Personal History Affects Reference Points: A Case Study of Codeforces

Takeshi Kurashima[1], **Tomoharu Iwata**[2], **Tomu Tominaga**[1], **Shuhei Yamamoto**[1],
Hiroyuki Toda[1], **Kazuhisa Takemura**[3]
[1]NTT Human Informatics Laboratories, NTT Corporation
[2]NTT Communication Science Laboratories, NTT Corporation
[3]Department of Psychology, Waseda University
{takeshi.kurashima.uf, tomoharu.iwata.gy, tomu.tominaga.hm, shuhei.yamamoto.ea}@hco.ntt.co.jp,
toda.hir.xg@yokohama-cu.ac.jp, kazupsy@waseda.jp

Abstract

Humans make decisions based on their internal value function, and its shape is known to be distorted and biased around a point, which the research community of behavior economics refers to as the *reference point*. People intensify activities that come to lie within the reach of their reference point, and abstain from acts that would incur losses once they've crossed the point. However, the impact of past experiences on decision making around the reference point has not been well studied. By analyzing a long series of user-level decisions gathered from a competitive programming website, we find that history has a clear impact on user's decision making around the reference point. Past experiences can strengthen, and sometimes weaken, the decision bias around the reference point. Experiences of past difficulties can strengthen the tendency towards loss aversion after achieving the reference point. When a person crosses a reference point for the first time, the cognitive decision bias is significant. However, repeating this crossing gradually weakens the effect. We also show the value of our insights in the task of predicting user behavior. Prediction models incorporating our insights may be used for motivating people to remain more active.

Introduction

Background. According to *Prospect Theory*, people make decisions based on a psychological value function; people estimate the value of outcomes relative to a *reference point*, and perceive outcomes above (beyond) the point as gains, and outcomes below (before) the point as losses (Kahneman and Tversky 1979). A reference-dependent decision model can be applied to any situation involving uncertainty (Locke et al. 1981; Heath, P.Larrick, and Wu 1999; Pope and Simonsohn 2011; Abeler et al. 2011; Allen et al. 2014; Anderson and A. Green 2018; Gordon, Althoff, and Leskovec 2019). For example, chess players treat their personal best scores in the context of chess ratings, which represent their chess skill, as their reference points (Anderson and A. Green 2018); they try hard to exceed their best scores and attenuate participation after their achieving the increase. Reference points can take the form of objective goals such as round numbers; there are many marathon runners with race times of a little less than the round number of four hours (Pope and Simonsohn 2011). However, the difference in the effects of reference points among individuals with different past experiences is not well understood. Fortunately, large-scale data of user-level decisions in uncertain situations are now becoming available. This enables us to analyze decision making in the context of user history.

This work. The hypothesis underlying our research is that the user's past experience alters the reference point effect. This paper shows a case study of individual differences in the reference point effect by utilizing over 1.2 million decisions made by over 180 thousand users on Codeforces [1] which is a website that hosts programming contests. Codeforces users are assigned ratings representing their programming skills, and have to make decisions in uncertain (probabilistic) situations: Users decide whether or not to join the next contest. If they don't, they can keep their current rating. If they join, they can increase their rating depending on the contest results. Of course, they also have to consider the possibility of having their current rating lowered. The site assigns a color to each user based on her rating; the colors serve as the reference points for many users since they publicly indicate each user's skill most succinctly.

Our finding gained from the case study of Codeforces users is that user's effort and habituation explain her tendency toward loss aversion around a reference point (color boundary). The harder she has struggled to reach and cross the reference point, the more she values what she gets. As a result, she becomes risk averse (refrains from participating) for a while to avoid the chance of losing what she has gained. Past experiences sometimes weaken the motivation induced by the reference points. When a user crosses a reference point for the first time, the cognitive bias from the reference point is significant. However, repeatedly crossing the reference point weakens the impact of the reference point.

Investigating the importance of past experiences helps with answering the key open question on reference-dependent decision models: What determines the shape of the psychological value function? The impact of past experience on future decision-making has been noted by some researchers. In the research area of behavioral eco-

[1] https://codeforces.com/

nomics, it has been shown that when there is a financial cost associated with an action or decision, that cost, in a monetary sense, affects subsequent actions and decisions. Typical examples are the sunk cost bias (the fallacy of buried cost/revenue) (Thaler and Johnson 1990; Kahneman, Knetsch, and Thaler 1990; Arkes et al. 1994; Thaler 1999). We further explore this research topic by considering a different sense of costs such as the effort (in a non-monetary sense), and newly study the impact of sunk costs on the shape of the value function present in the Prospect Theory. Our experiments support the possibility that the effort that the user puts into the process of reaching a reference point determines the shape of the value function.

Our research contributions include insights about which of multiple potential reference points we assign highest priority to. In the context of Codeforces ratings, various metrics such as round numbers and user's personal best rating could be reference points. Furthermore, the Codeforces site tags each user with a color (title) as a step/staircase function indicative of skill. We analyze a large-scale dataset of online competitive programming contests, and show that users behave as if they consider the color boundaries to be reference points. Our study suggests the importance of site design for personal pages since it may influence what numbers users perceive as reference points.

We demonstrate the value of our insights through a task of predicting the decision making characteristics of a user who has just crossed a color boundary (predicting when a user will return to the site and join the next contest) from the reference points and personal history. By manually selecting important features for making predictions based on our insights, we can build a model that achieves better prediction accuracy than a baseline model (time-series deep learning model) even if only a small amount of training data is available.

Our insights appear valid for services with a wide range of decision making events with uncertainty, and they are particular useful for site operators whose goal is to educate users and improve their skills (e.g., Coursera [2] and Duolingo [3]). Such sites need to encourage users and keep them from leaving the service. For example, based on the predicted return time of each user, the site operator might be able to determine whether or not to intervene (engage) with her before she actually quits the site. Also, since her personal history indicates which reference points she is (should be) focusing on, the site operator may improve motivation by explicitly highlighting the preferred point to her.

Preliminary

Competitive programming contests. Codeforces provides a website where participants compete using their programming skills. The site holds programming contests on a regular basis, and users (contestants) are free to decide whether or not to join the contest. Users are asked to solve presented questions within a time limit. The site evaluates the programming skills of users based on the contest results, and updates their ratings. The site uses a step/staircase function to assign a color and title to each user to indicate her rating (see Table 1). For example, a user with a rating between 1900 and 2099 points has the color of violet. Users are rated by a system similar to the popular Elo rating used by official chess federations (Elo 1978). Since Elo rating was originally designed for games with two participants, Codeforces extended it to support games (contests) with multiple participants. We start with a set of users $\boldsymbol{U}$ joining a contest. According to the Elo rating system, the probability that user $i \in \boldsymbol{U}$ with rating R_i gets the better of user $\{j \in \boldsymbol{U} | j \neq i\}$ with rating R_j is calculated by

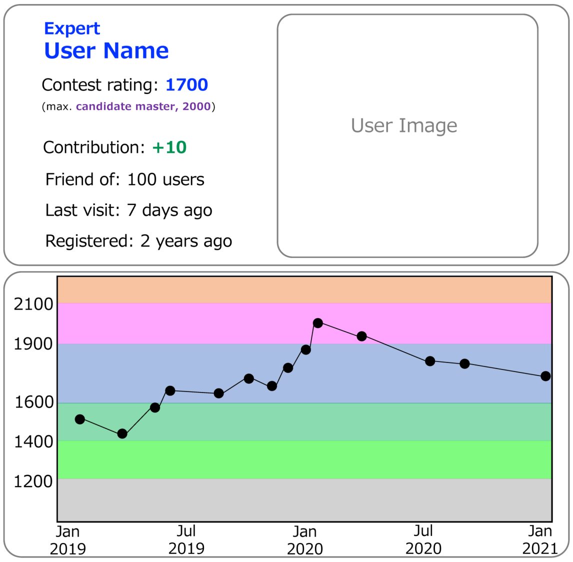
$$P_{ij} = \frac{1}{1 + 10^{\frac{R_j - R_i}{400}}}. \tag{1}$$

Expected rank r_i of i among $\boldsymbol{U}$ is calculated by the sum, over all other users, of the probabilities of winning (securing a higher position than) i;

$$r_i = \sum_{\{j \in \boldsymbol{U} | j \neq i\}} P_{ji} + 1, \tag{2}$$

Rating	Title	Color
3000+	Legendary grandmaster	Black+red
2600 - 2999	International Grandmaster	Red
2400 - 2599	Grandmaster	Red
2300 - 2399	International master	Orange
2100 - 2299	Master	Orange
1900 - 2099	Candidate Master	Violet
1600 - 1899	Expert	Blue
1400 - 1599	Specialist	Cyan
1200 - 1399	Pupil	Green
0 - 1199	Newbie	Gray

Table 1: Rating, title, and color table of Codeforces.



Expert
User Name
Contest rating: 1700
(max. candidate master, 2000)
Contribution: +10
Friend of: 100 users
Last visit: 7 days ago
Registered: 2 years ago

Figure 1: Basic structure of user's personal page.

[2] https://www.coursera.org/

[3] https://www.duolingo.com/

Dataset Statistics	Codeforces
Observation period	23 months June 1, 2018 - April 17, 2020
# unique users	184,161 (146,945)
# total contests	259
# total participating	1,225,240 (1,049,266)
Avg. # contests per user	11.6 times (18.2 times)
Avg. participation time interval per user	19.7 days (16.9 days)

Table 2: Basic dataset statistics. Numbers in brackets are statistics after removing players who participated in fewer than 10 contests over the observation period.

where the last term is a constant that ensures the rank value is greater than or equal to 1. After the contest, ratings of contestants are updated based on the differences between expected and actual rankings. The general idea is to increase the rating if actual rank is better than expected rank. Each user has the initial rating of 1500 upon registering with the site, and the rating is updated according to the results of the contests that the user participates in. For further details of the rating system of Codeforces, refer to its website [4].

User's personal pages. Figure 1 shows the basic layout of each user's personal page. In the upper half of the page, demographic information such as user's name, current rating, and her personal best rating is displayed. Codeforces offers social networking features to help users interact with each other (registering as friends, posting blogs, commenting on blogs, etc.). The method for calculating "contributions" of each user is not made clear on the site, but it appears that it is based on her actions on the social networks (blog posts, comments, etc.). The lower half shows her rating history from the first contest entry to the present.

Basic statistics. The datasets used in this paper were collected by downloading contest metadata from Codeforces using the site's public API [5]. Basic dataset statistics are shown in Table 2. The dataset includes 1.2 million actions (participations) by over 180 thousand users within the 23 month observation period. Note that we collected and used a set of rating histories (a set of timestamp of participation and rating after each participation for each user). Anyone can access the open information, making it easy to reproduce our datasets.

Prospect Theory. According to research into decision making under risk, differences in potential outcomes of decisions are perceived to be biased when they cross a reference point that creates two regions: psychological loss and gain (Kahneman and Tversky 1979). Contrary to the Expected Utility Theory (Von Neumann and Morgenstern 2007), which assumes decision making by perfectly rational agents, Prospect Theory aims to describe the actual behavior of people. The decision model is one of the main areas of interest in the research field of behavioral economics.

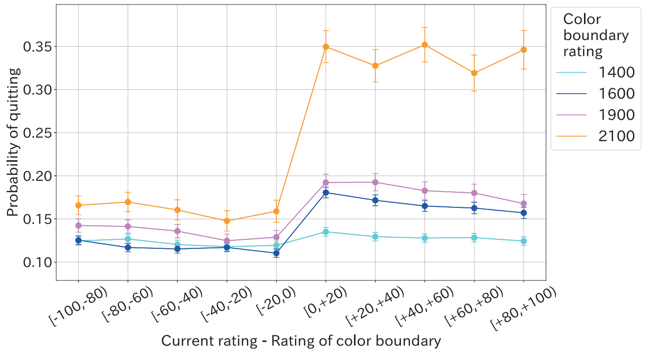


Figure 2: Probability of quitting for at least 20 days around a color boundary, with 95% confidence intervals.

According to Prospect Theory, people behave as if they were following utility function V when making a decision with N choices; they choose the alternative with highest value. V is calculated by

$$V = \sum_{i=1}^{N} \pi(p_i) v(x_i), \quad (3)$$

where x_i is the potential outcome of decision i and p_i is its probability. Function π is a probability weighting function that mirrors the overreaction of people to small probability events and their under-reaction to large probability events. Function v, called the *value function*, assigns a value to each outcome. The theory assumes that the value function passes through reference point r, and is S-shaped and asymmetrical. Value function $v(x)$ of outcome x of a decision is assumed to be a concave function (concave downwards) in the region greater than reference point r ($x - r > 0$), and a convex function (convex function downwards) in the region lower than reference point r ($x - r < 0$): for $x - r > 0$, $v''(x - r) < 0$, and for $x - r < 0$, $v''(x - r) > 0$. In addition, the value function is steeper for losses ($x - r < 0$) than gains ($x - r > 0$) indicating that losses outweigh gains. The value function, which is bilaterally asymmetric, well mirrors actual phenomena; individuals are risk averse in the region slightly above r ($x - r > 0$), but risk accepting in the region slightly under r ($x - r < 0$).

Behavioral Comparison of Multiple Potential Reference Points

Codeforces users have to make decisions in an uncertain (probabilistic) situation as assumed by Prospect Theory. The decision to participate in the next contest is made by the user herself. If the user enters the contest, the rating representing her skill could change. If user doesn't, she can keep her current rating. Earlier literature suggested that various rating contextual factors such as round numbers (Pope and Simonsohn 2011; Allen et al. 2014) and users' personal best (Anderson and A. Green 2018) can be reference points in such situations. In some cases, such as the Codeforces case, a value standard (e.g., see Table 1) other than a numerical rating is presented to users. When there are multiple

[4] https://codeforces.com/blog/entry/20762

[5] https://codeforces.com/apiHelp

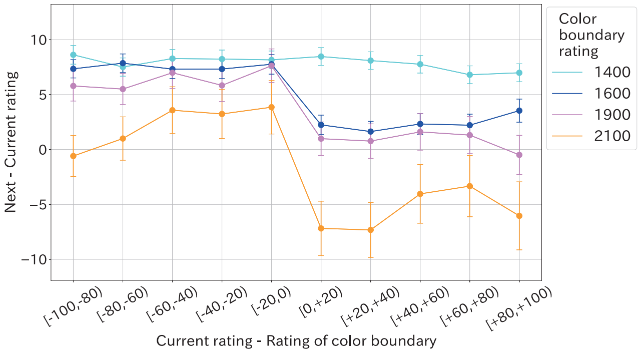


Figure 3: Performance improvement around a color boundary, with 95% confidence intervals.

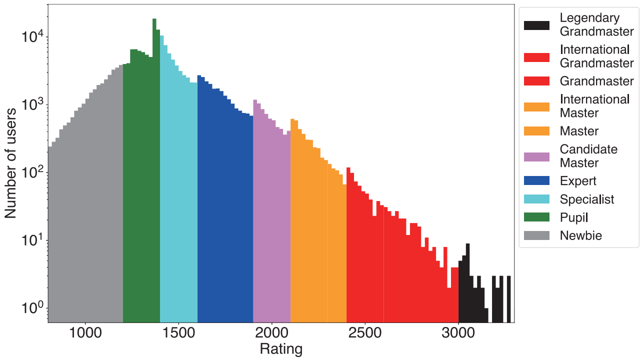


Figure 4: Rating distribution as of April 17, 2020. Only users who participated in contests within the observation period are considered in this plot.

metrics that can be reference points, to what extent is each metric considered as a reference point? This section details our case study on Codeforces data.

Quitting and performance around color boundary. What happens when the user's situation is around a color boundary? Figure 2 shows how the probability of quitting varies with the distance between the Codeforces user's current rating and the color boundary. We define quitting as not joining any new contest within 20 days of finishing the most recent contest (the avg. participation time interval per user is 19.7 days). We drop all contests before the user's 10th contest in order to remove non-serious decisions (i.e., we ignore her decisions until her rating matches her current programming skill). As shown, the probability of quitting decreases gradually when approaching a color boundary and jumps when the color boundary is crossed; users whose rating is slightly above the rating boundary have a higher probability of quitting, while users with ratings slightly below the boundary rating have a lower rate of quitting. The probability of quitting jumps 13-120% when the boundary is crossed. This sudden surge in probability strengthens as the boundary rating increases; for the boundary rating of 2100, the probability of quitting increases from 0.16 (at [-20,0]) to 0.35 (at [0,+20])).

The value function of Prospect Theory provides an interpretation of the observations. According to Prospect Theory, the value function passes through the reference point, and is S-shaped and asymmetrical; the degree of psychological impact in losing (the rating) is greater than that in gaining. This causes people to be "loss averse" which refers to peoples' tendency to prefer avoiding losses to acquiring an equivalent gain. As a result, people avoid putting themselves at risk after securing a better color, i.e., not entering the next contest. The jump-up effect induced by the rating boundary increase with each higher boundary. One hypothesis explaining this phenomenon is that users perceive a higher value for higher ratings, which may create a larger bias as a result (note that this hypotheses is not directly tested in this paper). Figure 4 shows a snapshot of the rating distribution as of April 17, 2020. The phenomenon of user concentration above the color boundary is observed in the region above the initial point (1500). As described in the Preliminary section, the concept of color/title, shown in Table 1, is not considered when updating each user's rating. Therefore, the phenomenon of user concentration cannot be explained in terms of the rating mechanism. We speculate that this is because people who want to avoid losing their current colors hold off on participating in the next contest.

When a color (rating) boundary appears to be reachable, do users try harder and thus achieve better performance as a result? Previous work suggests that users improve their performance when a reference point is within reach (Anderson and A. Green 2018). Do we observe the same phenomenon around color boundaries in the context of competitive programming? Figure 3 shows the difference between next and current ratings as a function of user's current rating. Cyan, blue, violet, and orange lines represent average performance improvement at ratings of around 1400, 1600, 1900, and 2100, respectively. As shown, the performance improvement increases as the color boundary is approached, but the moment the user crosses a color-boundary (thereby securing a more valued color), the performance improvement drastically decreases. This effect strengthens as the absolute rating level increases; the higher the hurdle, the greater is the drop in the effort expended. Note that the reference point of 1400 does not yield any significant difference in performance effort. The reason for this remains unclear, but one hypothesis is that the user's initial rating is set to 1500 points making it difficult for 1400 points to be seen as a reference point.

Adjusting the reference points not only reduces the withdrawal rate, but also strengthens the performance improvement. Our hypothesis is as follows: Those who are currently slightly below the reference point estimate the value of maintaining their status less than they should, and estimate the value of being slightly above the reference point more than they should. As a result, they may be encouraged to participate, strive for better performance, and achieve the higher rating. In other words, the reference point can act as an incentive. It is considered that this is why site operators provide users with reference points.

Why do users slightly above the color boundary perform relatively poorly? Since people valued achieving the higher rating (keeping their new color), they are expected to try hard to keep their new rating. High quitting rates as shown in

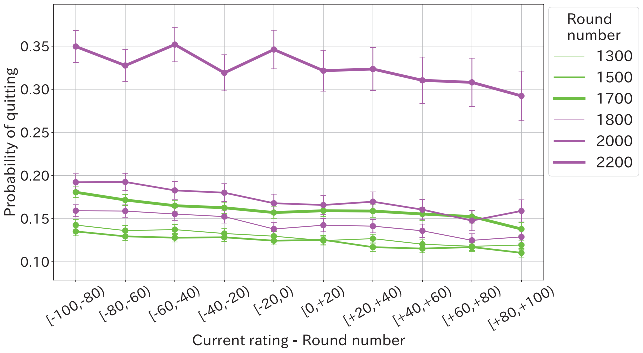


Figure 5: Probability of quitting (not joining the next contest within 20 days of finishing her most recent contest) for round number boundaries (except for color boundaries), with 95% confidence intervals.

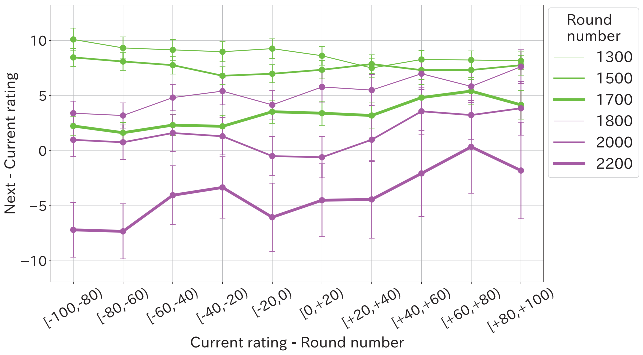


Figure 6: Performance improvement around round number boundaries (except for color boundaries), with 95% confidence intervals.

Figure 2 provides an explanation of the observed data trends. Once users attain the goal, they don't participate in any contests for a while. This quiescent period may lead to poor performance. The results suggest a need to deal with this negative aspect of using reference points (color boundaries); the necessity of encouraging people who have just crossed a color boundary to enter new contests as soon as possible, in order to keep their skills fresh.

Round number and personal best in the context of ratings. Are round numbers effective reference points? We measured the quitting rates and performance of users whose current ratings were near round numbers (scores that were multiples-of-100). Figure 5 plots average quitting rates for the user's current rating, and Figure 6 shows the difference between next and current ratings as a function of user's current rating. For both measures, being compared with the results of the color-boundary, there was little difference before and after the round number boundaries. We also measured the quitting rate and performance improvements of users who set new personal bests; the results do not show a clear change in activity around personal bests. The results of these analyses are listed in Appendix.

Discussions. Our analysis suggests that the colors strongly and clearly serve as reference points and completely suppress the effects of other potential reference points; the numbers on the original ratings such as round numbers and personal best ratings are likely to be ignored. The design of the site tends to reinforce the user behavior of acting in a color-conscious manner. For example, in Codeforces, usernames are highlighted in their current colors on all pages. When visiting a user's personal page, color (title), current rating, and personal best rating are displayed in that order (see Figure 1). Personal best rating is displayed in a smaller font size than others. The color borders are explicitly indicated by auxiliary lines in the graph of rating history. In the screen of user-search results, only the information of user color is shown. These contextual factors boost the users' dependence on color changes. Note that the concept of color as shown in Table 1 is not taken into account in the process of evaluating/calculating user skills/ratings based on the Elo rating system. Given this fact, we speculate that the color-conscious behavior of many users' is influenced by the site's design. Our study implies the importance of site design as regards personal pages since it may influence what metrics people perceive as reference points.

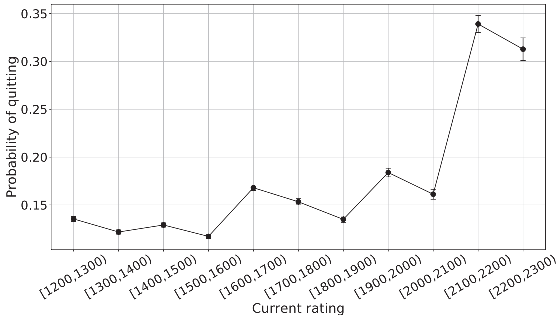


Figure 7: Quitting probability versus current rating.

Making color the most prominent feature on the personal page is successful in temporarily increasing participation-frequency and performance. However, it also causes a drop in participation-frequency and performance immediately after obtaining a better color. To address this issue, it might be a good idea to make other evaluation metrics more appealing than they are now, such as round numbers and personal bests related to ratings and user rankings. This would provide more incentives (different kinds of incentives) for users to participate in their next contests. In particular, it might be a good way to make people more aware of their personal bests than they are now regarding colors and ratings (e.g., highlighting them on personal pages and assigning colors based on them). This is because a more incremental metric diverts the user's attention away from the risk of losing her newly acquired color, and encourages her to look toward incremental performance improvement. Another suggestion is to hold the most recent color regardless of the results of the next contest. These changes are merely speculation, but we believe they are worth examining in a future study.

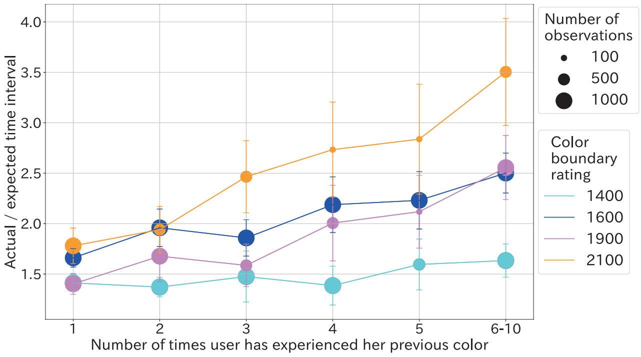


Figure 8: Loss aversion tendency after crossing color boundary given different amounts of effort (with 95% confidence intervals). X-axis plots C, the number of times user has experienced her previous color. Y-axis plots the averaged ratio of the observed to the expected time interval between contest participation.

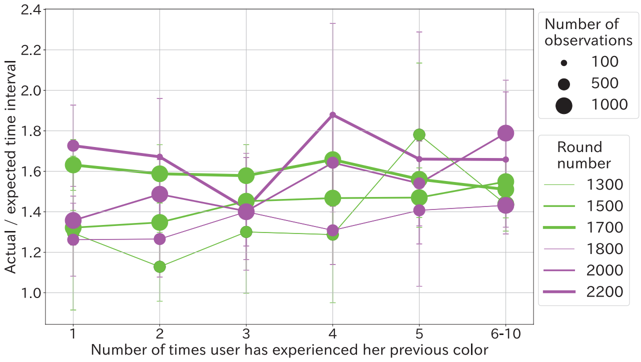


Figure 9: Loss aversion tendency after crossing round number for different amounts of effort (with 95% confidence intervals). X-axis plots C, the number of times user has experienced her previous color. Y-axis plots the averaged ratio of the observed to the expected time interval between contest participation.

Influences of Past Experiences

The hypothesis underlying our analyses in this section is that the user's past experience alters the impact of the reference point. We use "effort" and "habituation" to characterize past experiences. The previous section showed that color boundaries are effective reference points for many Codeforces users, so we consider their behaviors around these boundaries.

Effort. Is the decision of the user around reference points (color boundaries) influenced by the difficulty of past experiences? The hypothesis is that the harder she had to struggle to achieve the boundary, the more she values what she gets. This increase in strength of loss aversion; the user takes longer than usual to enter the next contest. We focus on users who successfully crossed the color boundary in their last contest, and study their subsequent behaviors. We define effort, C, incurred in obtaining the current color as the number of consecutive times the user has experienced their previous (worse) color in a row just before participating in the last contest. The definition is written in Table 3, and our thinking behind the definition is that the most recent (direct) effort it took to obtain the current better color would be straightforwardly reflected in C. For example, suppose a user enters a contest and her color changes from Cyan to Blue (better) as a result. In the next (her last) contest, her color changes from Blue to Violet (better). The effort C incurred in obtaining Violet for the user is small, and is assigned the value of one ($C = 1$). Note that we only count the latest number of times user has experienced her previous color in a row since the amount of recent effort is considered to have a significant impact on her decision making. For example, if the color history is "Cyan → Blue → Violet → Blue → Blue → Violet (current color)", the effort C incurred in obtaining the current color (Violet) is 2. To evaluate the propensity to avoid losses, we use the ratio of the observed to expected time interval; the average of the last five time-intervals is used as the expected value for each user [6]. We observe that the quitting propensity substantially varies with the current rating as shown in Figure 7, and users with high ratings are more likely to quit. In order to reduce the effect of individual differences in time intervals, we measured the time interval to participate relative to its expectation.

Symbol	Description
C	Number of times she has experienced her previous color in a row just before obtaining the current (better) color in her most recent contest. $C > 0$ quantifies effort of user who successfully crossed the color boundary in her last contest.
D	Number of times she has experienced the color change (from worse to better) that occurred with her last contest. $D > 0$ quantifies the degree of habituation to the change.

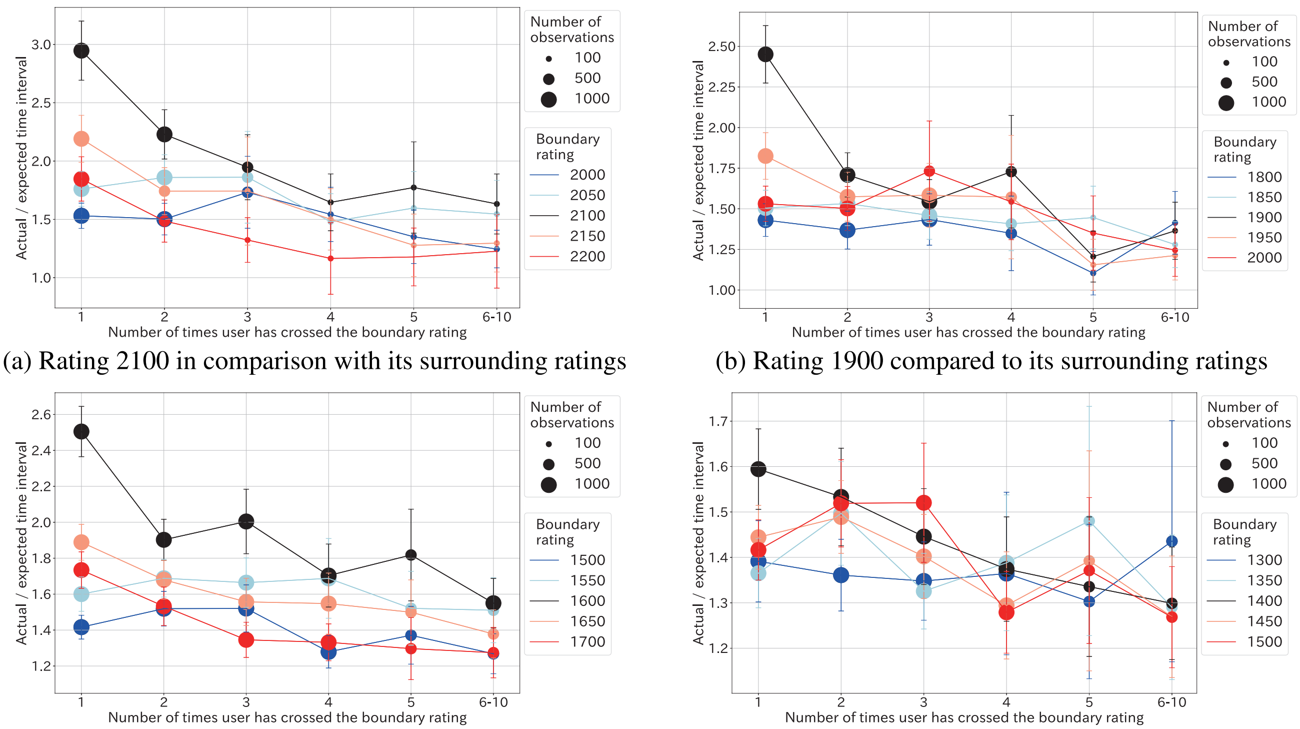
Table 3: Definitions.

Rating	Regression coefficient (slope)	p-value
1400	0.042	**
1600	0.158	***
1900	0.217	***
2100	0.335	***

Table 4: Results of linear regression analysis for color boundaries. The dependent variable is the loss aversion tendency (actual / expected time interval between contest participation), and the independent variable is the effort C (p-value: *** ...$p < 0.001$, ** ...$p < 0.01$).

Figure 8 shows the averaged ratio of the observed to the expected time interval between contest participation as a function of C. More intuitively, the metric shows how many times longer the user stays in the current color than usual.

[6] We also used the average of the last ten time-intervals as the expected value for each user, and confirmed that the results do not depend on which parameter is selected.



(a) Rating 2100 in comparison with its surrounding ratings

(b) Rating 1900 compared to its surrounding ratings

(c) Rating 1600 compared to its surrounding ratings

(d) Rating 1400 compared to its surrounding ratings

Figure 10: loss aversion tendency for different levels of habituation (with 95% confidence intervals). X-axis plots D, the number of times user crossed the color boundary (from worse to better color). Y-axis plots the averaged ratio of the observed to the expected time interval between contest participation.

Rating	Regression coefficient (slope)	p-value
1400	-0.066	***
1600	-0.196	***
1900	-0.257	***
2100	-0.329	***

Table 5: Results of linear regression analysis for color boundaries. The dependent variable is the loss aversion tendency (actual / expected time interval between contest participation), and the independent variable is D, the number of times user has crossed the color boundary from worse to better color (p-value: *** ...$p < 0.001$).

As shown, the time-interval to participate increases as C increases. We performed statistical tests to see if there was a significant difference in the mean of quitting propensity between users with $C = 1$ and users with $C = [6, 10]$. For each color boundary rating, a significant difference was confirmed (two-sided t-test, $p < .05$). This shows that past struggles to secure the latest color strengthen the tendency towards loss aversion. We also performed simple linear regression analysis in order to assess the relationship between the loss aversion tendency (dependent variable) and the effort C (independent variable). The results (regression coefficients and p-values) are shown in Table 4. As shown, for each color boundary, effort C has a statistically significant effect on the loss aversion tendency (regression coefficient > 0 and $p < .01$). Note that we also confirmed the statistically significant effect when the quitting rate (probability of quitting for at least 20 days) is used as an indicator for evaluating the loss aversion propensity (regression coefficient $>$ 0 and $p < .05$); The results are listed in the Appendix.

For comparison, we also determined the average ratio of the observed values to the expected values of users who crossed round number boundaries (multiples-of-100 except for color boundary numbers). The results are shown in Figure 9. As expected, ratings of 1300, 1700, 1800, and 2200 exhibit no significant differences in scores between users with small efforts ($C = 1$) and users with large efforts ($C =$ 6-10) (two-sided t-test, $p > .05$). This is because most users do not consider these ratings as reference points as described in the previous section. However, interestingly enough, we did confirm significant increases in scores for ratings of 1500 and 2000 (two-sided t-test, $p < .05$). Moreover, the linear regression analysis confirmed that effort C has a statistically significant effect on the loss aversion tendency for ratings of 1500 and 2000 (regression coefficient > 0 and $p < .01$). These results suggest that round numbers, which are not usually taken to be effective reference points, may actually be effective for some users depending on their experiences. This interesting anomaly shows the importance of considering history in identifying personal reference points.

One finds greater value in the rewards depending on the costs (efforts) incurred in achieving the rewards. The more worth she ascribes to the reward, the steeper is the curve of the value function in Prospect Theory. The relative impact of prospective losses increases, which leads to increased loss aversion. This phenomenon is also related to the *sunk cost* biases studied in behavioral economics (Thaler and Johnson 1990; Kahneman, Knetsch, and Thaler 1990; Arkes et al. 1994; Thaler 1999). A sunk cost is defined as a cost (in a monetary sense) that has already been incurred and thus cannot be recovered. The sunk cost effect is the tendency to incorporate costs incurred in the past into one's plans for the future even when these past costs are no longer relevant to optimal planning. Also, our observations can be explained by the effect called *effort justification*, which has been studied in social psychology. Effort justification is a kind of cognitive bias and stems from Leon Festinger's theory of cognitive dissonance (Festinger 1957). Effort justification is a person's tendency to attribute a value to an outcome that needed significant effort to achieve, that exceeds the objective value of the outcome. The effect of effort justification is thought to contribute to the sunk cost effect.

Habituation. Does the effect created by the reference point last? Does the value user places on acquiring a certain color stay the same no matter how many times user experiences it? In common with the previous experiments of effort analysis, we studied the behaviors of users who received a better color, and assessed whether the number of times they experienced the color-change influenced their future decisions. We evaluated the propensity to avoid losses by measuring observed time intervals relative to expected intervals; the average of the last five observed time intervals is used as the expected interval.

Figures 10 (a), (b), (c), and (d) show averaged ratio of observed time interval to expected time interval upon exceeding the rating values of 2100, 1900, 1600, and 1400, respectively. X-axis plots D, the number of times the user crossed the color boundary (from worse to better). D is designed for quantifying the degree of habituation to the color change, and its definition is listed in Table 3. The value at $D = 1$, for example, represents the characteristics of users who have acquired the color for the first time. Users with $D > 1$ have experienced a loss of the (better) color in the past. As shown, those who experienced color-change improvement for the first time, waited a long time to participate in the next contest. A user exceeding the color boundaries of 2100, 1900, 1600 and 1400, took 2.9, 2.5, 2.5, and 1.6 times longer than usual to enter the next contest, respectively. The decrease in the quitting propensity (observed time intervals relative to expectations) as D increases suggests that the effect of the reference point fades. We confirmed the hypothesis using linear regression analysis. As shown in Table 5, for each color boundary, D is significantly associated with the loss aversion propensity (regression coefficient < 0 and $p < .01$). We also compared the behaviors around color-boundaries to those around the surrounding round number boundaries (color boundary rating ± 50, ± 100). Note that the previous experiments show that the round numbers are relatively ineffective as reference points for many Codeforces users. We performed statistical tests to see if there was a clear difference in the mean of quitting propensity between two user groups; users who gained the color for the D-th time as a result of their performance in recent contests, and users who crossed the round number boundary for the D-th time as a result of a recent increase in rating. The results of the statistical tests are as follows; we confirmed significant differences in the scores of quitting between 1600 and each of its surrounding round numbers when D is between 1 and 3 (two-sided t-test, $p < .05$). For the color boundary given by rating 2100, significant differences in scores are confirmed when D is 1 and 2. With regards the color boundaries given by ratings of 1400 and 1900, we see a significant difference only when D is 1. By comparison, the value the user feels from acquiring a certain color decreases once she achieves it, and the reference point effect is valid only a few times (up to 3 times).

Past experiences, how many times she has acquired the color, affects the tendency for loss aversion created by the color boundary. When the user crosses the color-boundary and gets the (better) color for the first time, the tendency for loss aversion (i.e., increase in quitting rate) is high. One hypothesis, which has not been verified, is that she may see great value in retaining the new color. However, after the user has been through the color-change a few times, her obsession with the color may fade. The phenomenon, decreased loss aversion effect with the number of times experienced, can be explained by the psychological effect of "habituation" (Sokolov 1963; Groves and Thompson 1970). Repeated experience of a stimulus will cause it to lose its novelty and thus strength. The more one encounters something, the less likely one is to react to it. A similar trend was observed in our analyses; the stimulus effect of the reference point fades due to habituation. Our observation that users securing a better color for the first time (D is 1) have higher quitting rates, can be explained by the loss aversion effect of beating her personal best. However, the effect is maintained only for a few times (up to 3 times), which appears to be due to habituation.

Summary of insights. The harder the user struggles to reach and cross a reference point, the more she values what she gets. As a result, this increases her loss aversion immediately after achieving the goal. Also, when she crosses a reference point for the first time, its impact on cognitive bias (giving excessive importance to achieving the reference point) is significant. However, repeatedly crossing a reference point weakens the reference point effect.

Discussions. What does the Codeforces case suggest about the design of sites intend to help users improve their skills? As indicated, expending significant effort to achieve rating thresholds may increase loss aversion tendencies, which need to be addressed if too many users leave the site. Displaying the rating history in the prominent place on the personal page, as shown in the lower part of Figure 1, may trigger the recall of her own past efforts and may increase her loss aversion tendency. A solution might be to customize their UI so that their rating histories are hidden or obscured.

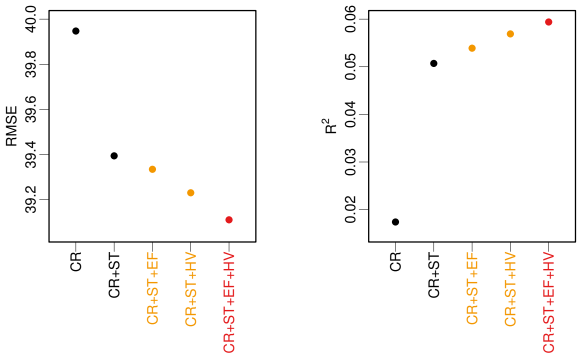


Figure 11: Comparing different model components in terms of RMSE or R^2 when predicting next participation (the time variable has units of days). Comparing our model with insights (orange and red) to baselines (black).

Predicting User Return

Next, we utilize the insights of the previous sections to predict the time to participate in the next contest at the level of each individual using a standard machine learning technique. The task is to predict when a user who won a better color as a result of the most recent contest will enter the next contest. Codeforces provides a place/site for users to improve their programming skills. In order for users to interact with each other actively and compete with many others through contests, it is necessary to encourage user participation. The prediction model can be used for finding potential users who are likely to quit (fail to join any contest for some itme), and this may lead to a decision about whether to intervene (e.g., sending emails to users who are predicted to leave the site in order to encourage participation in the contest). Note that there are a number of studies on techniques for predicting user "lifespan" in online service/site (Kapoor et al. 2014; Dror et al. 2012; Neslin et al. 2006; Yang et al. 2010; Ribeiro 2014; Althoff and Leskovec 2015; Lin, Althoff, and Leskovec 2018).

Features used for learning. We define a series of models that use different sets of features based on the factors explored in the previous sections. We focus on four features:

Current rating (CR) We simply consider the current rating after the most recent contest since the tendency for quitting varies in accordance with it (see Figure 7).

Basic statistics of user history (ST) Statistics on participation time intervals of each user. We use the average of the last (recent) 5 time-intervals between contests attended.

Effort (EF) Based on the analysis reported in the previous section, we equate user effort with the number of times the user has experienced her previous color before her most recent contest.

Habituation (HV) Insights gained in the previous section demonstrate that it's important to know how many times user has experienced the acquisition of the current color.

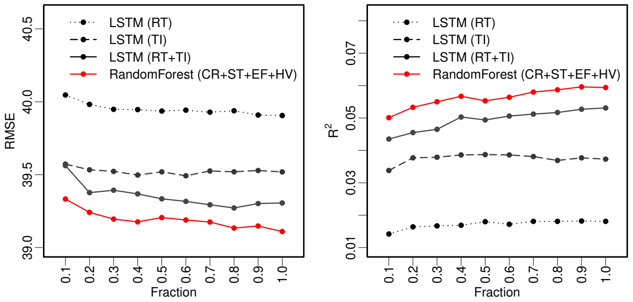


Figure 12: Root mean squared error (RMSE) and Coefficient of determination (R^2) when predicting time of next participation (the unit of time variable is days). Comparing our model (red) to LSTM models with different features (black) while changing the size of training data.

Experimental setup. We report the performance achieved with the Random Forest (Random Forest Regressor of scikit-learn[7]). We use root mean square error (RMSE) and coefficient of determination (R^2) as evaluation metrics. We collected cases where predictions were feasible from the data shown in Table 2; the number of data (cases) totaled 68,397. We randomly split the data into 70% (47,879 data) as training data, 10% (6,839 data) as validation data, and 20% (13,679 data) as test data. We choose the trained model with best parameter values (e.g., the number of trees in the forest and the maximum depth of the tree) based on the results of prediction performance on validation data.

Summary of results. Our results are shown in Figure 11. They demonstrate that capturing all properties (CR, ST, EF and HV) is essential to predicting the time of next participation. The model that considers all our insights, the red plot, outperforms baselines (black plots) by 0.28-0.84 in terms of RMSE and 0.009-0.042 in terms of R^2.

Comparison with deep learning. Recent developments in the field of deep learning enable us to automatically learn the relationship between past history and the behavior around the reference points. Considering the recent success of deep learning in various research fields, applying the technique to the user return prediction task is, at least, one of the reasonable approaches. However, a lot of data must be prepared for training and the cost of parameter tuning is high. By manually selecting important features for making predictions based on our insights, we can build a model with good prediction performance even if only a small amount of training data is available.

To illustrate this fact, we evaluate the prediction accuracy on test data when varying the amount of training data for both Random Forest with our insights (CR+ST+EF+HV) and a modern time-series deep-learning model. The basic flow of the experiment is the same as in the previous experiment. However, in this experiment, sub-sets of the training data are randomly formed using $X\%$ (from 10 to 100% at

[7] https://scikit-learn.org/stable/index.html

10% intervals) of the original training data. Thus, as X increases, the accuracy (error) of predictions is expected to increase (decrease).

To assess prediction performance, we compare our prediction model with LSTM, which is a state-of-the-art time-series deep learning model (Hochreiter and Schmidhuber 1997). LSTM has a feedforward neural network structure wherein outputs from the hidden units at the prior time step are used as the inputs for the current time step. We learn a series of LSTM models that use two different features as follows.

- Sequence of current and past ratings (RT)
- Sequence of past time intervals (TI)

The best parameter values are determined using validation data. MSE is used as the training objective; the Adam optimizer (Kingma and Ba 2015) is adopted to compute and update all the training parameters. The number of LSTM layers, LSTM units, epochs are set to $\{1, 2\}$, $\{32, 64\}$, and $\{300, 500, 1000\}$ respectively; best parameter values are chosen from them based on the results of prediction performance on validation data. The framework we used is Chainer.5.4.0 [8]. Experimental results are shown in Figure 12. X-axis plots the fraction of data used for training and Y-axis plots the prediction error. As shown, our model (red plot) always outperforms LSTM-based models (black plots) in all conditions (for all fractions of training data). For both evaluation metrics, our model offers noticeably lower prediction error than the baselines when the fraction of training data is relatively small (between 0.1 and 0.4). The accuracy of LSTM may improve if more data is available. However, the advantage of our methodology, which explicitly specifies important features for making predictions, is that it allows us to build predictors even if a small amount of data is available. Also, there is no need to choose the best parameters from among the myriad of candidates.

We showed one aspect of the value of our insights as we achieve the better predictors than the deep learning models. On the other hand, the value of R^2 (RMSE) itself is not so high (small), and further research and development for the prediction task is needed to raise accuracy to a practical level.

Related Work

Prospect Theory is one of the most significant theories in recent psychological research into behavioral economics and decision making. The value function and probability weighting function have attracted the attention of many researchers as they form the basis of Prospect Theory. Kahneman et al. show the basic properties of these functions (Kahneman and Tversky 1979), and their subsequent work proposes specific functions that satisfy the properties needed (Tversky and Kahneman 1992; Lattimore, Baker, and Witte 1992; Prelec 1998; Gonzalez and Wu 1999; Rieger and Wang 2006). Although our work does not discuss the specific shape and formula of the functions, we newly analyze how past experiences enhance and weaken the psychological biases around reference points of the value function (i.e., the impact of past experiences on the value function has not been studied in those prior studies).

Another line of research is to study what people consider to be reference points in various situations. Prospect Theory, was initially developed from the existence of counter-examples to the Expected Utility Theory (Von Neumann and Morgenstern 2007) in economics, which assumes agents are rational, thus the main subject of research has been decisions that involve money. However, (Tversky and Kahneman 1992) generalize this proposal and show that it can be applied to any situation involving uncertainty. Typical examples of uncertain reference points are round numbers. Pope et al. (Pope and Simonsohn 2011) found that professional baseball players, as the end of the season approaches, act as if they want to end with a batting average slightly higher than .300. Allen et al. (Allen et al. 2014) analyzed a massive dataset of finishing times of marathon runners, and found that round numbers such as 4 hours often serve as reference points. Some reference points come from internal judgements. Heath et al. (Heath, P.Larrick, and Wu 1999) point out that goals at an abstraction level can serve as reference points; the goal alters the value of outcomes as suggested by the psychological principles underlying Prospect Theory's value function. Gordon et al. (Gordon, Althoff, and Leskovec 2019) analyzed large-scale action datasets from activity tracking applications to determine the relationship between user goal-setting and resulting behavior. In addition, Abeler et al. (Abeler et al. 2011) showed that expectations can serve as reference points; Anderson et al. (Anderson and A. Green 2018) studied reference points in the context of chess ratings by using data of online chess games, and found that players act as if their personal best ratings are reference points. Although these studies have shown that a variety of metrics can be treated as reference points, they did not discuss the superiority of these potential reference points. More importantly, those studies did not study the impact of past experiences on peoples' decision making. Our study, by contrast, finds that peoples' efforts and habituation (components of personal history) help to explain individual differences in the effects of reference points.

The badge system is similar to the rating system of Codeforces in that badges (equivalent to color or title) are used for motivating users (Anderson et al. 2013; Kusmierczyk and Gomez-Rodriguez 2018). For example, (Anderson et al. 2013) studied how badges influenced user behavior on a site with a badge system based on the behavior data of the question-answering site. Badges are given to users for particular contributions to the site, such as performing a certain number of actions (e.g., answering questions). However, they assume the situation wherein the users face no risk to their ratings. Our study examines decision making wherein their ratings are at risk, and show insights into user behavior on sites that provide user with a color (like a badge) as a step/staircase rating function.

The impact of past experience on future decision-making has been noted by some researchers. In the research area of behavioral economics, it has been shown that when there is a financial cost associated with an action or decision, that cost

[8] https://docs.chainer.org/en/stable/

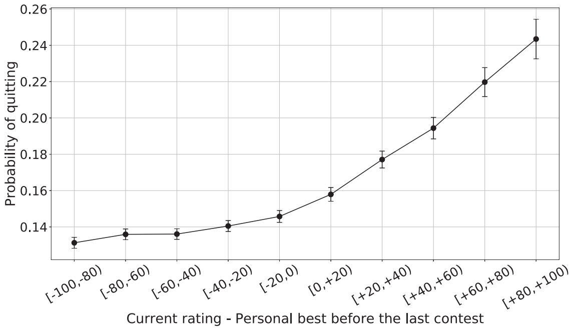


Figure 13: Probability of quitting for at least 20 days around personal best ratings, with 95% confidence intervals.

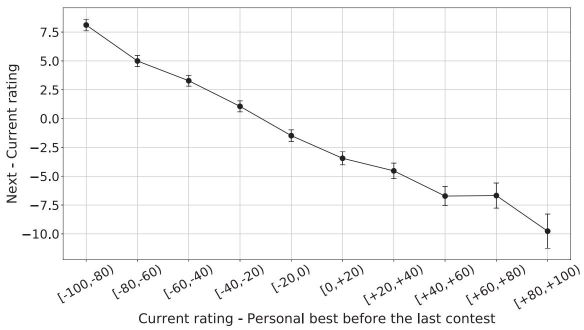


Figure 14: Performance improvement around personal best ratings, with 95% confidence intervals.

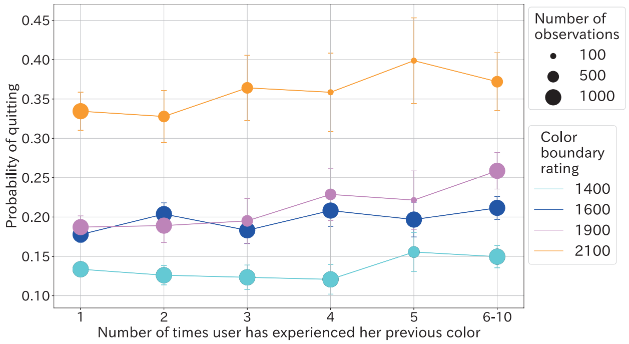


Figure 15: Loss aversion tendency after crossing color boundary given different amounts of effort (with 95% confidence intervals). X-axis plots C. Y-axis plots the probability of quitting (not joining the next contest within 20 days of finishing the most recent contest).

affects subsequent actions and decisions (i.e., the sunk cost bias (Thaler and Johnson 1990; Kahneman, Knetsch, and Thaler 1990; Arkes et al. 1994; Thaler 1999)). However, the extensive research into sunk costs has focused primarily on cost in a monetary sense. We further explore this research topic by considering a broader range of costs such as effort, and newly study the impact of sunk costs on the shape of the value function present in the Prospect Theory.

Conclusion

This paper showed that past experiences affect decisions around reference points. Examining 1.2 million actions of 180 thousand users, we found that the harder a user struggles to reach and cross a reference point, the more she is likely to quit once she has achieved it. Repeatedly crossing the reference point weakens the reference point effect. Based on our insights, we can build a model that makes better predictions about when users who have just crossed the color boundary will next participate than the deep-learning based model, particularly if only a small amount of training data is available.

Our work is a case study of one of the world's most popular competitive programming sites; the case of Codeforces shows that history has a clear impact on user's decision making around reference points. We believe that the effects found will also be demonstrated in similar cases where participants compete using their skills, and our future work is to explore the extent to which our foundational idea is valid.

Appendix

Personal best in the context of ratings. We measured the quitting rate and performance improvements of users who set new personal bests. Figure 13 shows how the probability of quitting varies with the distance between the user's current rating and her personal best rating from her last contest. As shown, quitting probability increases as ratings increase, but no surge in quitting is found for the color boundaries. Peaks in quitting around the personal best have been reported in analyses of chess ratings (Anderson and A. Green 2018), but this phenomenon was not clearly observed in our data. The upward trend shows that people who break their own records tend to be more likely to quit, however, we should also consider the fact that users with higher ratings are likely to leave the site completely (see Figure 7). Figure 14 shows how a user's performance changes with the distance between hers current rating and her personal best rating from her last contest. The downward trend in this figure is mainly due to the Elo rating system where the higher user's own rating, the harder it is to increase her rating. Similar to the results for the quitting rate, the result does not show a clear change in activity around personal bests.

Effort analysis based on quitting probability. Figure 15 shows the probability of quitting (not joining the next contest within 20 days of finishing the most recent contest) as a function of effort C. We also performed simple linear regression analysis, and confirmed the statistically significant effects for all color boundaries (regression coefficient > 0 and $p < .05$). The dependent variable is whether she quits the site or not, and the independent variable is the effort C. The fact shows that past struggles strengthen the tendency towards loss aversion.

Ethics Statement

The findings of this study are intended to be used to assist people in continuously improving their skills. The only concern is that it might not always be good to encourage users of a site to be more active; excessive its use could be problematic. This should be kept in mind when using the results of this study.

The data used in this study is publicly available with Codeforces users' permission. Because we only used the public data, we did not recruit any human subjects for this research. Anyone can access the open information, making it easy to reproduce our datasets.

References

Abeler, J.; Falk, A.; Goette, L.; and Huffman, D. 2011. Reference points and effort provision. *American Economic Review*, 101(2): 470–492.

Allen, E. J.; M. Dechow, P.; G. Pope, D.; and Wu, G. 2014. Reference-dependent preferences: Evidence from marathon runners. *Management Science*, 63(6): 1657–1672.

Althoff, T.; and Leskovec, J. 2015. Donor retention in online crowdfunding communities: A case study of donorschoose.org. In *Proceedings of the World Wide Web Conference*, 34–44.

Anderson, A.; and A. Green, E. 2018. Personal bests as reference points. *Proceedings of the National Academy of Sciences of the United States of America*, 115(8): 1772–1776.

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2013. Steering user behavior with badges. In *Proceedings of the World Wide Web Conference*, 95–106.

Arkes, H. R.; Joyner, C. A.; Pezzo, M. V.; Nash, J. G.; Siegel-Jacobs, K.; and Stone, E. 1994. The psychology of windfall gains. *Organizational Behavior and Human Decision Processes*, 59(3): 331–347.

Dror, G.; Pelleg, D.; Rokhlenko, O.; and Szpektor, I. 2012. Churn prediction in new users of yahoo! answers. In *Proceedings of the World Wide Web Conference*, 829–834.

Elo, A. E. 1978. *The rating of chess players, past and present*. Arco.

Festinger, L. 1957. *A theory of cognitive dissonance*. Stanford University Press.

Gonzalez, R.; and Wu, G. 1999. On the shape of the probability weighting function. *Cognitive Psychology*, 38(1): 129–166.

Gordon, M.; Althoff, T.; and Leskovec, J. 2019. Goal-setting and achievement in activity tracking apps: A case study of MyFitnessPal. In *Proceedings of the World Wide Web Conference*, 571–582.

Groves, P. M.; and Thompson, R. F. 1970. Habituation: A dual-process theory. *Psychological Review*, 77(5): 419–450.

Heath, C.; P.Larrick, R.; and Wu, G. 1999. Goals as reference points. *Cognitive psychology*, 38(1): 79–109.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*, 9(8): 1735–1780.

Kahneman, D.; Knetsch, J. L.; and Thaler, R. H. 1990. Experimental tests of the endowment effect and the Coase theorem. *Journal of political Economy*, 98(6): 1325–1348.

Kahneman, D.; and Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2): 263–292.

Kapoor, K.; Sun, M.; Srivastava, J.; and Ye, T. 2014. A hazard based approach to user return time prediction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1719–1728. ACM.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*.

Kusmierczyk, T.; and Gomez-Rodriguez, M. 2018. On the causal effect of badges. In *Proceedings of the World Wide Web Conference*, 659–668.

Lattimore, P. K.; Baker, J. R.; and Witte, A. 1992. The influence of probability on risky choice: a parametric examination. *Journal of Economic Behavior and Organization*, 17: 377–400.

Lin, Z.; Althoff, T.; and Leskovec, J. 2018. I'll be back: on the multiple lives of users of a mobile activity tracking application. In *Proceedings of the World Wide Web Conference*, 1501–1511.

Locke, E. A.; Shaw, K. N.; Saari, L. M.; and Latham, G. P. 1981. Goal setting and task performance: 1969–1980. *Psychological bulletin*, 90(1): 125.

Neslin, S. A.; Gupta, S.; Kamakura, W.; Lu, J.; and Mason, C. H. 2006. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of marketing research*, 43(2): 204–211.

Pope, D.; and Simonsohn, U. 2011. Round numbers as goals: Evidence from baseball, SAT Takers, and the Lab. *Psychological Science*, 22(1): 71–79.

Prelec, D. 1998. The probability weighting function. *Econometrica*, 66(3): 497–527.

Ribeiro, B. 2014. Modeling and predicting the growth and death of membership-based websites. In *Proceedings of the 23rd international conference on World Wide Web*, 653–664.

Rieger, M. O.; and Wang, M. 2006. Cumulative prospect theory and the St. Petersburg paradox. *Economic Theory*, 28: 665–679.

Sokolov, E. N. 1963. Higher nervous functions: The orienting reflex. *Annual Review of Physiology*, 25(1): 545–580.

Thaler, R. H. 1999. Mental accounting matters. *Journal of Behavioral decision making*, 12(3): 183–206.

Thaler, R. H.; and Johnson, E. J. 1990. Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management Science*, 36(6): 643–660.

Tversky, A.; and Kahneman, D. 1992. Advances in prospect theory: Cumulative representations of uncertainty. *Journal of Risk and Uncertainty*, 5: 297–323.

Von Neumann, J.; and Morgenstern, O. 2007. *Theory of games and economic behavior*. Princeton University Press.

Yang, J.; Wei, X.; Ackerman, M.; and Adamic, L. 2010. Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1): 186–193.