Usage Patterns and Performance Gains in Gamified Online Judges: A Data-Driven Analysis Informed by Cognitive Psychology in CS1

Luiz Rodrigues[1 [0000−0003−0343−3701]], Andres Felipe Zambrano[2 [0000−0003−0692−1209]], Maciej Pankiewicz[2,3 [0000−0002−6945−0523]], Amanda Barany[2 [0000−0003−2239−2271]], and Ryan Baker[2 [0000−0002−3051−3232]]

[1] Technological Federal University of Paraná (UTFPR) - Apucarana, Brazil
luizrodrigues@utfpr.edu.br
[2] University of Pennsylvania, Philadelphia, USA
[3] Warsaw University of Life Sciences, Warsaw, Poland

Abstract. This study examines how students interact with a Gamified Online Judge through cognitive psychology strategies, repeated and distributed practice, and how these patterns contribute to performance improvements. Using a data-driven approach, we analyze students' usage patterns while considering their prior knowledge and HEXAD profiles to understand how different patterns relate to performance gains. Feature selection resulted in the choice of Number of Attempts and Grouped Hours Between Attempts for further analysis, as these were significantly related to performance gains. Clustering analysis revealed six student usage patterns with varied practice and spacing behaviors, highlighting the complex interplay between repeated and distributed practice strategies by demonstrating that usage patterns influence performance gains differently depending on students' prior knowledge but not HEXAD user profiles. Thus, this work offers insights into how adaptive learning environments can better support diverse student needs, emphasizing the importance of aligning gamified assessment design with cognitive strategies to personalize learning experiences effectively.

Keywords: Student Modeling · Gamification · Testing Effect · Spacing Effect

1 Introduction

Students from Introductory Programming Courses (CS1) often struggle with logic, syntax, and problem-solving, leading to high dropout and failure rates as well as motivating educators' search for engaging and effective learning approaches [13, 3]. In that context, Online Judges (OJs) are promising solutions, offering immediate assessment and feedback that support deliberate practice [7]. Despite that, these tools' effectiveness depends on students engaging with them, where gamification has been adopted as a means to enhance their motivation to do so [5]. While Gamified OJs (GOJ) can foster student motivation, gamification's effectiveness often varies depending on individual differences [16].

Particularly, the HEXAD framework, which categorizes users into six types (e.g., Achievers, Socializers, and Philanthropists), provides a useful model for understanding how different students respond to gamification [17].

Besides motivation, learning is also shaped by cognitive strategies, such as *Repeated Practice* (RP) and *Distributed Practice* (DP), which reinforce learning through frequent problem-solving and spaced exposure [4, 8]. While GOJs encourage repeated attempts that might be spaced in varied ways, no single strategy is likely to support every student [11]. Notably, it remains unclear how different combinations of RP and DP contribute to student learning in STEM settings like CS1, especially in the context of GOJs [8], as previous studies on CS1 education often explored the benefits of OJs and gamification separately.

Research on OJs emphasizes their role in automated assessment and scalable learning [9, 5] but studies on gamification highlight its motivational potential as well as its inconsistent impact [15, 6]. Additionally, few studies integrate cognitive psychology principles to analyze how students' behaviors in GOJ translate into learning outcomes in light of their individual differences (e.g., [14, 12]), but fail to acknowledge previous knowledge and motivational orientations like HEXAD profiles. That is, there is a knowledge gap in understanding how individual differences relate to students' interactions with GOJ in the context of CS1 [8].

Towards addressing that gap, this paper adopts a data-driven approach to investigate how CS1 students interact with a GOJ in light of cognitive psychology strategies. We examine how usage patterns derived from RP and DP relate to performance gains while considering students' prior knowledge and HEXAD profiles. Accordingly, our Research Question (RQ) asks: **How do different usage patterns in a GOJ contribute to performance improvements among CS1 students with varying levels of prior knowledge and motivational orientations?** By identifying effective usage strategies based on empirical data, our findings offer insights into personalized strategies for guiding CS1 students toward productive behaviors and the design of adaptive learning environments.

2 Method

This study explores how CS1 students' interactions with a GOJ relate to performance gains. For this, we analyzed data from RunCode [10], an online platform for automated code execution and testing, used voluntarily by CS1 students to solve C# programming assignments. Students submitted code via an online editor, with unlimited submissions per task evaluated using unit tests. Platform usage did not impact course grades.

The dataset includes submission records from 586 undergraduate CS1 students (27% female) enrolled at a large European university between 2021 and 2024, who provided consent via a questionnaire at the start of the semester. A pre-test assessed baseline knowledge, and a graded post-test evaluated final performance. Performance gain was calculated as the difference between pre- and post-test scores. At semester's end, students also completed the HEXAD gamification profile questionnaire [17].

First, we extracted ten features from the dataset based on RP and DP. For example, to capture retrieval practice, we included the number of attempts and

the number of tasks solved. To reflect session-based DP, features included the total number of sessions (defined as attempts made within an hour of one another), time between attempts (in hours and grouped by every 30 minutes), and attempts per session. We then conducted a data cleansing process to remove outliers, defined as values beyond three interquartile ranges from the median, in line with best practices for handling extreme values [1]. Next, we performed feature selection, where features showing a strong correlation ($r > 0.5$) with others were excluded. Then, to assess the remaining features' relevance to performance gains, we used Spearman's correlation for numeric features and analysis of variance (ANOVA) for categorical features, applying a 95% confidence threshold.

For usage pattern modeling, we aimed to understand how combinations of RP and DP features relate to performance gains. We standardized the selected features using z-score normalization and applied k-means clustering due to its efficiency and wide adoption in educational data mining [2]. The optimal number of clusters was determined using the elbow method based on within-cluster sum of squares, ensuring meaningful groupings in the dataset. Lastly, to analyze these clusters, we conducted a series of ANOVAs to understand how performance gains, RP and DP strategies, prior knowledge (i.e., pre-test scores), and HEXAD profiles differed across clusters. When significant differences were found, we conducted pairwise comparisons using the Mann-Whitney test, which is robust to non-normal distributions and unequal variances [18], adjusting alpha values due to multiple comparisons using the Benjamini-Hochberg procedure.

3 Results

Number of Attempts, *Hours Between Attempts*, *Attempts per Session* and *Grouped Hours Between Attempts* were selected after feature selection. *Number of attempts* has a small-to-moderate statistically significant correlation with performance gains ($r = 0.339$; $p < 0.001$), as well as the average performance gain differs depending on *Grouped Hours Between Attempts* ($F(10, 575) = 5.093$; $p < 0.001$; $\eta_p^2 = 0.081$). Differently, *Hours Between Attempts* ($r = 0.094$; $p = 0.236$) and *Attempts per Section* ($r = -0.044$; $p = 0.291$) correlation with performance gains are non-significant. Thereby, we chose *Number of Attempts* and *Grouped Hours Between Attempts* to be inputted into the clustering analysis, where the elbow method suggested the optimal number of clusters was six.

Table 1 demonstrates that clusters differ in the *Number of Attempts* and *Grouped Hours Between Attempts*, and pairwise tests (see Table 2) reveal all clusters differ from one another with a single exception: the *Number of Attempts* between C2 and C4. Despite that, the clusters present varied combinations of testing and spacing practices. For instance, C1 is the number one in attempts but the second-to-last in spacing, while the opposite is true for C2.

Additionally, Table 1 demonstrates that clusters differ in terms of performance metrics, where pairwise results are more nuanced (see Table 3). For *performance gains*, three out of the 15 pairwise comparisons did not yield statistically significant differences based on the adjusted alpha values. For *previous knowledge*, four out of the 15 pairwise comparisons led to non-significant results. On the other hand, Table 1 suggests no significant differences among clusters in

Table 1: Cluster comparison for overall usage and performance metrics.

	Clusters						ANOVA			
Stat	**C0**	**C1**	**C2**	**C3**	**C4**	**C5**	**F-stats**	**p-val**	**alpha**	η_p^2
N	44	62	188	98	81	113				
*Number of Attempts**										
Mean	145.205	729.548	112.415	299.449	97.309	444.460	776.144	< **0.001**	0.04	0.870
Std	79.425	127.124	74.239	45.966	89.727	67.291				
*Hours Between Attempts**										
Mean	5.682	3.081	10.862	7.735	1.556	4.584	1662.29	< **0.001**	0.05	0.935
Std	1.272	0.816	0.416	1.001	0.775	1.307				
*Performance Gain**										
Mean	1.447	3.070	1.643	2.460	1.053	2.498	13.147	< **0.001**	0.02	0.102
Std	1.570	1.792	1.861	2.006	1.493	2.130				
*Previous Knowledge**										
Mean	4.353	1.973	4.220	3.336	4.487	3.059	14.22	< **0.001**	0.03	0.109
Std	2.396	1.789	2.376	2.416	2.145	2.170				
Final Performance										
Mean	5.409	4.952	5.452	5.541	4.827	5.381	2.063	0.068	0.01	0.017
Std	1.884	1.703	1.842	1.917	1.842	2.072				

* statistically significant difference among clusters based on adjusted alpha value.

Table 2: Pairwise comparisons for overall usage patterns.

		Number of Attempts			Hours Between Attempts		
A	**B**	**A: M (SD)**	**B: M (SD)**	**P-val**	**A: M (SD)**	**B: M (SD)**	**P-val**
0	1	145.205 (79.425)	729.548 (127.124)	< **0.01**	5.682 (1.272)	3.081 (0.816)	< **0.01**
0	2	145.205 (79.425)	112.415 (74.239)	**0.016**	5.682 (1.272)	10.862 (0.416)	< **0.01**
0	3	145.205 (79.425)	299.449 (45.966)	< **0.01**	5.682 (1.272)	7.735 (1.001)	< **0.01**
0	4	145.205 (79.425)	97.309 (89.727)	< **0.01**	5.682 (1.272)	1.556 (0.775)	< **0.01**
0	5	145.205 (79.425)	444.460 (67.291)	< **0.01**	5.682 (1.272)	4.584 (1.307)	< **0.01**
1	2	729.548 (127.124)	112.415 (74.239)	< **0.01**	3.081 (0.816)	10.862 (0.416)	< **0.01**
1	3	729.548 (127.124)	299.449 (45.966)	< **0.01**	3.081 (0.816)	7.735 (1.001)	< **0.01**
1	4	729.548 (127.124)	97.309 (89.727)	< **0.01**	3.081 (0.816)	1.556 (0.775)	< **0.01**
1	5	729.548 (127.124)	444.460 (67.291)	< **0.01**	3.081 (0.816)	4.584 (1.307)	< **0.01**
2	3	112.415 (74.239)	299.449 (45.966)	< **0.01**	10.862 (0.416)	7.735 (1.001)	< **0.01**
2	4	112.415 (74.239)	97.309 (89.727)	0.072	10.862 (0.416)	1.556 (0.775)	< **0.01**
2	5	112.415 (74.239)	444.460 (67.291)	< **0.01**	10.862 (0.416)	4.584 (1.307)	< **0.01**
3	4	299.449 (45.966)	97.309 (89.727)	< **0.01**	7.735 (1.001)	1.556 (0.775)	< **0.01**
3	5	299.449 (45.966)	444.460 (67.291)	< **0.01**	7.735 (1.001)	4.584 (1.307)	< **0.01**
4	5	97.309 (89.727)	444.460 (67.291)	< **0.01**	1.556 (0.775)	4.584 (1.307)	< **0.01**

terms of *final performance*. Lastly, we investigated how clusters differ in terms of HEXAD profiles (see Table 4), which revealed statistically non-significant differences and / or practically irrelevant effect sizes ($\eta_p^2 < 0.025$) for each profile. Thereby, while clusters differ from one another in terms of overall usage, final performance is similar despite the several differences in previous knowledge and performance gains, which seems not to be related to students' HEXAD profiles.

Based on these results, we interpret our clusters as follows. Cluster C1 represents *High-Gain Novices*. It demonstrates a scenario where students with the lowest initial knowledge managed to achieve the highest performance gain by engaging in a high number of attempts with small to moderate spacing. Clusters C3 and C5 concern *Spaced Practicers* and *Intensive Practicers*, respectively. These groups had moderate levels of initial knowledge but achieved moderate

Table 3: Pairwise comparisons for performance metrics.

		Performance Gains			Previous Knowledge		
A	**B**	**A: M (SD)**	**B: M (SD)**	**P-val**	**A: M (SD)**	**B: M (SD)**	**P-val**
0	1	1.447 (1.570)	3.070 (1.792)	< **0.01**	4.353 (2.396)	1.973 (1.789)	< **0.01**
0	2	1.447 (1.570)	1.643 (1.861)	0.847	4.353 (2.396)	4.220 (2.376)	0.702
0	3	1.447 (1.570)	2.460 (2.006)	< **0.01**	4.353 (2.396)	3.336 (2.416)	**0.023**
0	4	1.447 (1.570)	1.053 (1.493)	0.068	4.353 (2.396)	4.487 (2.145)	0.868
0	5	1.447 (1.570)	2.498 (2.130)	< **0.01**	4.353 (2.396)	3.059 (2.170)	< **0.01**
1	2	3.070 (1.792)	1.643 (1.861)	< **0.01**	1.973 (1.789)	4.220 (2.376)	< **0.01**
1	3	3.070 (1.792)	2.460 (2.006)	**0.029**	1.973 (1.789)	3.336 (2.416)	< **0.01**
1	4	3.070 (1.792)	1.053 (1.493)	< **0.01**	1.973 (1.789)	4.487 (2.145)	< **0.01**
1	5	3.070 (1.792)	2.498 (2.130)	**0.037**	1.973 (1.789)	3.059 (2.170)	< **0.01**
2	3	1.643 (1.861)	2.460 (2.006)	< **0.01**	4.220 (2.376)	3.336 (2.416)	< **0.01**
2	4	1.643 (1.861)	1.053 (1.493)	< **0.01**	4.220 (2.376)	4.487 (2.145)	0.492
2	5	1.643 (1.861)	2.498 (2.130)	< **0.01**	4.220 (2.376)	3.059 (2.170)	< **0.01**
3	4	2.460 (2.006)	1.053 (1.493)	< **0.01**	3.336 (2.416)	4.487 (2.145)	< **0.01**
3	5	2.460 (2.006)	2.498 (2.130)	0.939	3.336 (2.416)	3.059 (2.170)	0.492
4	5	1.053 (1.493)	2.498 (2.130)	< **0.01**	4.487 (2.145)	3.059 (2.170)	< **0.01**

performance gains with distinct strategies. *Spaced Practicers* (C3) engaged in fewer attempts but adopted a strategy of more DP, whereas *Intensive Practicers* (C5) engaged in a higher number of attempts with minimal spacing. Interestingly, both strategies yielded similar gains, albeit through different paths, for students who started with a moderate knowledge.

C0, C2, and C4 concern *Frequent Refreshers*, *Strategic Spacers*, and *Minimum Engagers*, respectively. They represent students with high initial knowledge but divergent performance gains and usage patterns. *Minimum Engagers* (C4), those with the smallest number of attempts and spacing, also presented the smallest performance gains. In contrast, *Frequent Refreshers* and *Strategic Spacers*, who outperformed *Minimum Engagers* in performance gains, differ in usage patterns. *Frequent Refreshers* (C0) engaged more frequently but with little spacing, while *Strategic Spacers* (C2) favored fewer, more spaced attempts.

4 Discussion and Final Remarks

Given the need for studies on RP's role in ecological settings, especially in STEM [8], our findings expand the literature with evidence of RP's role in CS1. Furthermore, by clustering and comparing CS1 students' usage data, our analysis revealed how different strategies might be more or less effective depending on the students' knowledge at the start of the course. Thus, our analysis also is an answer to recent literature by exploring how RP and DP relate [8].

Altogether, our findings have implications for both research and educational practice. These findings can guide the design of instructional resources and teaching strategies tailored to diverse student needs. For instance, educators can build on the *High-Gain Novices* pattern to foster learning environments that support frequent, moderately spaced practice for students needing foundational knowledge boosts, which can be operationalized by encouraging consistent engagement with structured schedules for frequent yet moderately spaced practice.

Table 4: Descriptive statistics and cluster comparison for HEXAD profiles.

	Clusters						ANOVA			
Stat	**C0**	**C1**	**C2**	**C3**	**C4**	**C5**	**F-stats**	**p-val**	**alpha**	η_p^2
N	44	62	188	98	81	113				
Philanthropists										
Mean	18.302	20.935	20.522	20.379	19.222	19.972	2.780	0.017	0.017	0.024
Std	5.596	4.044	4.599	4.468	4.207	4.547				
Socializers										
Mean	18.349	19.226	18.891	18.874	18.037	19.404	0.900	0.480	0.042	0.008
Std	4.730	4.543	5.398	4.425	4.635	4.699				
Free Spirits										
Mean	21.233	20.968	21.196	21.232	21.111	21.468	0.196	0.964	0.050	0.002
Std	3.785	3.750	3.383	3.237	3.560	3.387				
Achievers										
Mean	19.744	22.194	21.821	22.000	20.469	21.431	3.067	0.010	0.008	0.026
Std	5.287	4.068	4.345	4.207	4.287	4.315				
*Disruptors**										
Mean	16.093	14.145	14.690	14.453	16.395	14.661	2.675	**0.021**	0.025	0.023
Std	5.250	4.472	5.184	4.697	4.748	4.315				
Players										
Mean	19.581	21.452	21.114	21.011	19.901	21.064	2.098	0.064	0.033	0.018
Std	4.573	4.329	4.347	4.294	4.076	3.677				

* statistically significant difference among clusters based on adjusted α value.

Similarly, the *Spaced Practicers* and *Strategic Spacers* patterns can inform how educators promote DP in learning and assessment schedules, perhaps through flexible timing for assignment / revisiting topics. Moreover, our findings suggest that supporting the *Minimum Engagers* requires careful consideration. This is specially important because, despite the gamification, these students were still not motivated to engage with it more actively. Hence, by aligning instructional strategies with these empirically derived patterns, educators can better support varied learning needs and optimize outcomes across every student groups.

Furthermore, clusters such as *Spaced Practicers* and *Strategic Spacers* highlight the effectiveness of DP in light of previous knowledge and compared to RP. Research in that line could further investigate reasons why DP is beneficial for students with different initial knowledge levels, including novice and experienced ones, as well as how it compared to other usage patterns. This would not only inform refinements to existing learning theories but, given that our study domain is CS1, such studies would shed light into the understanding of metacognition and retention of complex material.

Importantly, using a single dataset that does not account for external factors may limit the findings' generalizability. Also, while HEXAD profiles and prior knowledge were considered, other characteristics may also influence engagement and learning outcomes. Therefore, we call future research employing experimental designs, expanding to diverse learning platforms, incorporating a broader range of learner characteristics, investigating how usage patterns evolve over time, and how adaptive interventions can optimize learning strategies.

Acknowledgement. This work was supported in part by a 2024 AIED-DEIA Fellowship grant from the IAIED Society (iaied.org).

References

1. Aguinis, H., Gottfredson, R.K., Joo, H.: Best-practice recommendations for defining, identifying, and handling outliers. Organizational research methods **16**(2), 270–301 (2013)
2. Baker, R.S., Martin, T., Rossi, L.M.: Educational data mining and learning analytics. The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications pp. 379–396 (2016)
3. Bennedsen, J., Caspersen, M.E.: Failure rates in introductory programming: 12 years later. ACM Inroads **10**(2), 30–36 (2019)
4. Goldstein, E.B.: Cognitive psychology: Connecting mind, research and everyday experience. Nelson Education (2014)
5. Maryono, D., Budiyono, S., Akhyar, M.: Implementation of gamification in programming learning: literature review. Int. J. Inf. Educ. Technol **12**(12), 1448–1457 (2022)
6. Mellado, R., Cubillos, C., Vicari, R.M., Gasca-Hurtado, G.: Leveraging gamification in ict education: Examining gender differences and learning outcomes in programming courses. Applied Sciences **14**(17), 7933 (2024)
7. Paiva, J.C., Leal, J.P., Figueira, Á.: Automated assessment in computer science education: A state-of-the-art review. ACM Transactions on Computing Education (TOCE) **22**(3), 1–40 (2022)
8. Pan, S.C., Dunlosky, J., Xu, K.M., Ouwehand, K.: Emerging and future directions in test-enhanced learning research. Educational Psychology Review **36**(1), 20 (2024)
9. Pankiewicz, M., Baker, R., Ocumpaugh, J.: Using intelligent tutoring on the first steps of learning to program: affective and learning outcomes. In: International Conference on Artificial Intelligence in Education. pp. 593–598. Springer (2023)
10. Pankiewicz, M., Bator, M.: On-the-fly estimation of task difficulty for item-based adaptive online learning environments. In: Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE. p. 317 – 323 (2021). https://doi.org/10.1145/3430665.3456305
11. Park, S., Yun, H.: Relationships between motivational strategies and cognitive learning in distance education courses. Distance Education **38**(3), 302–320 (2017). https://doi.org/10.1080/01587919.2017.1369007, https://doi.org/10.1080/01587919.2017.1369007
12. Pereira, F.D., Fonseca, S.C., Oliveira, E.H., Cristea, A.I., Bellhäuser, H., Rodrigues, L., Oliveira, D.B., Isotani, S., Carvalho, L.S.: Explaining individual and collective programming students' behavior by interpreting a black-box predictive model. IEEE Access **9**, 117097–117119 (2021)
13. Robins, A.V.: Novice programmers and introductory programming. The Cambridge Handbook of Computing Education Research, Cambridge Handbooks in Psychology pp. 327–376 (2019)
14. Rodrigues, L., Pereira, F., Toda, A., Palomino, P., Oliveira, W., Pessoa, M., Carvalho, L., Oliveira, D., Oliveira, E., Cristea, A., et al.: Are they learning or playing? moderator conditions of gamification's success in programming classrooms. ACM Transactions on Computing Education (TOCE) **22**(3), 1–27 (2022)

15. Rodrigues, L., Toda, A.M., Oliveira, W., Palomino, P.T., Avila-Santos, A.P., Isotani, S.: Gamification works, but how and to whom? an experimental study in the context of programming lessons. In: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education. pp. 184–190 (2021)
16. Sailer, M., Homner, L.: The gamification of learning: a meta-analysis. Educational Psychology Review (Aug 2019). https://doi.org/10.1007/s10648-019-09498-w, https://doi.org/10.1007/s10648-019-09498-w
17. Tondello, G.F., Mora, A., Marczewski, A., Nacke, L.E.: Empirical validation of the gamification user types hexad scale in english and spanish. International Journal of Human-Computer Studies **127**, 95–111 (2019)
18. Wilcox, R.R.: Introduction to robust estimation and hypothesis testing. Academic press (2011)