Clustering learners' feedback processing patterns based on their response latency

Wei Chu and Philip I. Pavlik Jr.
University of Memphis
wchu, ppavlik@memphis.edu

ABSTRACT
In intelligent tutoring systems (ITS) abundant supportive messages are provided to learners. One implicit assumption behind this design is that learners would actively process and benefit from feedback messages when interacting with ITS individually. However, this is not true for all learners; some gain little after numerous practice opportunities. In the current research, we assume that if the learner invests enough cognitive effort to review feedback messages provided by the system, the learner's performance should be improved as practice opportunities accumulate. We expect that the learner's cognitive effort investment could be reflected to some extent by the response latency, then the learner's improvement should also be correlated with the response latency. Therefore, based on this core hypothesis, we conduct a cluster analysis by exploring features relevant to learners' response latency. We expect to find several features that could be used as indicators of the feedback usage of learners; consequently, these features may be used to predict learners' learning gain in future research. Our results suggest that learners' prior knowledge level plays a role when interacting with ITS and different patterns of response latency. Learners with higher prior knowledge levels tend to interact flexibly with the system and use feedback messages more effectively. The quality of their previous attempts influences their response latency. However, learners with lower prior knowledge perform two opposite patterns, some tend to respond more quickly, and some tend to respond more slowly. One common characteristic of these learners is their incorrect response latency is not influenced by the quality of their previous performance. One interesting result is that those quick responders forget faster. Thus, we concluded that for learners with lower prior knowledge, it is better for them not to react hastily to obtain a more durable memory.

Keywords
response latency, clustering, fluency, feedback, intelligent tutoring system

1. INTRODUCTION
Abundant supportive messages are provided in intelligent tutoring systems (ITS). One implicit assumption behind the prevalent presentation of instructional feedback messages is that all learners will actively process these messages by investing cognitive effort and then benefit from the system's instructional feedback. Consequently, learners' learning process could be enhanced by this scaffolding information and obtain desired learning gains after interaction with the system. Unfortunately, this assumption does not always hold as shown in previous research. For example, some learners tend to game the system [2, 3] instead of cognitively processing the presented instructional feedback messages; other learners tend to overuse supportive messages [1] to obtain the correct answers without applying cognitive efforts. The ineffective interactions between learners and ITS lead to slow improvement, or even worse, learners may fail to achieve their learning goals. Therefore, it is useful to investigate the different feedback message usage patterns of learners in ITS. It is reasonable to expect that some learners will invest more cognitive efforts than others when processing feedback messages.

Response latency is an indicator with a long history of measuring people's cognitive processes that cannot be directly observed, such as fact memory and paired-association learning [7, 8]. Correct response latency for a specific item is an index to indicate the amount of information in memory about it. The shorter correct latency indicates more presumed information about the item in memory. On the other hand, incorrect response latency is an index to measure the learner's willingness to continue searching for memory about the item [8, 9]. It must be noted that response latency is not a sensitive indicator to all levels of cognitive processes and cannot be used to infer all level changes in people's cognitive states. For instance, three cognitive and learning processes have been proposed in the Knowledge-Learning-Instruction (KLI) framework [6]: memory and fluency-building processes, induction and refinement processes, and understanding and sense-making processes. Different levels of cognitive processes correspond to various kinds of knowledge components. A knowledge component (KC) is defined as an acquired unit of cognitive function or structure inferred from performance on a set of related tasks. It implies that cognitive processes are different in complexity and a higher-level cognitive process requires more interconnected knowledge components, as a result, the change of learners' cognitive states during such complicated processes may not be reflected accurately by their response latency. Therefore, to ensure the validity of the clustering based on response latency relevant features, we only focused on the memory and fluency processes that require basic and simpler declarative or conceptual fact KCs (e.g., foreign vocabulary memory; paired association learning).

In most ITS, for content that depends on memory and fluency processes, feedback messages are typically provided by the system after learners' incorrect attempts. These feedback messages are important for learners to correct their wrong memory traces or strengthen their existing correct but unstable memory traces to improve fluency. If learners invest cognitive efforts and actively process these feedback messages after failed attempts, they should benefit from the instructional information and their performance would gradually improve as practice opportunity increases. Otherwise, the time assigned to feedback messages would be wasted since passively receiving feedback contributes little to learners' memory trace strength [10]. To investigate how learners process the feedback differently, we conducted cluster analyses by exploring features calculated according to learners' response latency.

2. CLUSTERING

2.1 Hierarchical Clustering

Clustering techniques are widely used in educational data mining to assess the existence of separate groups of learners according to the similarity of their cognitive or behavioral patterns [4,5]. Learners in the same cluster have similar attributes to each other but have dissimilar attributes to the learners belonging to other clusters. For the explorational goal of the current research, we opted to use the Hierarchical Cluster Analysis (HCA) method since HCA allowed us to build a hierarchy of clusters without having a pre-specified optimal number of clusters. Specifically, Ward's method [12] was used in our clustering since the strongest clustering structure was identified by this method with a relatively high agglomerative coefficient (*ac* = 0.93). All the following cluster and statistical analyses were conducted in R.

2.2 Data and Data Cleaning

The data set we used here was collected from the authors' experiment designed to understand how native English speakers learn aural Chinese vocabulary. Three practice conditions were designed: only picture illustration context, only English translation context, and the combination of picture illustration and English translation. Specifically, we assumed that for learners, English translations in practice context would interfere with the acquisition of aural Chinese vocabulary. Participants were recruited from the Amazon Mechanical Turk (MTurk) platform, and they were asked to learn 27 aural Chinese words during the practice session. All practice trials are multiple-choice retrieval questions, and the choices for each trial consist of one correct answer and three alternatives randomly selected from the remaining 26 Chinese words. The formats of choices correspond to different practice conditions. After hearing the target aural Chinese word for each trial, the participants were given 5 seconds to select the corresponding meaning of the aural Chinese word. Feedback messages were given after incorrect attempts and learners have 5 seconds to review the messages. Specifically, in our experiment, during the feedback, the aural Chinese words were played one more time with the presentation of the correct choice. A total of 191 learners finished the practice session.

Before conducting clustering, we filtered out observations in raw datasets that may distort our analyses. Firstly, we cleaned the learners and KCs without enough observations. Learners were omitted if their observations were less than 25. KCs were omitted if their observations were less than 25. If a learner only experienced an example of a KC once, the observation was excluded. Secondly, the missing trial duration values and missing trial response latency values were inputted with the overall median values. Finally, all trials whose duration beyond the range of 0 to 5 seconds were removed.

After data cleaning, a new data frame was built for clustering. Each row in this data frame is one unique learner. The columns consist of features that we have mentioned in Table 1. All features were standardized before clustering. Learners whose absolute standardized values of any one of the features were larger than 3 were considered outliers and were omitted. A total of 175 learners remained for clustering.

2.3 Features Relevant to Response Latency

We calculated numerous features based on their sequential response latency to infer the learners' cognitive efforts investment when finishing retrieval practice and feedback review during the practice session with an online flashcard memory system. We hypothesized that if the learner actively engaged in the practice and reviewed the feedback messages carefully, then as the practice opportunity increases, the learner's accuracy on a certain KC should increase. At the same time, the response latency should decrease as the memory strength of the same KC increases. In other words, the learners' response fluency should increase with practice opportunity accumulation. To measure the sequential changes in learners' response latency, we calculated several features considering the learner's correctness of the earlier attempt. To inspect learners' performance after finishing the practice session, we included one feature for learning gains. Furthermore, learners have different prior knowledge levels before interacting with ITS, and this factor would affect learners' progress during practice, so it was also included in our features. Details of all features are shown in Table 1.

Table 1. Features relevant to latency and performance

Features relevant to response latency	
Feature label	**Description**
cor-cor	The average response latency change of consecutive correct and correct responses of each KC for each learner $RT_{n(correct)} - RT_{n-1(correct)}$
cor-incor	The average response latency change of consecutive correct and incorrect responses of each KC for each learner $RT_{n(incorrect)} - RT_{n-1(correct)}$
incor-cor	The average response latency change of consecutive incorrect and correct responses of each KC for each learner $RT_{n(correct)} - RT_{n-1(incorrect)}$
incor-incor	The average response latency change of consecutive incorrect and incorrect responses of each KC for each learner $RT_{n(incorrect)} - RT_{n-1(incorrect)}$
corRT	The average correct response latency on all KCs for each learner
incorRT	The average incorrect response latency on all KCs for each learner
$corRT_{(cor)}$	The average correct response latency on all KCs for each learner given the previous attempt for the same KC is correct
$corRT_{(incor)}$	The average correct response latency on all KCs for each learner given the previous attempt for the same KC is incorrect
$incorRT_{(cor)}$	The average incorrect response latency on all KCs for each learner given the previous attempt for the same KC is correct
$incorRT_{(incor)}$	The average incorrect response latency on all KCs for each learner given the previous attempt for the same KC is incorrect
$ratio_{(corRT)}$	$ratio_{(corRT)} = \frac{corRT_{(cor)}}{corRT_{(incor)}}$
$ratio_{(incorRT)}$	$ratio_{(incorRT)} = \frac{incorRT_{(cor)}}{incorRT_{(incor)}}$
ratio	$ratio = \frac{corRT}{incorRT}$
Features relevant to performance	
prior knowledge	The correct percentage for each learner after finishing the first attempts of all KCs

Features relevant to response latency	
Feature label	**Description**
learning gains	For each learner, the average correct percentage on the last two attempts of all KCs minus the average correct percentage on the first two attempts of all KCs

2.4 Feature Selection

Since one goal of the current exploration is to investigate the connection between learners' response latency changes during practice and their final performance at the end of the practice. The learning-gains feature was used as the dependent variable for our feature selection, and it was not included in the clustering since usually this value cannot be obtained during practice. First, we used all response latency-relevant features and the prior knowledge feature as predictors to build a linear regression model for the learning gains and the overall regression was statistically significant, $R^2 = 0.23$, F (14,160) = 3.39, p <.0005. Both the *prior knowledge* feature (β = -0.36, p < .0005) and the *cor-incor* feature (β = -0.30, p = 0.03) can significantly predict learners' learning gains. The $incorRT_{(cor)}$ feature (β = -0.80, p = 0.08), $incorRT_{(incor)}$ feature (β = 0.57, p = 0.09) and $ratio_{(incorRT)}$ feature (β = 0.94, p = 0.09) marginally predict learners' learning gains. Second, we inspected the correlation matrix for all features to avoid the multicollinearity problem as much as possible. The heatmap for correlations was shown in Figure 1. We found that only the $ratio_{(corRT)}$ feature (r=0.14, p=0.05) weakly correlated with the learning gains feature but it also significantly correlated with *prior knowledge* feature (r=-0.15, p=0.04), so it was not included in the following clustering. We also included the *ratio* feature as a baseline indicator for learners' response latency. Besides, since features *cor-cor* (r=0.70, p<.0005), *incor-cor* (r=0.64, p<.0005), *incor-incor* (r=0.70, p<.0005) highly correlated with feature *cor-incor*, they were also not used for clustering but the *cor-cor* feature was used for the result interpretation to reflect learners' fluency improvement. In summary, four features were used for clustering: *cor-incor, prior knowledge*, *ratio*, and $ratio_{(incorRT)}$.

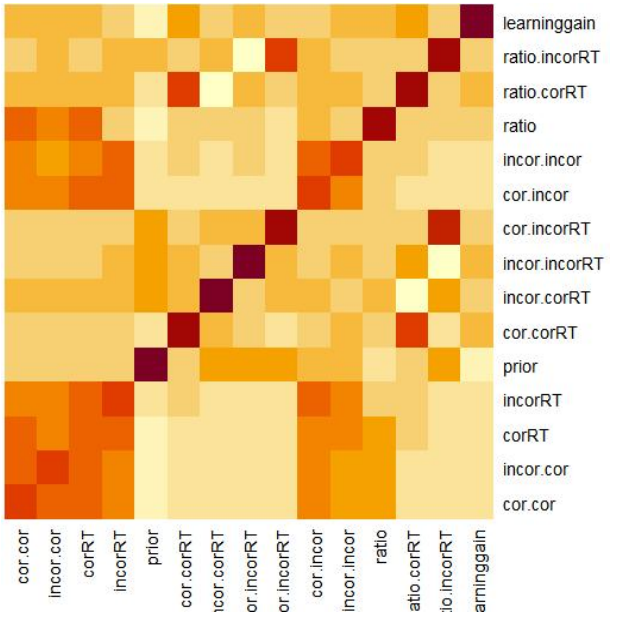

Figure 1. Heatmap for features' correlation matrix

3. RESULTS

By inspecting the dendrogram, we decided that a 4-cluster representation is appropriate for our dataset. To further justify this decision, we also referred to the silhouette method result and the average silhouette scores for 2-cluster and 4-cluster were 0.24 and 0.19, respectively. Even though the value was slightly lower for 4-cluster representation to explore more interesting patterns we still maintained the 4-cluster representation decision. Learning curves for all 4 clusters were shown in Figure 2. The patterns of response latency features, learning gains, and prior knowledge level for each cluster were shown in Figure 3. ANOVA tests were also conducted to compare the features' differences among 4 clusters and the results were shown in Table 2. All the differences in posthoc comparisons showed high significance with p <.0005 and were not explicitly mentioned in the following sections.

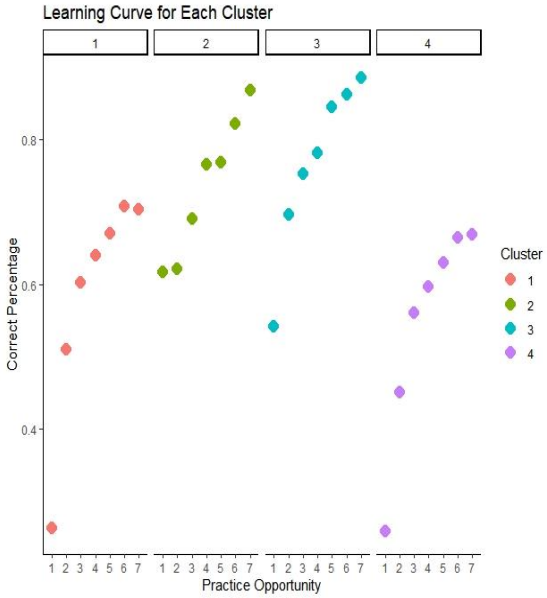

Figure 2. Learning curve for each cluster

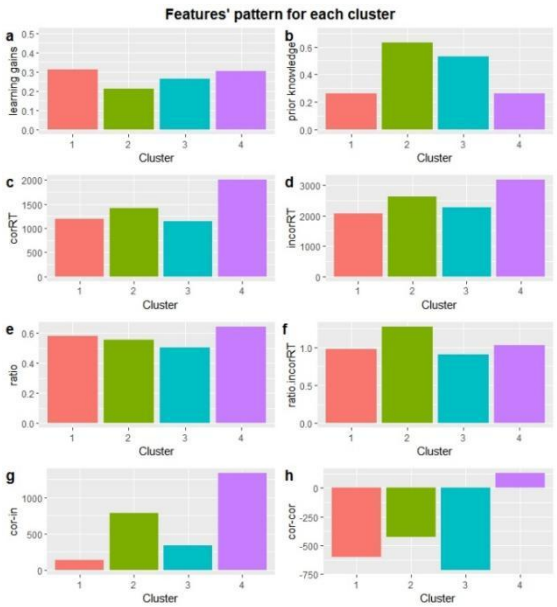

Figure 3. Features' pattern of each cluster and learning gains

Table 2. Descriptive and ANOVA results for 4 clusters

Features	Clusters				*F (df)*	*p*
	1 (*n*=69)	2 (*n*=15)	3 (*n*=36)	4 (*n*=55)		
prior knowledge	0.26 (0.11)	0.63 (0.14)	0.54 (0.11)	0.25 (0.11)	87.82 (3, 171)	<.0005
learning gains	0.31 (0.10)	0.20 (0.14)	0.25 (0.09)	0.31 (0.11)	5.71 (3, 171)	<.0005
correct RT (ms)	1206 (310)	1444 (654)	1143 (268)	2052 (488)	55.61 (3, 171)	<.0005
incorrect RT (ms)	2060 (545)	2595 (906)	2277 (571)	3133 (531)	36.40 (3, 171)	<.0005
ratio	0.59 (0.09)	0.55 (0.10)	0.50 (0.07)	0.65 (0.09)	18.36 (3, 171)	<.0005
$ratio_{(incorRT)}$	0.97 (0.11)	1.27 (0.22)	0.90 (0.16)	1.02 (0.16)	21.42 (3, 171)	<.0005
cor-cor	-581 (368)	-393 (641)	-702 (322)	139 (528)	36.10 (3, 171)	<.0005
cor-incor	125 (501)	818 (1085)	302 (634)	1317 (504)	44.31 (3, 171)	<.0005

3.1 Cluster 1

Learners in cluster 1 have low prior knowledge about aural Chinese vocabulary compared with learners from both cluster 2 and cluster 3. Their average correct response time (*corRT*) was as short as their counterparts with higher prior knowledge, but their average incorrect response time (*incorRT*) was even shorter than learners in cluster 2. Combined with the higher ratio (*ratio*=0.59) of cluster 1, the pattern of cluster 1 suggested that these learners might tend to respond as quickly as possible. The ratio of incorrect response latency ($ratio_{(incorRT)}$) which was approximately equivalent to 1 showed that their failed attempts' response latency was relatively stable and not influenced by the quality of their earlier attempts. The negative average sequential correct response latency change (*cor-cor*) reflected that for learners in cluster 1, their response latency was decreased as practice opportunities increased for the aural Chinese words they have learned. In other words, their response fluency was gradually increased for knowledge components with stable memory strength and minor risk of forgetting. The small positive value of the average sequential correct and incorrect response latency change (*cor-incor*) suggested that for aural Chinese words that have not been permanently learned, learners in cluster 1 seem unwilling to spend time actively searching from their memory before responding and reviewing the feedback messages provided by the system. When inspecting these characteristics of cluster 1 together, the falling trend of the learners' accuracy after six practice opportunities showed in the learning curve seems to be interpreted as some aural Chinese words that were temporally learned have started to be forgotten. The fastest forgetting characteristic distinguished cluster 1 from the other three clusters.

3.2 Cluster 2 & Cluster 3

Learners with some prior knowledge were assigned to cluster 2 and cluster 3. Both learning curves gradually increased as practice accumulated. The values of the average correct response latency (*corRT*), average incorrect response latency (*incorRT*), and the baseline of response latency (*ratio*) for both cluster 2 and cluster 3 were not significantly different. The values of the *ratio* feature for cluster 2 and cluster 3 were 0.55 and 0.50 respectively and this suggested that their average incorrect response latency was almost twice as slow as their average correct response latency. The negative average sequential correct response latency change (*cor-cor*) for both clusters showed that their fluency for already learned aural Chinese words increased as the practice accumulated. The most obvious differences between cluster 2 and cluster 3 were captured by the ratio of average incorrect response latency given the quality of the earlier attempt ($ratio_{(incorRT)}$) and the average sequential correct and incorrect response latency (*cor-incor*) features. For learners from cluster 2 (*ratio*=1.27), their incorrect response latency for a certain knowledge component (here each KC referred to one aural Chinese word) was influenced by the quality of their earlier attempt for the same KC. If the KC was answered correctly at the previous trial but cannot be recalled at the current trial, learners in cluster 2 performed a tendency to retrieve from their memory before responding for the temporally learned KC. This tendency might be used to explain the two little plateaus shown in the learning curve (see Figure 2.). In contrast, learners in cluster 3 (*ratio*=0.90) tended to respond faster to KC which they answered correctly in the previous trial compared to KC which they answered incorrectly in the previous trial. One likely interpretation for this pattern was that learners in cluster 3 applied some metacognitive skills and they made quick decisions about KCs that have been forgotten and used the feedback messages as another learning opportunity. However, this assumption cannot be verified at this time for the absence of questionnaire data.

3.3 Cluster 4

Learners in cluster 4 have a similar low prior knowledge level as learners in cluster 1. By inspecting the values of *corRT*, *incorRT*, and *ratio* features, compared with all three other clusters, learners in cluster 4 tended to respond more slowly. The higher *ratio* value (*ratio*=0.65) implied that the response speed of learners in cluster 4 was more consistent regardless of the accuracy of their responses. The ratio of incorrect response latency ($ratio_{(incorRT)}$) for cluster 4 was approximately equal to 1 and this pattern was consistent with cluster 1 which suggested that these learners' incorrect response latency was not influenced by the quality of their previous attempts. The positive average sequential correct response latency change (*cor-cor*) showed that for learners in cluster 4, their fluency has not improved after finishing the practice session. Furthermore, learners in cluster 4 have the maximum average sequential correct and incorrect response latency change (*cor-incor*) among all clusters and this indicated that they have a willingness to invest cognitive effort and retrieve from memory first about uncertain knowledge components before reviewing feedback. This probably explained the monotonic increasing learning curve for cluster 4. However, both cluster 1 and cluster 4 have not reached a high accuracy level after 7 practice opportunities. This implies we should investigate more detailed differences between cluster 1 and cluster 4. In our future research, we need to set other criteria for stopping practice (such as 95% accuracy) for all learners to obtain learning curves with the same asymptote, then we would observe how these learners reach the same learning goal differently. We expected to observe more dynamic changes from such learning curves.

4. DISCUSSION

With the development of intelligent tutoring systems and the prevalence of online learning, it is useful for us to investigate factors that influence learners' success or failure. The cognitive effort investment of learners is one of factors that directly influences their learning outcomes. Unfortunately, not all learners are willing to invest mental efforts when interacting with ITS individually. For instance, research has found that some learners tend to game the system, and this phenomenon is prevalent across many educational ITS [2]. When gaming the system, the learners try to correctly answer questions by systematically guessing or abusing hints [1] instead of paying attention to the learning materials. In the current research, we explore several features relevant to learners' response latency when interacting with an online flashcard learning system to learn foreign vocabulary. All practice trials are multiple-choice questions, and this test format is widely used in most foreign languages learning websites and apps, such as Duolingo. Even though multiple-choice questions are relatively easy to respond to, it is also more likely for learners to make their choice without mental effort. Therefore, we assume that learners' response latency for these multiple-choices questions could reflect their engagement during the practice.

Our results suggested that some learners (cluster 1) tended to respond as quickly as they can even though they do not have enough prior knowledge. Their rapid response pattern resulted in weaker final learning. In our experiment, each knowledge component was repeated seven times and no criteria were embedded. Thus, the practice time was shorter compared to some learning sessions that happened in real classrooms. Therefore, we expected that if a fixed practice time was used in the system as a criterion for stopping the practice, learners in cluster 1 would have a greater probability of gaming the system. Based on the pattern of cluster 1, we recommended that to obtain more durable memory about simple knowledge components, such as foreign vocabulary or declarative fact knowledge, it is better for learners with lower prior knowledge not to react hastily, but first, actively invest cognitive effort and try to search from their memory about knowledge components independently. Furthermore, in contrast with cluster 1, some learners with low prior knowledge tended to respond more slowly (cluster 4). For these learners, we expected that if they have no response after a certain time (e.g., 3 seconds), some support hints that remind them to try to guess and use feedback messages would be helpful to improve the use of practice time. Cluster 4 may represent students doing their best despite being fairly unskilled at metacognitive memory skills.

Finally, we found that compared with learners who have higher prior knowledge levels, and learners with lower prior knowledge levels, in both cluster 1 and cluster 4, their average incorrect response latency seems not influenced by the quality of their previous attempts for the same knowledge component. This result indicated that learners' prior knowledge was related to their interaction with ITS. Learners who have higher prior knowledge might also have better strategies and metacognitive skills which assure they can interact with the system more flexibly and use feedback messages more effectively. Additionally, the average *ratio* of correct response latency and incorrect response latency was around 0.5 for learners in both cluster 2 and cluster 3, and this suggested that generally fast learners allocate a greater proportion of time spent studying feedback and searching from memory rather than responding.

However, some limitations need to be noted. One is that only one data set was tested in the current research, and it is not clear if these findings can be generalized to other datasets. For example, all practice trials in our dataset are multiple-choice questions, and this question type is more likely to capture learners' response latency differences than other formats, such as cloze questions or other more time-consuming questions. Another limitation is the constraint of knowledge component complexity. These response latency-relevant features we investigated here may not be appropriate and sensitive for more complicated knowledge components that require higher levels of cognitive processes, such as reasoning and sense-making.

5. ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation Learner Data Institute (NSF #1934745).

6. REFERENCES

[1] Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help Seeking and Help Design in Interactive Learning Environments. *Review of Educational Research, 73*(3), 277–320. DOI= https://doi.org/10.3102/00346543073003277

[2] Baker, R. S., De Carvalho, A. M. J. A., Raspat, J., Aleven, V., Corbett, A. T., & Koedinger, K. R. (2009, June). Educational software features that encourage and discourage "gaming the system". *In Proceedings of the 14th international conference on artificial intelligence in education,* 475-482.

[3] Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19*(2), 185-224.

[4] Bouchet, F., Harley. JM., Trevors, GJ., Azevedo, R. (2013). Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *Journal of Educational Data Mining 5*(1), 104-146.

[5] Fang, Y., Lippert, A., Cai, Z., Chen, S., Frijters, J. C., Greenberg, D., & Graesser, A. C. (2021). Patterns of Adults with Low Literacy Skills Interacting with an Intelligent Tutoring System. *International Journal of Artificial Intelligence in Education.* DOI= https://doi.org/10.1007/s40593-021-00266-y

[6] Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the Science-Practice chasm to enhance robust student learning. *Cognitive Science, 36*(5), 757–798. DOI= https://doi.org/10.1111/j.1551-6709.2012.01245.x

[7] MacLeod, C. M., & Nelson, T. O. (1984). Response latency and response accuracy as measures of memory. *Acta Psychologica, 57*(3), 215–235. DOI= https://doi.org/10.1016/0001-6918(84)90032-5

[8] Millward, R. (1964). Latency in a modified paired-associate learning experiment. *Journal of Verbal Learning and Verbal Behavior, 3*(4), 309–316. DOI= https://doi.org/10.1016/s0022-5371(64)80071-2

[9] Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109–133. DOI= https://doi.org/10.1037/0033-2909.95.1.109

[10] Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning & Memory, 4*(3), 210–221. DOI= https://doi.org/10.1037/0278-7393.4.3.210

[11] Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory, 5*(6), 607–617. DOI= https://doi.org/10.1037/0278-7393.5.6.607

[12] Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, 58*(301), 236–244. DOI= https://doi.org/10.1080/01621459.1963.10500845