evaluate MAB algorithms.

In the remainder of this paper, we first present some background on player modeling, adaptive games, and MABs. We then present our MAB player modeling framework, followed by our methodology for creating simulated players. Finally, we present empirical results from simulations (with the simulated players) and a real user study.

2 BACKGROUND AND RELATED WORK

This section briefly introduces some basic concepts of player modeling, adaptive games, multi-armed bandits, and simulated player-based evaluations.

FDG '20, September 15–18, 2020, Bugibba, Malta

2.1 Player Modeling and Adaptive Games

Adaptive games leverage knowledge of the player to automatically adapt to better serve specific users or specific design goals [3, 22, 24, 28, 34]. These methods often rely on player modeling to detect or predict a set of characteristics of the player that can inform the AI's decisions [31]. Previous work has shown applications in improving learning outcomes [25] and health outcomes [33], adjusting game difficulty [1, 14], managing user interfaces [12], or even adapting game narratives [19, 23, 28].

Player models are often designed to leverage either *a priori* theory and heuristics ("top-down") or assumption-free statistical methods ("bottom-up") to perform their task of differentiating or defining players, where this dichotomy has also been suggested to define a spectrum [31]. Our work is informed by both of these approaches; though our application domain is derived from psychology theory, we also wish to leverage the statistical power of context-agnostic modeling in the form of an MAB strategy. Specifically, we based our work on the psychology theory of social comparison (see Section 3.1). Though heuristic-based models exist for classifying users based on their social comparison tendencies [13], our work employs a bottom-up approach using MAB strategies.

2.2 Multi-Armed Bandits

A multi-armed bandit (MAB) problem [2] is a class of sequential decision problem where an agent needs to iteratively choose one among k actions (called *arms*), after which it receives a stochastic reward. This mirrors the problem faced by a player in a casino deciding on which of the different gambling machines each of their tokens should be spent. The goal of an MAB strategy is to balance exploration and exploitation, assimilating new knowledge from rewards to "converge" on the arm with the maximum expected return as quickly as possible. Popular MAB strategies include the ϵ-greedy strategy, where the arm that has historically returned the highest reward is always selected except in a portion of iterations (designated by ϵ) where a random arm is chosen. Another is UCB1 [2], which considers the upper confidence bound of expected rewards.

MAB strategies are interesting for adaptive games if we consider the different adaptation options as the arms in an MAB. Investigations into this have already begun in the context of adaptive interventions such as those that promote behavior change [10] with promising results. However, to the best of our knowledge, MABs have not been used in the context of adaptive games or player modeling.

Applying MABs to player modeling raises an important challenge, however. There is a very large collection of MAB strategies proposed in the literature, each with their own practical and theoretical properties. These strategies are usually evaluated by their behavior "in the limit" (i.e., with large numbers of interactions with the environment). However, in a player modeling situation, we cannot expect the system to interact with players for this long. This means that MAB strategies that work with very few interactions with users are needed, and thus we had to design an MAB strategy that satisfies these constraints before carrying out the study.

Therefore, our work pushes the state of the art in two separate ways. First, we present a novel player modeling framework based on MAB strategies. Second, we present an approach for evaluating strategies via simulated players to design an MAB strategy that is effective with very few player interactions.

2.3 AI Tuning via Simulation

Carterette et al. [6] present a conceptualization of system evaluations as a continuum between *systems-based* approaches involving automated tests that evaluate predetermined scenarios and *user studies* involving real user interactions with the system. The former are viewed to have the advantages of stability, repeatability, and low costs at the risk of oversimplifying assumptions that could invalidate results. The latter are capable of answering more questions with potentially higher accuracy, but in exchange they carry a burden of higher expense and variability. This is referred to as the *bias-variance tradeoff* [6].

For adaptive games and player modeling, it is difficult to escape the requirement of genuine user studies; however, researchers have found value in simulations for a number of situations. These may include the rarity of real players [7], the complexity of the test space [27], a desire to maintain specific control over how a model is trained [17], or the need to train an AI via techniques that require very large data sets [30].

3 PLAYER MODELING VIA MABS

The key idea behind our multi-armed bandit approach is that the MAB strategy serves both as (1) the method by which we model players and (2) the AI that adapts the game to guide player experience in real time. Our approach consists of 3 main components:

- The **arms**: the set of possible ways in which the game can be adapted and from which the MAB strategy chooses each time it needs to adapt the game to the player.
- The **reward**: the numerical quantity the MAB strategy will aim to maximize, such as the player's daily steps in an exergame designed to encourage walking.
- The **MAB strategy**: the algorithm that chooses an arm, observes the reward resulting from that choice, and updates its internal model of the player to make the next decision.

The execution cycle (Figure 1) works as follows:

(1) Initially, the MAB strategy has no information about the player at hand (however, pre-existing individual or population information could be used to initialize the strategy).
(2) The MAB strategy selects one of the possible arms, and the game is adapted as designated by the arm.
(3) The player continues to interact with the game within this adaptation, which results in some measurable metric or metrics that render a "reward" value.
(4) The MAB strategy observes this reward and updates its internal model of the player.
(5) The cycle repeats, and the MAB strategy chooses again.

MAB strategies aim to balance exploration and exploitation, deciding when to *exploit* the arm that is currently believed to be the best for the player (according to the objective encoded in the reward function) and when to *explore* a different arm in order to learn more about the current player. Therefore, in addition to making the necessary decisions to adapt the game, MAB strategies naturally

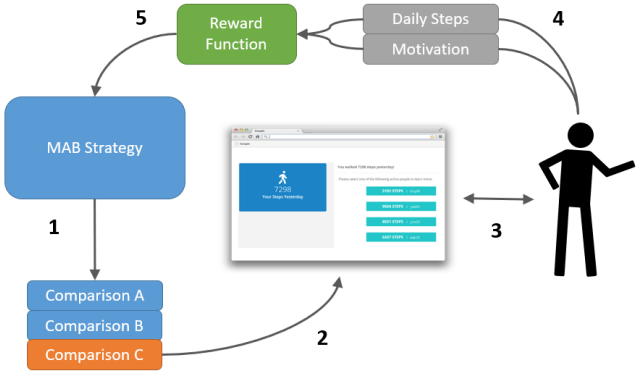

Figure 1: Execution cycle for our MAB strategy: 1) MAB strategy selects among multiple configurations of comparisons to present to the player. 2) Comparison options are presented. 3) Player interacts with the software. 4) Metrics are generated (daily steps and self-reported motivation). 5) Reward is observed and recorded by the MAB strategy.

facilitate methods for obtaining the necessary data to update the player model.

3.1 Modeling SCO via Multi-armed Bandits

Social comparison is a psychological process in which individuals use comparisons to others, often subconsciously, to assess their degree of success (self-evaluation), to plan for future success (self-improvement), or to view themselves in a favorable light (self-enhancement) [29]. Even when objective standards are available, comparisons to salient others can still be preferred and potentially even more influential [18]. This carries into gaming, where leveraging social comparison in team competitions has been shown to be effective in influencing participant motivation toward increasing physical activity (PA) [32].

The details that govern the ways in which a person conducts these comparisons are regarded as individual traits that can be described in aggregate as the person's social comparison orientation (SCO) [13], which includes their tendency to perform comparisons, their preference in seeking out targets, and the influence that such comparisons have on their future behavior [4]. Specifically, our research is interested in modeling the degree to which an individual tends to seek out comparison targets performing better than they are (i.e., *upward* comparisons) or worse than they are (i.e., *downward* comparisons). As discussed later, our simulated players model these features.

In our broader research of leveraging motivation psychology toward improving engagement and efficacy of game-related interventions, particularly social exergames [5], we seek to model player SCO within adaptive games in a way that can provide dynamic and individualized experiences. This paper is a first step in this direction, where we evaluated our SCO player modeling approach in a simpler web-based intervention that gave players an opportunity to log in and compare themselves against other profiles. In our application, we instantiated the three elements of the MAB player modeling framework as follows (Figure 1):

- **Arms:** The MAB strategy had an opportunity each day to choose which comparison opportunities were displayed. Our setup had 3 arms: arm "A" presented the player with zero upward comparison opportunities (i.e., all other displayed profiles walked fewer steps than the player); arm "B" presented the player with two upward and two downward, and arm "C" offered the player four upward comparisons. It is expected that a player's act of comparing themselves to these profiles, depending on their individual SCO, would result in a change in motivation. Once the configuration was chosen, profiles were presented, and the player was given an opportunity to investigate more details of only one of the profiles.
- **Reward:** The player's eventual steps s that day following the session as well as a self-reported motivation score m on a 5-point Likert scale following the session (players reported their motivation before the session as well) were used to calculate a reward score r_t using the following formula (where μ and σ represent mean and standard deviation with respect to all previously observed data for that player):

$$r_t = \frac{\frac{s_t - \mu_s}{\sigma_s} + \frac{m_t - \mu_m}{\sigma_m}}{2}$$

- **MAB strategy:** we evaluated a large collection of strategies enumerated in Section 4.4.

The next section describes the approach we used to evaluate the different MAB strategies and parameters by using simulated players. As detailed later in the paper, we then evaluated the best performing strategy with real players.

4 SIMULATED PLAYERS

The purpose of creating simulated players was to evaluate different MAB strategies while modeling players that exhibit similar statistical trends as real users (i.e., same variance in numbers of steps per day). This was crucial in our case, as it was unclear whether any MAB strategy would converge fast enough given the expected duration of the study and the large degree of noise present in real human data. Our simulated players had three main components:

(1) Step Model: A probabilistic model that simulated the number of steps typical humans take in a day.
(2) SCO Data Model: A representation of a player's tendency toward upward and downward comparisons.
(3) SCO Behavioral Model: A set of functions implementing player behavior given the step model, the SCO data model, and the player's social comparison activities.

4.1 Step Model

In order to obtain a realistic step model, and in consideration for the bias-variance tradeoff discussed by Carterette et al. [6], we opted to leverage existing behavioral data. Specifically, we obtained data from a publicly available Mechanical Turk survey conducted over three months in 2016 by Furberg et al. [11]. After omitting days with zero steps, we confirmed via D'Agostino-Pearson and Shapiro-Wilk tests that the data was not from a normal distribution (both $p < 0.01$). Previous research has suggested human walking patterns align with gamma distributions [21], reflecting a common

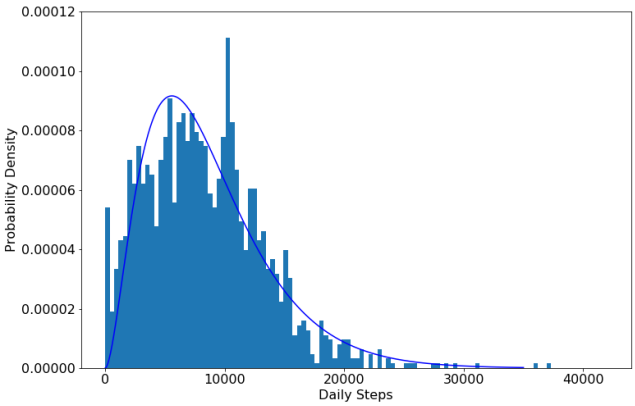

Figure 2: Daily step data from the Mechanical Turk experiment [11]. Histogram is overlayed with a probability density function curve for a Gamma distribution with k=2.8, θ=3100.

trend among intermittent human behaviors [16], which we used to fit the data (Figure 2).

4.2 SCO Data Model

We considered the following SCO traits for the simulated players:

(1) Direction: A propensity for a player to more often make (deliberately or subconsciously) upward or downward comparisons. This is referred to as the player's directional *preference* for social comparison [29].
(2) Intensity: The general degree of influence that SCO activities have in a simulated player's motivation and behavior.

To achieve this, two parameters $0 \leq u \leq 1, 0 \leq d \leq 1$ were used that represent the simulated player's affinity toward upward and downward comparisons on a linear scale. This was chosen as two separate variables to reflect the design of the common psychology instrument used to measure SCO–namely, the upward and downward comparison subscales of the Iowa-Netherlands Comparison Orientation Measure (INCOM) [13].

The propensity to prefer one comparison over another (1) is modeled as the proportion defined by the simulated player's u and d values. E.g., an assignment of $(0.8, 0.4)$ would indicate a 2x preference toward upward comparisons. A simulated player's general sensitivity to either comparison (2) is modeled by the magnitude of the value. E.g., an assignment of $(0.0, 0.5)$ would designate a simulated player not influenced at all by upward comparisons but moderately influenced by downward comparisons.

4.3 SCO Behavioral Model

The simulated players were given a programmatic version of the same exercise intended for real human users in an upcoming user study. The details of this exercise are explained in further detail in Section 5.2.1, and involve a repeated interaction over the course of 21 days (i.e., time steps). In each time step, the simulated player is given a list of four *profiles* depicting the PA behavior and other details for four realistic (but fabricated) people. The PA performance of these profiles would be strategically generated to provide upward or downward comparisons for the player, according to the simulated player's own steps the previous day and the MAB strategy's assessment of their preference for social comparison. The simulated player then chooses to view one of the profiles in detail, and (presumably influenced by that experience of comparing their PA output to that of another) afterward generates a value for their "steps" that day.

The decisions made in this process and the value of the generated steps were influenced by the simulated player's internal SCO data model via the behavioral models described below. Specifically, each simulated player was equipped with three behavior models: *selector*, *step simulator*, and *motivation*.

The *selector* component considers the list of the four potential player profiles for comparison and chooses one of them. The choice (resulting in a *comparison target* for that day) is determined by the simulated user's underlying (u, d) values, where a direction preference is stochastically selected, weighted by u and d. E.g., if the simulated player had values $(0.4, 0.2)$, they would be twice as likely to choose an upward comparison than a downward comparison. Once the direction is determined, the user selects randomly among the choices available in that direction. If no choices are available in that direction (e.g., the simulated user chose to select downward today but the MAB strategy chose Arm C and provided four upward comparisons), a random choice is made from the remainder.

The *step simulator* component reports the simulated user's daily steps following this selection and resulting comparison event. To do so, the step model is queried to sample a number of steps s'_t from the gamma distribution discussed in Section 4.1. This is further modified by the comparison target's performance, s^c_t as well as (u, d). Specifically, the number of steps s_t reported for an upward comparison is:

$$s_t = s'_t \left(1 + u \frac{s^c_t - s'_t}{s'_t}\right)$$

and

$$s_t = s'_t \left(1 + d \frac{s'_t - s^c_t}{s'_t}\right)$$

for a downward comparison.

The *motivation* component enables the simulated player to self-report their motivation both before and after a comparison. Because no public data existed for user motivation reporting, our simulated players select uniformly at random from values 2, 3, and 4 in the 5-point custom Likert scale used for self-reporting motivation. Motivation reported after the comparison is determined by an aggregate affect value calculated as $u - d$ for an upward comparison and $d - u$ for a downward comparison. If this aggregate affect value is positive, then a motivation score higher than or equal to the initial value is randomly selected; if it is negative, than a motivation score lower than or equal to the initial value is randomly selected; and if it is zero, then a motivation score equal to or adjacent to (higher and lower) the initial motivation is randomly selected.

4.4 MAB Strategies

We conducted experiments on our simulation reflecting the user flow of the anticipated user study, using simulated players (instead

of real players) and multiple MAB strategy variants. Specifically, we compared the following MAB strategies:

- Random: used as a baseline strategy for evaluation where the arms are always selected at random.
- UCB1 [2]: calculates a score for each arm based on past average reward and a confidence interval (arms selected less often have less data and therefore lower confidence in expected value). UCB1 balances exploiting arms with proven rewards and exploring arms not yet selected enough to create tight confidence intervals.
- ϵ-greedy: selects the best historically performing arm except for a certain percentage of the time (designated by ϵ) where the strategy will randomly explore another arm.
- ϵ-first: selects randomly at the beginning of the experiment until a point specified by ϵ, after which the best historically performing arm is always selected.
- ϵ-decreasing with linear decay: similar to ϵ-greedy, except that the ϵ parameter starts higher and gradually decreases to a lower value over a specified number of steps.
- ϵ-decreasing with exponential decay: an ϵ-decreasing implementation that decreases the ϵ factor according to a specified exponential decay curve rather than a linear schedule.

4.4.1 Regression Strategy Variants. Each of the MAB strategies listed above maintains an internal expectation of the reward for each arm, calculated as the average of all rewards previously reported when selecting that arm. Since the amount of variance we observed in the publicly available data was very large, we designed another set of strategies equipped with better score estimation techniques.

Specifically, we used linear regression involving reported steps and motivation over the previous days to predict the expected score that would be obtained when selecting a given arm in the next decision. We are aware that such a modification challenges the status of these approaches as MAB strategies, because MAB strategies are stateless; rather, the introduction of reward prediciton via linear regression aligns more with a general reinforcement learning paradigm. But for simplicity, we present these strategy variants as modified MAB strategies.

Running linear regressions for each day in the Mechanical Turk data set and observing p-values for the correlations of each feature, we performed iterative backward elimination to ultimately end with a model in which all features correlated with a significance of $p < 0.05$. Our final model for predicting a person's daily steps from historical data consisted of their steps 1, 2, 3, 4, 6, and 7 days prior, as well as whether the day of the week was Monday or Friday. Because no such data existed for motivation, we chose an initial approach of building our regression model from the three previously reported motivation values.

We incorporated this regression model into the ϵ-based strategies, where we replaced the *mean* operation with this linear regression, resulting in "regression-based" variants for each of them (e.g., "regression-based ϵ-decreasing with exponential decay").

5 RESULTS

In this section, we examine both the results of our simulator runs to prepare for the user study and the results of the user study deploying the MAB strategy that worked best in simulation.

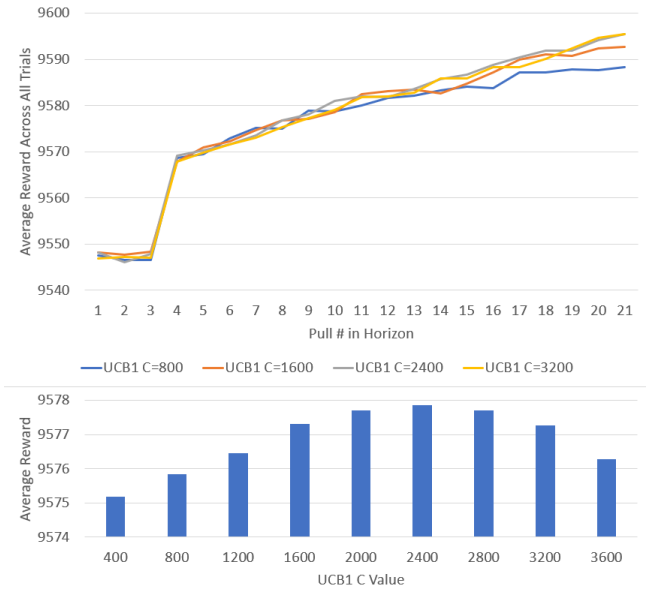

Figure 3: Average reward (vertical axis) obtained using a UCB1 MAB strategy for different values of C over time (horizontal axis, top), and global averages (bottom).

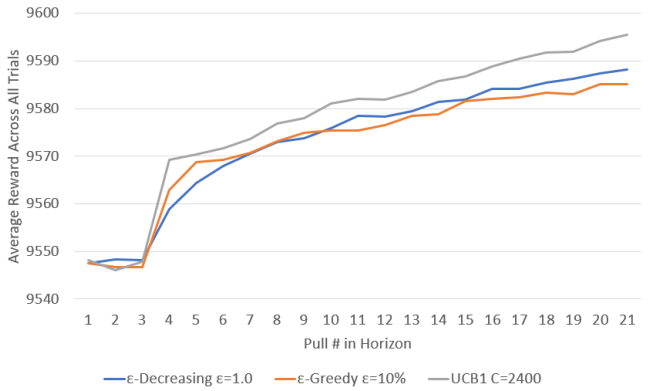

Figure 4: Average reward (vertical axis) for three different MAB strategies (UCB1, ϵ-greedy, and ϵ-decreasing) over time (horizontal axis).

5.1 Results of Simulations

We performed three sets of experiments, where in each a specific MAB configuration was tested with a given simulated player. Each experiment conducted N experimental trials, where an experimental trial consisted of the simulated player interacting with the MAB strategy over M steps (simulated days). In each step, the MAB strategy was queried for its decision, which would be given to the simulated player, and the player would simulate behavior in response (e.g., selecting one of the four presented profiles and generating resulting steps for that day). This would be reported to the MAB strategy, which would update its internal statistics. The MAB state was reset at the beginning of each trial. We report the average

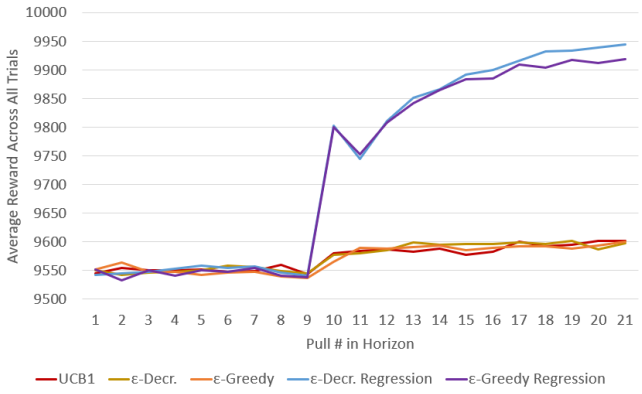

Figure 5: Average reward (vertical axis) for three different MAB strategies over time (horizontal axis) with a nine-step forced exploration period, comparing strategies with linear regression and without.

number of steps performed by the simulated players each of the simulated days over all the experimental trials. We used $u = 0.3$ and $d = 0.6$ for all the simulated players in our experiments (estimating the ranges of u and d values that most resemble human behavior is part of our future work).

5.1.1 UCB1 C-Value Experiment. Some strategies needed to fine-tune their parameters to the specific task, such as the ϵ parameter in ϵ-greedy strategies or the C parameter in UCB1. In our simulation experiment depicted in Figure 3, we report on a set of runs that set out to to tune the C parameter for the UCB1 strategy. Notice the uncommonly high values for C, as a result of the fact that the reward function also returns very high values (far from the usual $[0 - 1]$ interval).

In this experiment, we ran $N = 50$ million trials of UCB1 for $C = k * 400$ for $k \in \{1, ..., 9\}$. The experiments targeted a horizon of $M = 21$ steps in anticipation of a user study lasting 21 days. Our UCB1 implementation requires that every arm be evaluated one time before the strategy engages, which is the cause for the lower results in the first three pulls. From this data, it appears that $C = 2400$ achieved the largest reward in our experimental setup.

In the interest of space, we do not report the detailed results for tuning ϵ-greedy strategies, but the best parameters were found to be $\epsilon = 0.1$ for ϵ-greedy, and $\epsilon = 1.0$ for ϵ decreasing.

5.1.2 Strategy Comparison Experiment. In this experiment, we compared three of the most promising strategies to determine the top candidates to investigate for deployment in our user study.

Figure 4 shows the results of $N = 50$ million trials for each of UCB1 ($C = 2400$), ϵ-decreasing ($\epsilon = 1.0$), and ϵ-greedy ($\epsilon = 0.1$). As before, we targeted a horizon of $M = 21$ steps with a requirement that all test each arm once (steps 1-3) before engaging their strategy (forced exploration). The results suggested an advantage in the UCB1 strategy over the others. However, of the two ϵ-class strategies, ϵ-decreasing appeared to perform the best.

5.1.3 Regression-Based Experiment. The last experiment compared our best performing strategies from Experiment 2 to regression-based strategies. Figure 5 graphs the results of the experiment with $N = 1$ million trials. In this experiment, we also introduced a nine-step *forced exploration* period in which each of the three arms were pulled three times (in random order) before the MAB strategy was engaged, an approach that has been found to be advantageous in short-horizon MAB scenarios [15]. Our results showed that the regression variants performed significantly better than the non-regression models (experiments comparing different forced exploration periods are not reported in the interest of space).

Results from simulated experiments ultimately led to our selection of MAB strategy for the user study: an ϵ-decreasing strategy implemented with an exponential decay of $1/x^{\epsilon}$, an ϵ value of 1.0, and a nine-step forced exploration period.

5.2 Results with Real Users

Finally, we evaluated the best performing MAB strategy from the simulated experiments with real users via a 3-week study. Although the long-term goal of this project is to design a full-fledged game with this technology, we limited ourselves in this study to a web-based activity as a first step toward that goal.

5.2.1 Methodology. The participants were recruited from psychology and digital media courses at Drexel University, where they were informed they would be participating in a study regarding attitudes toward health. They were set up with pedometers (i.e., smartphones equipped with accelerometers and Fitbit software) to track their daily steps and were then directed to engage in daily sessions with a web-based software application. It was requested from each participant that they complete one session (around 5 min.) each day for 21 days, which consisted of the following:

After logging in, participants were asked to rank their motivation to exercise on a scale from "very low" (1) to "very high" (5), after which they were presented with their own step count from the previous day. They were then shown four buttons representing profiles of other people that they could investigate. These profiles were created by the research team and did not represent real people, but they were presented as real and the participants were not informed that they were fabricated. Participants were requested to select a profile among the four options to view additional details regarding that profile beyond simple step count (e.g., diet, hobbies, exercise habits, profession, etc.). The MAB strategy's choice would dictate which profiles would be given in order to offer those comparisons. After the players were done inspecting the selected profile, they were asked again to report their motivation to exercise. Participants were divided into two conditions: experimental (with MAB strategy engaged) and a control condition (that received random arm selections).

5.2.2 Results. A total of 53 people enrolled in the study, but five participants did not finish enough sessions to qualify as having completed the study (at least 14 days). Of the remaining 48 participants, 25 were in the control condition and 23 were in the experimental condition.

Results are shown on Table 1, where we found that participants in the control group saw an average of 42 extra steps on the days

Table 1: Difference in step counts (between previous day and current day) and reported motivation (before and after session) during the intervention period for participants with and without the MAB strategy (2-tailed T-test, α=0.05).

Condition	Steps change	Motivation change
Control	42	0.013
Experimental	160	0.111
T-score:	0.3007	1.9908
p-value:	0.764	0.047

of their sessions (with respect to the day before) compared to 160 extra steps for participants in the experimental group. Though perhaps representing a trend, this finding was not found to be statistically significant (p=0.764) via two-sample T-test at $\alpha = 0.05$ (two-tailed, dof=445). However, the change in motivation before and after inspecting the selected user profile did demonstrate a statistically significant difference (p=0.047) via two-sample T-test at $\alpha = 0.05$ (two-tailed, dof=388), where participants in the control group saw an average motivation score increase of 0.013 compared to an increase of 0.111 in the experimental group.

6 DISCUSSION

Though metrics such as daily steps and player motivation might be affected by many factors, the increase in steps and the statistically significant increase in self-reported motivation suggest that our bandit-driven manipulation of selecting individualized social comparison targets for users was more effective than random assignment. This in turn appears to support our approach of defining our player modeling problem as an MAB problem, to which we were able to apply a wealth of theory and solutions already developed by that field. To our knowledge, this is the first case in which an MAB-based approach has been applied toward player modeling in order to implement an adaptive game.

In our case, this technique for player modeling via MAB strategies allowed us to engage in both a top-down and bottom-up approach to player modeling simultaneously [31]. Theory borrowed from the psychology field of social comparison helped to define the arms for our MAB problem (i.e., enjoying the theory-based insight inherent in top-down modeling), while avoiding the need to provision our classification system with context-specific heuristics to define players (i.e., the advantages of bottom-up modeling). Rather, in our case the mechanism is the model, and the context-agnostic operation of the MAB strategy allowed the system to assess players based simply on a reinforcement loop (i.e., user response in terms of steps and self-reported motivation) instead of the researcher's perceptions or interpretations of player behavior.

Further, these results support the proposed benefits of the technique of implementing an AI-based intervention first as a simulation in order to explore the potential options for the AI. In this practice, simulated users were constructed with data and behavioral models (based on psychology theory) that allowed them to exhibit behaviors on which we conducted multiple experiments. The results of these experiments, which were achieved with greater speed and lower cost than preliminary user studies, informed our decisions prior to recruiting human players for our planned user study.

7 CONCLUSION

This paper presented a new approach to player modeling based on multi-armed bandits (MABs). MABs naturally model both the problem of exposing the player to different situations to build an accurate player model and the problem of adapting a game to maximize features of interest to the designer. We also presented a method for creating simulated players to evaluate and fine-tune these MAB techniques before deploying with real users.

Our results indicated that an ϵ-decreasing strategy with a nine-step forced exploration period and a linear regression model to estimate both steps and motivation performed the best in simulation and therefore was used with real users. Our results showed the difference in step count increment with respect to the previous day and motivation change were both higher for the experimental condition using the MAB, although only the latter was found to be statistically significant. This is no easy task, as the degree of variance in both step and motivation data is very high, and the MAB was able to select arms that achieved positive results in just 21 interactions with the users.

As part of our future work, we plan to improve our simulated user framework to obtain more realistic user behavior models. We also plan to investigate more sophisticated MAB approaches such as contextual bandits [8] or combinatorial bandits [20], which would allow us to integrate state knowledge or engage complex decisions. We are also interested in comparing the estimations built by the MAB strategy with results obtained from standard psychological SCO tests to measure agreement. Finally, our next step is to incorporate our new approach into our game prototype.

8 ACKNOWLEDGEMENTS

This work is partially supported by the National Science Foundation under Grant Number IIS-1816470. The authors would like to thank the participants of our user study and all current and past members of this project. Special thanks to Jennifer Villareale and Diane Dallal for facilitating the user study and user data collection which is used in this paper.

REFERENCES

[1] Justin T Alexander, John Sear, and Andreas Oikonomou. 2013. An investigation of the effects of game difficulty on player enjoyment. *Entertainment computing* 4, 1 (2013), 53–62.
[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. 2002. Finite-time analysis of the multi-armed bandit problem. *Machine learning* 47, 2 (2002), 235–256.
[3] Joseph Bates. 1992. Virtual reality, art, and entertainment. *Presence: Teleoperators & Virtual Environments* 1, 1 (1992), 133–138.
[4] Abraham P. Buunk and Frederick X. Gibbons. 2007. Social comparison: The end of a theory and the emergence of a field. *Organ. Behav. Hum. Decis. Process.* 102, 1 (2007), 3–21.
[5] Karina Caro, Yuanyuan Feng, Timothy Day, Evan Freed, Boyd Fox, and Jichen Zhu. 2018. Understanding the effect of existing positive relationships on a social motion-based game for health. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 77–87.
[6] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. 2011. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. 611–620.
[7] Zhengxing Chen, Su Xue, John Kolen, Navid Aghdaie, Kazi A Zaman, Yizhou Sun, and Magy Seif El-Nasr. 2017. Eomm: An engagement optimized matchmaking framework. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1143–1150.
[8] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 208–214.

[9] Anders Drachen, Alessandro Canossa, and Georgios N Yannakakis. 2009. Player modeling using self-organization in Tomb Raider: Underworld. In *2009 IEEE symposium on computational intelligence and games*. IEEE, 1–8.

[10] Evan M Forman, Stephanie G Kerrigan, Meghan L Butryn, Adrienne S Juarascio, Stephanie M Manasse, Santiago Ontañón, Diane H Dallal, Rebecca J Crochiere, and Danielle Moskow. 2019. Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *Journal of behavioral medicine* 42, 2 (2019), 276–290.

[11] Brinton J. Keating M. Furberg, R. and A. Ortiz. 2016. Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016. http://doi.org/10.5281/zenodo.53894

[12] Krzysztof Z Gajos and Krysta Chauncey. 2017. The influence of personality traits and cognitive load on the use of adaptive user interfaces. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 301–306.

[13] Frederick X. Gibbons and Bram P. Buunk. 1999. Individual differences in social comparison: Development of a scale of social comparison orientation. *J. Pers. Soc. Psychol.* 76, 1 (1999), 129–142.

[14] Robert C Gray. 2018. *Adaptive Game Input Using Knowledge of Player Capability: Designing for Individuals with Different Abilities*. Drexel University.

[15] Robert C Gray, Jichen Zhu, and Santiago Ontañón. 2020. Regression Oracles and Exploration Strategies for Short-Horizon Multi-Armed Bandits. *Proceedings of the 2020 IEEE Conference on Games (CoG)* (2020), accepted.

[16] J-L Guo, C Fan, and Z-H Guo. 2011. Weblog patterns and human dynamics with decreasing interest. *The European Physical Journal B* 81, 3 (2011), 341.

[17] C. Holmgård, A. Liapis, J. Togelius, and G. N. Yannakakis. 2014. Evolving personas for player decision modeling. In *2014 IEEE Conference on Computational Intelligence and Games*. 1–8.

[18] William M Klein. 1997. Objective standards are not enough: affective, self-evaluative, and behavioral responses to social comparison information. *Journal of personality and social psychology* 72, 4 (1997), 763.

[19] Michael Mateas and Andrew Stern. 2003. Façade: An experiment in building a fully-realized interactive drama. In *Game developers conference*, Vol. 2. 4–8.

[20] Santiago Ontañón. 2017. Combinatorial multi-armed bandits for real-time strategy games. *Journal of Artificial Intelligence Research* 58 (2017), 665–702.

[21] Michael S Orendurff, Jason A Schoen, Greta C Bernatz, Ava D Segal, and Glenn K Klute. 2008. How humans walk: bout duration, steps per bout, and rest duration. *Journal of Rehabilitation Research & Development* 45, 7 (2008).

[22] Mark O Riedl, Andrew Stern, Don Dini, and Jason Alderman. 2008. Dynamic experience management in virtual worlds for entertainment, education, and training. *International Transactions on Systems Science and Applications, Special Issue on Agent Based Systems for Human Learning* 4, 2 (2008), 23–42.

[23] Manu Sharma, Santiago Ontañón, Manish Mehta, and Ashwin Ram. 2010. Drama management and player modeling for interactive fiction games. *Computational Intelligence* 26, 2 (2010), 183–211.

[24] David Thue and Vadim Bulitko. 2018. Toward a Unified Understanding of Experience Management. In *Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

[25] Josep Valls-Vargas, Andrew Kahl, Justin Patterson, Glen Muschio, Aroutis Foster, and Jichen Zhu. 2015. Designing and tracking play styles in solving the incognitum. In *Proceedings of the Games+ Learning+ Society Conference, to appear. Games+ Learning+ Society*.

[26] Josep Valls-Vargas, Santiago Ontanón, and Jichen Zhu. 2015. Exploring player trace segmentation for dynamic play style prediction. In *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*.

[27] Viktor Wendel, Johannes Alef, Stefan Göbel, and Ralf Steinemtz. 2014. A method for simulating players in a collaborative multiplayer serious game. In *Proceedings of the 2014 ACM International Workshop on Serious Games*. 15–20.

[28] Peter William Weyhrauch. 1997. *Guiding interactive drama*. Carnegie Mellon University.

[29] Joanne V Wood. 1989. Theory and research concerning social comparisons of personal attributes. *Psychol. Bull.* 106, 2 (1989), 231.

[30] Georgios N Yannakakis and John Hallam. 2004. Evolving opponents for interesting interactive computer games. *From animals to animats* 8 (2004), 499–508.

[31] Georgios N Yannakakis, Pieter Spronck, Daniele Loiacono, and Elisabeth André. 2013. Player modeling. (2013).

[32] J Zhang, D Brackbill, S Yang, J Becker, N Herbert, and D Centola. 2016. Support or competition? How online social networks increase physical activity: a randomized controlled trial. *Prev. Med. reports* 4 (2016), 453–458.

[33] Jichen Zhu, Yuanyuan Feng, Anushay Furqan, Robert C Gray, Timothy Day, Jessica Nebolsky, and Karina Caro. 2018. Towards Extending Social Exergame Engagement with Agents. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 349–352.

[34] Jichen Zhu and Santiago Ontañón. 2019. Experience management in multi-player games. In *2019 IEEE Conference on Games (CoG)*. IEEE, 1–6.