Identifying Off-Task Users in a Large-Scale, Game-Based Practice Assessment

Matthew Emery
Roblox Inc.
memery@roblox.com

David Laing
Roblox Inc.
dlaing@roblox.com

Philip Simmons
Roblox Inc.
psimmons@roblox.com

Jacob Seybert
Roblox Inc.
jseybert@roblox.com

Katrina Yu
Roblox Inc.
kyu@roblox.com

Erica Snow
Roblox Inc.
esnow@roblox.com

ABSTRACT

Kaiju Cats, a game-based practice assessment, was developed and launched on the Roblox platform to prepare job candidates for a game-based hiring assessment. Since September 2023, Kaiju Cats has been played over 50,000 times; however, many of these playthroughs were completed by non-candidates who were motivated to redeem avatar rewards rather than to prepare for the hiring assessment. To better gain insight into how players may be approaching and interacting with the practice test, we deployed a finite mixture model to distinguish between on-task and off-task playthroughs. This initial analysis provided useful insights and laid out next steps for the continued development and iteration of our game-based practice assessments.

Keywords

game-based assessment, clustering, finite mixture model, mixed effect model

1. INTRODUCTION

1.1 Game-Based Hiring Assessments at Roblox

Roblox is an online game platform and game creation system with a growing employee base. Each year, Roblox receives thousands of applications from new college graduates for entry-level positions in software engineering and product management. In 2021, Roblox developed and operationalized a game-based hiring assessment for these roles, designed to measure cognitive skills such as creative problem-solving and systems thinking.

1.2 Roblox Practice Assessment: Kaiju Cats

Leveraging evidence that practice tests can elevate both organizational and candidate success[3], we launched the Kaiju Cats practice assessment in 2023 to strategically prepare applicants for our hiring process. Kaiju Cats integrates user interface patterns from our hiring assessment, complete with introductory and concluding screens, an intuitive tutorial, and a real-time timer. It familiarizes candidates with the specific UI and gameplay mechanics they will encounter in the hiring assessment and immerses them in a strategic scenario that produces similar cognitive and decision-making challenges.

The game is designed to engage candidates in complex problem-solving tasks that require careful resource management and strategic planning, core competencies that are crucial across diverse roles at Roblox as identified through an exhaustive job analysis. Players must effectively allocate a limited budget and predict the outcomes of various strategies within the constraints of time. This setup illustrates the application of problem-solving skills, as candidates iterate their strategies in response to feedback, mimicking the dynamic and interactive nature of the problem-solving scenarios with which applicants must engage.

In Kaiju Cats, players are challenged to create a plan that directs a trio of cats across a city. The objective of the game is to maximize the cats' total power—the sum of the numeric score associated with each cat—by causing them to destroy buildings and reach their respective destinations, cat beds located at the opposite side of the city. Players start with a fixed amount of dollars, which they can spend to define their plan. Each time players submit a plan for testing, they see the resulting power gains. Players have fifteen minutes to submit a plan that maximizes the cats' power. Within this time limit, players can submit as many plans as they wish.

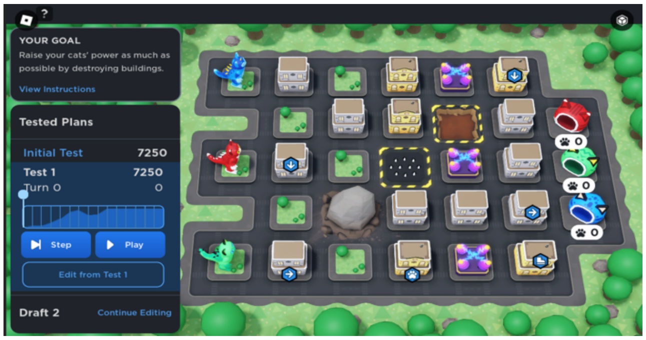
Kaiju Cats was designed and developed in collaboration with game designers and engineers over a period of six months. The development phase included regular "think aloud" sessions with test users, which helped us ensure that the game appealed to a wide audience and effectively facilitated our learning objectives. Moreover, these sessions allowed us to confirm that players engaged in the expected problem-solving processes while playing, mirroring the cognitive challenges of our actual assessments.

1.3 Game-Based Assessment Practice

Incorporating a practice test specifically designed to acclimate candidates to the innovative UI and gameplay mechanics of game-based assessments is essential within a comprehensive practice test framework, as evidenced by the detailed exploration of psychometric properties and assessment valid-

M. Emery, D. Laing, P. Simmons, J. Seybert, K. Yu, E. Snow, and J. Buckley. Identifying off-task users in a large-scale, game-based practice assessment. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 844–848, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729968



YOUR GOAL

Raise your cats' power as much as possible by destroying buildings.

Tested Plans

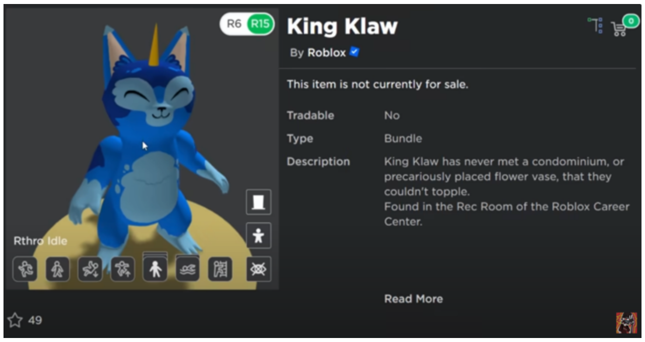
Initial Test 7250

Test 1 7250

Turn 0 0

Draft 2

Figure 1: Screenshot of the Kaiju Cats Board.



King Klaw

By Roblox

This item is not currently for sale.

Tradable No

Type Bundle

Description King Klaw has never met a condominium, or precariously placed flower vase, that they couldn't topple.
Found in the Rec Room of the Roblox Career Center.

Figure 2: Screenshot of the Kaiju Cats Avatar.

ity in game-based learning environments[1]. These practice tests serve a dual purpose: they mitigate the novelty effect—ensuring that performance reflects true ability rather than unfamiliarity with the format—and they enhance comfort with the game's operational aspects, allowing candidates to focus on the assessment's content[4]. By simulating real assessment conditions, these practice tests prepare candidates for the types of cognition and actions required under similar pressures and constraints encountered in the actual assessment.

1.4 Data Collection

Roblox experiences use a Client-Server architecture, where the user's local client sends telemetry to a server, updating the game state accordingly. In the Kaiju Cats experience, telemetry data was collected from all users, without capturing any personally identifiable information. The Roblox ID associated with each user was anonymized using a one-way hash function, ensuring that the telemetry data cannot be deanonymized, yet allowing identification of multiple playthroughs by the same user.

Consent for the use of this data was obtained through the Roblox User Agreement and Privacy Policy, which all players accept before participating in any Roblox experience. The policy explicitly states how data collected during gameplay may be utilized for various purposes, including analytics and research. By agreeing to these terms, users consent to the collection and use of their data in anonymized form for research and development purposes, such as this study aimed at enhancing user experience and game design. This process aligns with Institutional Review Board exemption criteria, emphasizing data collection from subjects who consent to the use of their information and the recording of this information in a manner that prevents the identification of individual subjects, directly or through identifiers linked to them.

1.5 Kaiju Cats Viral Moment

When Kaiju Cats was initially launched, it offered players exclusive avatar items as rewards for playing. The purpose of these rewards was to introduce job candidates to the culture of avatar customization and item collection that is central to the Roblox platform and business. In November 2023, Kaiju Cats experienced a sudden and unexpected influx in users following the release of several YouTube videos that guided users on how to redeem these avatar items. Produced by prominent YouTubers, whose content focuses on helping users acquire free Roblox items, these guides dramatically altered the game's player demographic[10]. This shift, which we are calling the “viral moment”, led to an influx of data from off-task users who engaged with the game primarily to acquire the exclusive items.

1.6 Related Work

Walonski et al.[11] proposed detecting gaming behaviour in intelligent tutoring systems using machine learning systems. Caballero-Hernández et al.[2] suggest using process mining techniques to assess skills in serious games. This technique is similar to the one described here, which also uses process data.

The use of finite mixture model-based clustering in education has been discussed in Scrucca et al.[9]. The use of a finite mixture model applied to longitudinal data to identify non-compliance was proposed in Pauler et al.[7].

1.7 Current Work

With the launch of Kaiju Cats and the viral moment that followed, we wanted to gain insight into how players may be approaching and interacting with the practice test. The work presented here aims to capture and quantify the ways in which players engaged in both on-task and off-task behaviors in the practice test. On-task behavior involves actively engaging with and focusing on the tasks aligned with the learning objectives of an assessment, such as using designated game mechanics to solve problems. Off-task behavior involves engaging in activities unrelated to the assessment's objectives, such as following specific instructions to do minimal actions to reach a reward.

2. METHODS

2.1 Data Collection

For the current work, raw telemetry data was collected through the Roblox platform. This data represented the users' actions, clicks, and decisions made during gameplay. The complete dataset represented 114,878 individual playthroughs from 53,792 unique users across 9 datasets of Kaiju Cats between 2023-08-03 and 2024-01-25.

2.2 Data Processing and Analyses

2.2.1 Data Cleaning

We excluded playthrough data in which the users did not submit at least one plan and instances in which a rare bug allowed users to continue playing past the 15-minute timer. The final dataset consisted of 56,419 playthroughs. We wrote a Python package to extract key measures from the telemetry of each playthrough, including user identifiers, dataset IDs, submission timestamps, and submission scores.

2.2.2 Data Analysis

The volume of playthroughs increased dramatically in November 2023, after the viral moment. Before the viral moment, the average weekly volume of playthroughs was 508. After, it was 4,131. We found that playthroughs in the pre-viral sample submitted on average 9 plans (SD=6.6) with a mean high score of 37,903 (SD= 27,277), while those in the post-viral sample submitted on average 5.6 plans (SD= 4.8) with a mean high score of 19,713 (SD= 12,811). These differences alerted us to the possibility of a change in population characteristics following the release of the YouTube videos.

2.2.3 Clustering

We suspected that the release of the YouTube videos led to an increase in off-task playthroughs that were motivated by acquiring avatar rewards. To identify these playthroughs, we applied a clustering algorithm based on the trajectory of scores and submission times.

2.2.4 Finite Mixture Model

We created a finite mixture model where each subpopulation was defined by a mixed effect model.[1] The fixed effect of the mixed effect model predicts the score of a submission as a function of the time it was submitted. Each playthrough was given a random intercept to account for the intrinsic individual differences between playthroughs. The EM algorithm iteratively updates the model parameters towards their maximum likelihood. This model was fitted using the flexmix R package[5, 8]. The finite mixture model can be described by the following equations[6]:

$$f(\ y_{ij} \mid \Theta) = \sum_{k=1}^{K} \pi_k \cdot f_k(y_{ij} \mid \Theta_k)$$

Where:

π_k is the mixing proportion of the kth cluster

$f_k(y_{ij} \mid \Theta_k)$ is the PDF of the kth mixed effect model component, parameterized by Θ_k

$f(\ y_{ij} \mid \Theta)$ is the PDF of the data. Θ is the set of parameters across all K clusters. y_{ij} describes the mixed effect model for the ith user's jth submission, as follows.

$$y_{ij} = \beta_{0k} + \beta_{1k} \cdot x_{ij} + u_i + \epsilon_{ij}$$

Where:

y_{ij} is the score of the ith user's jth submission

β_{0k} is the intercept for the kth cluster, representing the initial score at time = 0

$\beta_{1k}$is the slope of the kth cluster

x_{ij} is the ith user's jth submission time, in seconds

u_i is the random intercept of user i

ϵ_{ij} is the error term of the ith user's jth submission

2.3 Assumptions

In employing a finite mixture model with mixed effect components for clustering playthrough trajectories, it is imperative to acknowledge the underlying assumptions that influence the analysis and interpretation of results. This approach assumes homogeneity within each identified cluster and presumes independence of observations, which may not be true because a user can engage in the practice test multiple times. The model requires specifying the number of clusters K a priori. Furthermore, it assumes that data within each cluster follows a normal distribution, and that the relationship between predictors and the response variable is linear, with error terms that are independent and identically distributed.

3. RESULTS

We applied the clustering method to the combined pre- and post-viral sample. The "on-task" cluster represents candidates with higher top scores and higher submissions, and the "off-task" cluster represents those with lower top scores and fewer submissions. We found 4,846 on-task playthroughs in the pre-viral sample and 6,843 in the post-viral sample, consistent with the assumption that the viral moment had little effect on the volume of on-task users. By contrast, we found 3,282 off-task playthroughs in the pre-viral sample and 38,597 in the post-viral sample. The clusters differed in the number of playthrough submissions and their top scores. The on-task sample submitted an average of 9.5 plans and the off-task sample submitted an average of 5.2 plans. The average highest score for the on-task cluster was 40,827 (SD=21,056), while for the off-task cluster it was 10,991 (SD=4,398).

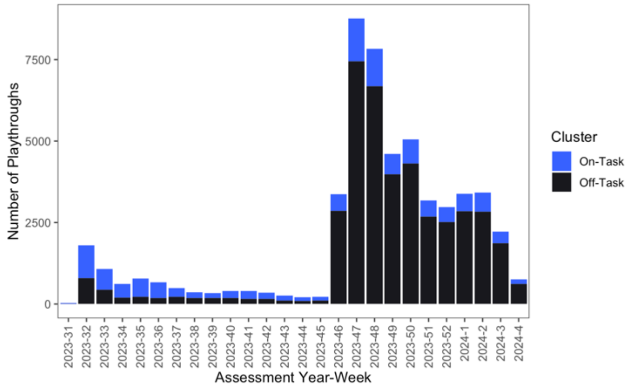


Figure 3: Cluster Assignment by Week.

Prior to the viral moment, on-task playthroughs outnumbered off-task playthroughs, week-over-week. After the viral moment, off-task playthroughs outnumber on-task ones. Although there were eight times as many playthroughs week-over-week in the post viral period, the number of on-task playthroughs only doubled, week-over-week.

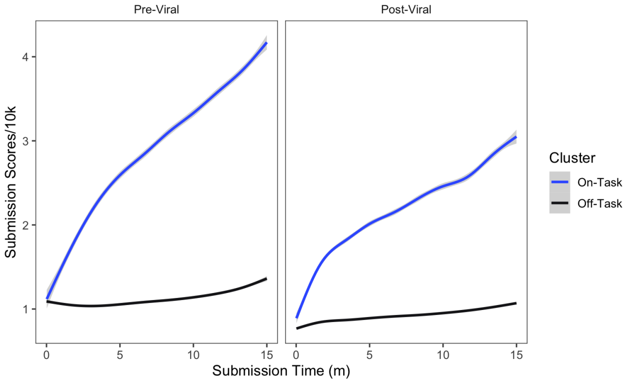


Figure 4: Mean Submission Score by Submission Time.

In both the pre- and post-viral periods, we observed a clear separation in average submission score trajectories across submission times between the two clusters. Users in the on-task cluster show rapid and steady improvement in submissions scores across the fifteen-minute period, while users in the off-task cluster show little to no improvement.

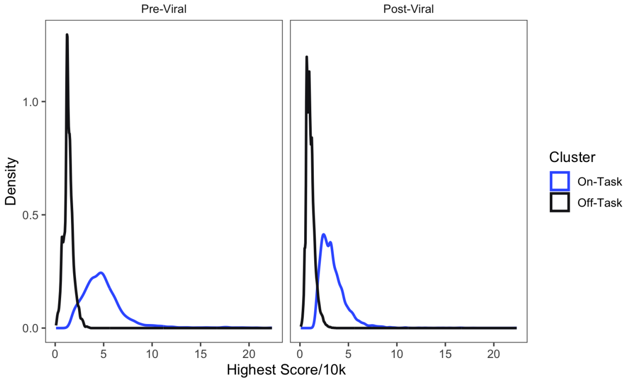


Figure 5: Highest Score Across Clusters in Pre- and Post-Viral Samples.

In both the pre- and post-viral periods, users in the on-task cluster achieve consistently higher final scores than users in the off-task cluster.

Users in the off-task cluster tend to submit their plans early—often submitting their first plan within one minute of beginning gameplay. Generally, users in the on-task cluster, by contrast, do not submit their first plans until 3–5 minutes have passed. This is consistent with the assumption that off-task users spend less time analyzing the problem and are more likely to act quickly, perhaps while directly mimicking strategies suggested in YouTube demonstrations.

4. DISCUSSION

4.1 Avatar Reward Incentives

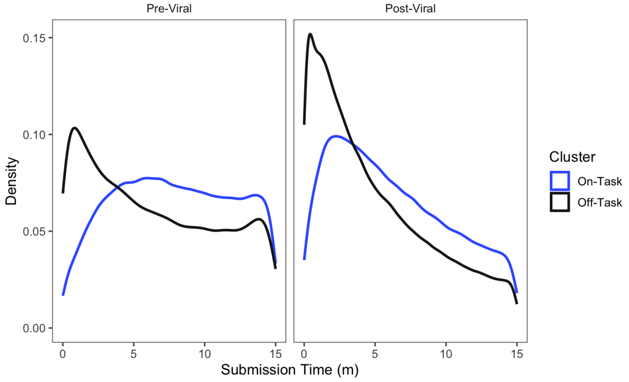


Figure 6: Submission Times Across Clusters in Pre- and Post-Viral Samples.

The purpose of offering avatar rewards was to introduce job candidates to Roblox's culture of avatar customization and item collection. However, since Kaiju Cats is accessible to anyone, the availability of these rewards caused many presumed non-candidate users to play the game. These users were motivated not to prepare for the hiring assessment by engaging earnestly with the problem-solving task presented by Kaiju Cats, but to mimic playthroughs displayed on YouTube for the purpose of reaching the checkpoints needed to acquire avatar rewards.

4.2 Further Research

A limitation of our present analysis is that we have not yet developed benchmarks or alternative methods for identifying off-task users. There are several ways we intend to seek out this information and further develop our analysis. One is to develop a criterion-based method for identifying off-task users. Since the criteria for acquiring the avatar rewards are precisely defined, we can use these criteria to identify which users successfully achieved the avatar rewards, and within that group, which users stopped playing immediately upon achieving the rewards.

Another approach will be to identify the intersection of users who played Kaiju Cats and users who completed our operational hiring assessment. This will allow us to directly examine the relationship between engagement with the practice assessment and performance on the hiring assessment. We are also collecting data in a controlled experiment that examines the causal effect of exposure to the practice assessment on performance in the hiring assessment. In this experiment, we are particularly interested in whether the effect of exposure to the practice test interacts with user attributes such as gender, ethnicity, or video game experience.

5. ACKNOWLEDGEMENTS

We would like to thank the Roblox Assessment Tools team for helping develop the Kaiju Cats assessment: Adrian Ionut Toriste, Brynn Balcom, Caitlin Goodwin, Dustin Haynes, Daniel Kopycienski, Josh Akana, Regis Frey, Rameez Saiyid, Seth Scoville, and Tom Smith. In addition, we would like to thank Yao Xiong for her initial work on metadata extraction.

6. ADDITIONAL AUTHORS

Additional author: Jack Buckley (Roblox Inc., email: jbuckley@roblox.com)

7. REFERENCES

[1] V. A. Brown. An Introduction to Linear Mixed-Effects Modeling in R. Advances in Methods and Practices in Psychological Science, 4(1):1–19, Jan. 2021. Publisher: SAGE Publications Inc.

[2] J. A. Caballero-Hernández, M. Palomo-Duarte, J. M. Dodero, and D. Gaševic. Supporting Skill Assessment in Learning Experiences Based on Serious Games Through Process Mining Techniques. International Journal of Interactive Multimedia and Artificial Intelligence, 8(6):146–159, June 2024. Section: 146.

[3] M. C. Campion, E. D. Campion, and M. A. Campion. Using practice employment tests to improve recruitment and personnel selection outcomes for organizations and job seekers. The Journal of Applied Psychology, 104(9):1089–1102, Sept. 2019.

[4] J. Eustace, M. Bradford, and P. Pathak. A Practice Testing Learning Framework to Enhance Transfer in Mathematics. volume 14, pages 88–95, Oct. 2015.

[5] B. Grün and F. Leisch. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. Journal of Statistical Software, 28(4):1–35, 2008.

[6] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake. Finite Mixture Models. Annual Review of Statistics and Its Application, 6(1):355–378, 2019. _eprint: https://doi.org/10.1146/annurev-statistics-031017-100325.

[7] D. K. Pauler and N. M. Laird. A Mixture Model for Longitudinal Data with Application to Assessment of Noncompliance. Biometrics, 56(2):464–472, June 2000.

[8] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[9] L. Scrucca, M. Saqr, S. López-Pernas, and K. Murphy. An introduction and tutorial to model-based clustering in education via Gaussian mixture modelling, June 2023. arXiv:2306.06219 [stat].

[10] SharkBlox. FREE AVATAR BUNDLES! HOW TO GET Meowza, Catzilla & King Klaw! (ROBLOX Kaiju Cats EVENT), Nov. 2023.

[11] J. Walonoski and N. Heffernan. Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. pages 382–391, June 2006.