Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

On Quitting: Performance and Practice in Online Game Play

Tushar Agarwal
Indian Institute of Technology Ropar
Punjab, India 140001
tushar.agarwal@iitrpr.ac.in

Keith Burghardt
University of California at Davis
Davis, CA 95616
kaburghardt@ucdavis.edu

Kristina Lerman
USC Information Sciences Institute
Marina del Rey, CA 90292
lerman@isi.edu

Abstract

We study the relationship between performance and practice by analyzing the activity of many players of a casual online game. We find significant heterogeneity in the improvement of player performance, given by score, and address this by dividing players into similar skill levels and segmenting each player's activity into sessions, i.e., sequence of game rounds without an extended break. After disaggregating data, we find that performance improves with practice across all skill levels. More interestingly, players are more likely to end their session after an especially large improvement, leading to a peak score in their very last game of a session. In addition, success is strongly correlated with a lower quitting rate when the score drops, and only weakly correlated with skill, in line with psychological findings about the value of persistence and "grit": successful players are those who persist in their practice despite lower scores. Finally, we train an ϵ-machine, a type of hidden Markov model, and find a plausible mechanism of game play that can predict player performance and quitting the game. Our work raises the possibility of real-time assessment and behavior prediction that can be used to optimize human performance.

Introduction

How much grit do you think you've got?
Can you quit a thing that you like a lot?
– "On Quitting" by Edgar Guest

How do people achieve mastery? What distinguishes high achievers from average performers? Performance generally improves with practice, as demonstrated on a variety of tasks in the laboratory setting and in the field (Newell and Rosenbloom 1981), suggesting that with enough practice even mediocre performers can approach the mastery of successful individuals. However, not all practice is equally effective in helping achieve mastery. Deliberate practice, which emphasizes quality, not quantity of practice, improves performance most (Duckworth et al. 2011; Ericsson and others 2006). The search for individual traits responsible for variations in the capacity for deliberate practice uncovered *grit*, a trait related to psychological constructs, such as persistence, resilience, and self-control, which enables individuals to persevere in their efforts to achieve their goals (Duckworth et al. 2007). Grit may explain the self-discipline to continue practicing, even when faced with temporary setbacks, such as a short-term drop in performance.

Recent proliferation of behavioral data collected "in the wild" enables longitudinal studies to explore and validate these findings. We carry out an empirical analysis of online game play to quantify individual traits associated with success. The data we study consists of records of over 850K players of a game called Axon. Following (Stafford and Dewar 2014), who first studied this data, we operationalize performance as player's score, and practice as playing rounds of the game. Like other behavioral data, Axon data presents analytic challenges. It is extremely *noisy*: requiring aggregating variables over the population. It is also *heterogeneous*: composed of differently-behaving subgroups varying in size according to the Pareto distribution. As a result, the trends observed in aggregated data may be quite different from those of the underlying subgroups (Vaupel and Yashin 1985). To address this effect, known as Simpson's paradox, we disaggregate data by user skill and activity. After disaggregating data, we can more accurately measure the relationship between performance and practice. While performance generally improves with practice, we find that players tend to quit after an abnormally high score, suggesting significant rewards in casual games may instead encourage players to leave. Interestingly, we find that players who are less likely to quit after a score drop tend to become more successful later. Quitting is not as strongly correlated with skill, suggesting that it is perseverance to poor outcomes, i.e., grit, that contributes to player success.

To identify a plausible mechanism of game play, we train an ϵ-machine, a type of a hidden Markov model, on the data, and find models that maximizes the accuracy of predicting players' performance. We find that players are most predictable when we model how their behavior is affected by changes in score from their previous game, instead of, for example, the change from their mean score. This model leads to insights not just in how players leave the game but the dynamics of performance as well.

Methods

The Axon game (http://axon.wellcomeapps.com) is a casual single player game where the player controls the growth of an axon. Performance is characterized by a score which

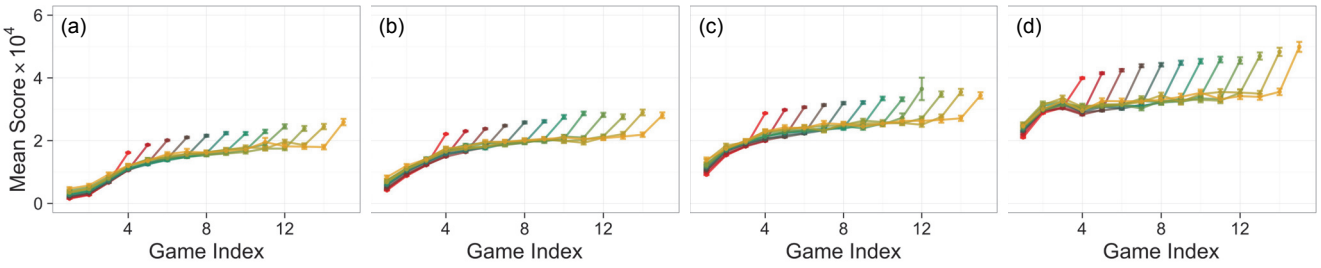

Figure 1: Score versus game play index for the (a) bottom, (b) second, (c) third, and (d) top quartiles by talent. Lines represent sessions of different length, from 4 games to 15 games played in the session. Error bars represent standard error.

represents the length of the axon, with stochasticity introduced by "power-ups" which can boost the score. The data (https://github.com/tomstafford/axongame) contains records of over 3M games played by more than 854K players. Each record has score and time of the game (hourly resolution), and a "machine identifier", an anonymized identifier derived from the web browser from where the game was accessed. Following Stafford & Dewar (Stafford and Dewar 2014), we assume that each machine identifier corresponds to a unique player. The code used for our study is available at https://github.com/agarwalt/AxonGame.

The vast majority of people played only a few games: 92% played fewer than eight games, with 28K playing more than 12 games. People who play few games may be systematically different from dedicated players who play many games; consequently, aggregating games across both groups can lead to Simpson's paradox. To address this challenge, we segment each player's activity into sessions, where a session is a sequence of games without a long break (two hours or longer) between consecutive games.

We also segment players by skill to partly control for variability between the best and worst players. We distinguish between two types of skill: (1) *talent*, or the initial skill of a player, which is operationalized as the median score of the first three games, and (2) *success*, which is measured by the median of the three highest remaining scores. For players with fewer than seven games, we take the median of all remaining scores, and for players with fewer than four games, talent and success are both defined as the median of all scores.

Results

Success and Practice

Figure 1 shows the evolution of performance (average score) over the course of a session among players of similar skill (grouped by talent). Lines represent sessions of different length, from 4 games to 15 games played in the session. There were 242K such sessions (out of the total 990K), representing 1.4M games, or approximately half of the total 3M games. The figures reveal interesting trends. First, performance generally increases with the number of games played, reflecting the benefits of practice. Second, eventual performance depends on skill: the most talented players (top quartile) have a better score, on average, on their very first game of a session than the least talented players (bottom quartile) have after practice. While the plot reflects performance averaged over all player sessions, these differences, also noted by Stafford & Dewar, remain strong when only the player's first session is considered (data not shown). Finally, the very last game of a session has an abnormally high score, on average. Aside from this last game, performance curves for sessions of different lengths within the same population overlap, suggesting that we properly captured the underlying behavior.

To rule out Simpson's paradox, we repeat the analysis on randomized data, where the indices of the games within each session are shuffled. In the randomized data (not shown) performance no longer depends on the order the games within the session are played. This suggests that we properly disaggregated data.

The high score in the very last game in the session partly explains the performance boost that Stafford & Dewar (Stafford and Dewar 2014) attributed to practice spacing. They found that players who split their first ten games over a period longer than a day had higher scores on average than those who played the games within the same 24-hour time period. However, those who spaced their games over a long period likely played the games over multiple sessions, while those who played them on the same day are more likely to have played just one session. Therefore, the higher average performance of the former group may be skewed by the high score of the last game of the session.

Quitting

Why does the last game of a session have a much higher score (on average)? Do players simply choose to stop playing, thus ending the session, after receiving an abnormally high score? To investigate this hypothesis, we empirically measure the probability to stop playing given the person played n games. We assume that this decision is based on a player's performance relative to his or her previous games. Though there is a variety of ways to measure relative performance, we choose to measure it as score difference from the previous game.

Figure 2 shows the quitting probability versus score difference from the previous game, Δ, for different populations of players when split by talent. The quitting rate is simply the number of users who quit at score difference Δ, divided by the number who ever reach Δ over a given range of game

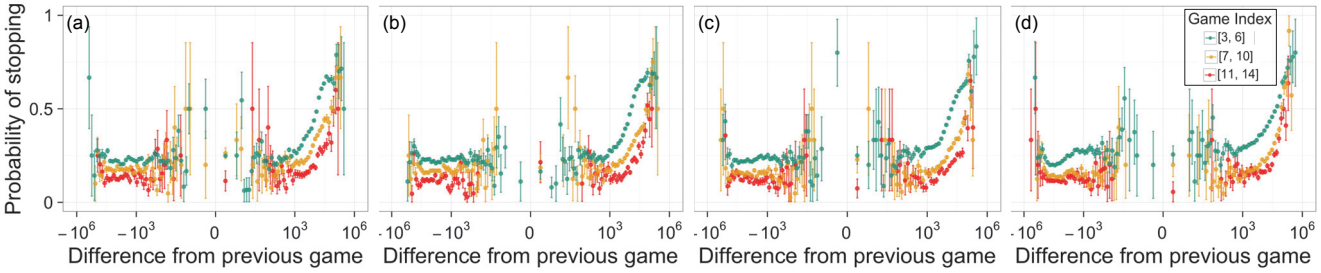

Figure 2: Probability to end a session (quit) versus score difference from the previous game for (a) the bottom, (b) second, (c) third, and (d) top quartiles of talent. Colors indicate over which indices the probability was calculated: $3-6$ (green), $7-10$ (yellow), or $11-14$ (red). Error bars represent standard error. We notice that the plots approximately overlap which suggests that the rate of quitting is nearly a stationary process. However, at early indices, the quitting rate is higher than at later ones.

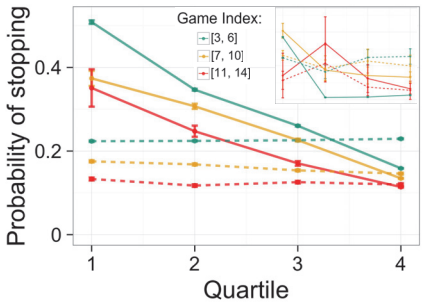

Figure 3: The probability to quit when scoring less than in the previous game. The lines represents quartiles of players split by success (solid line) and talent (dashed line). Error bars show standard error. The first quartile represents the least talented (successful) players and the 4th quartile the most talented (successful) players. Inset: the probability to quit when scoring more than the previous game.

indices. For 10K$< \Delta <$ 15K, players are more likely to stop playing (Figure 1), even though large Δ does not correlate directly with any single game feature, such as power-ups. However, a concerted use of power-ups in succession can result in an increase of more than 10K points. Surprisingly, for $\Delta < 0$, the quitting rate is not strongly dependent on Δ (Figure 2). Should the designers of such games, then, avoid adding game elements which "satisfy" a player and potentially cause them to lose motivation to play? Answering this requires controlled experiments and is beyond the scope of this study.

Success and Persistence

Why do some people quit while others continue to play even when doing poorly (i.e., obtaining a worse score)? These *persistent* players may possess a trait psychologists call *grit*, which has been linked to high achievement and success (Duckworth et al. 2007). To investigate the impact of persistence on performance, we first need to quantify persistence, which we operationalize as the probability to stop playing after underperforming, i.e., obtaining a score less than the previous game's score. Figure 3 shows the average persistence—or probability to quit playing after getting a worse score—for different quartiles of players as split by success or talent. Interestingly, we see a relationship between performance and persistence only in subpopulations of players segmented by success: the more successful players (those who achieve higher best scores) are less likely to stop playing after a setback, i.e., receiving a worse score. In contrast, the relationship does not appear to be very strong when players are split by talent (their initial skill), or when we measure the effect of score increases rather than decreases (inset in Figure 3). Thus, consistent with psychology research on grit, persistence is associated with high performance and success, and not talent. Furthermore, successful players do not simply play longer; rather their ability to persevere despite lower scores distinguishes them from the less successful players.

Modeling Performance

We model game play activity, including quitting and performance, using an ϵ-machine, a type of Hidden Markov Model that is optimally predictive (it produces the least uncertainly about the future behavior (Shalizi and Crutchfield 2001)) and minimally complex (it requires the fewest number of effective states (Crutchfield 2012)). We fit an ϵ-machine to our data using the Causal State Splitting Reconstruction algorithm (Shalizi and Klinkner 2004), which groups past behaviors together into a single effective state if they make similar predictions. The outcome is the simplest, most parsimonious model with the highest predictive power.

We created models with four states: one state for quitting the game, and the remaining states for "poor", "good", and "very good" performance, defined as the score difference, Δ from either the player's previous score. Thus, a player can have "poor" ($\Delta < 0$), "good" ($0 \leq \Delta < \Theta$), and "very good" performance ($\Theta \leq \Delta$), for some threshold Θ. We further tested whether this was the best type of model, by comparing it against and median, or mean of the previous games. We also explored how the past game behavior (e.g., score changes between successive games) affect model prediction accuracy, but slight improvement in accuracy did not justify the dramatic increase in the number of states.

To evaluate the model's prediction accuracy, we first used 90% of the data for training and reserved 10% for testing, in order to maximize the amount of training data of the data-intensive ϵ-machine (Shalizi and Klinkner 2004). We then created receiver operating characteristic (ROC) curves by testing whether the model correctly predicted state X or not state X, found the corresponding area under the curve (AUC) for every state X. Finally, we took the mean AUC weighted by the frequency of each symbol X, as in (Provost and Domingos 2000). We bootstraped the testing data to determine the confidence intervals of the AUC values.

After we trained the models separately on each quartile of players, split by talent, we found that using score difference from the previous game, where thresholds $\Theta = 300,\ 8K,\ 16K$, or $22K$ for each respective quartile, optimizes the ϵ-machine's prediction accuracy. The average AUC of the model was approximately 0.64. This suggests that player behavior may depend on how their scores change from their last play, a type of peak-end effect, rather than from their typical play.

By training the model, we also learned transition probabilities between states. The model shows that not only does the quitting probability increase with score difference, in agreement with Figure 2, but also that players transition in unexpected ways between states before they eventually quit. For example, players who perform poorly ($\Delta < 0$) are very likely to perform well in the next game. Similarly, there is an unexpected probability to transition from a "very good" state in the last game to a "poor" state in the next one, which suggests that players undergo periods of score volatility. Finally, the transition rates from negative to positive states are greater than the opposite transition rates in several quartiles, which suggests that players tend to improve over time.

Conclusion

We empirically investigated factors affecting mastery of an online game using digital traces of activity of many players. The large dataset enabled us to investigate sources of individual variability and their impact on practice and performance.

Skilled (or talented) players, who get high scores already in their first games, are more successful overall. However, continued practice improves the scores of all players. We identified a factor, related to grit, which captures the likelihood the player will keep practicing, i.e., playing the game, even when performing poorly. The more likely the player is to continue playing after a drop in performance, the more successful he or she eventually becomes. However, the ability to persevere and continue practicing is not related to the player's initial skill.

We modeled this behavior using an ϵ-machine and found that the model in which players based their decisions on how well they did compared to their previous game best predicted whether they will continue playing and their performance. Surprisingly, when players did very well compared to their last game, they were highly likely to quit, but when they performed poorly, their quitting probability remained low.

Our analysis relied on identifying and accounting for the sources of heterogeneity in game play data. Unless this is done, analysis can fall prey to Simpson's paradox, in which false trends can be observed when aggregating over heterogeneous populations. Initial skill, or talent, is a major source of behavioral heterogeneity. Players who score well on their first games continue to improve and outperform the poorest players. Another significant source of heterogeneity is the temporal structure of game play: players have periods, or sessions, of continuous activity with breaks in between. After accounting for sessions, a clearer picture of performance emerged.

While empirical analysis of behavioral data cannot replace controlled experiments, the sheer size of the data allows for the study of individual variability that is not possible with smaller laboratory experiments. Such data can be used to explore alternate hypotheses about behavior, which can then be validated in the laboratory setting. Moreover, the types of quantitative methods explored in this paper could be used to predict performance and for psychological and cognitive assessment of individuals from their observed behavior. Future human-computer interfaces could continuously observe and predict users' behavior and adapt so as to optimize their performance.

Acknowledgements. This work was supported in part by NSF (#SMA-1360058), ARO (#W911NF-15-1-0142), and the USC Viterbi-India and ISI summer internship programs.

References

Crutchfield, J. P. 2012. Between order and chaos. *Nature Physics* 8:18–24.

Duckworth, A. L.; Peterson, C.; Matthews, M. D.; and Kelly, D. R. 2007. Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology* 92(6):1087.

Duckworth, A. L.; Kirby, T. A.; Tsukayama, E.; Berstein, H.; and Ericsson, K. A. 2011. Deliberate practice spells success: Why grittier competitors triumph at the national spelling bee. *Social Psychological and Personality Science* 2(2):174–181.

Ericsson, K. A., et al. 2006. The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge handbook of expertise and expert performance* 38:685–705.

Newell, A., and Rosenbloom, P. S. 1981. Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition* 1:1–55.

Provost, F., and Domingos, P. 2000. Well-trained pets: Improving probability estimation trees. *CDER Working Paper*.

Shalizi, C., and Crutchfield, J. P. 2001. Computational mechanics: Pattern and prediction, structure and simplicity. *J Stat Phys* 104(3):817–879.

Shalizi, C. R., and Klinkner, K. L. 2004. Blind construction of optimal nonlinear recursive predictors for discrete sequences. In Chickering, M., and Halpern, J. Y., eds., *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI 2004)*, 504–511. Arlington, Virginia: AUAI Press.

Stafford, T., and Dewar, M. 2014. Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological Science* 25(2):511–518.

Vaupel, J. W., and Yashin, A. I. 1985. Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician* 39(3):176–185.