# Clustering the Learning Patterns of Adults with Low Literacy Skills Interacting with an Intelligent Tutoring System

Ying Fang[1], Keith Shubeck[1], Anne Lippert[1], Qinyu Cheng[1], Genghu Shi[1], Shi Feng[1], Jessica Gatewood[1], Su Chen[1], Zhiqiang Cai[1], Philip Pavlik[1], Jan Frijters[2], Daphne Greenberg[3] and Arthur Graesser[1]

[1]University of Memphis, [2]Brock University, [3]Georgia State University

{yfang2, kshubeck, alippert, qcheng, gshi, sfeng, jdgatewood16, schen4, zcai, ppavlik, grasser}@memphis.edu, jfrijters@brocu.ca, dgreenberg@gsu.edu

## ABSTRACT

A common goal of Intelligent Tutoring Systems (ITS) is to provide learning environments that adapt to the varying abilities and characteristics of users. To do this, researchers must identify the learning patterns exhibited by those interacting with the system. In the present work, we use clustering analysis to capture learning patterns in over 250 adults who used the ITS, *CSAL* (Center for the Study of Adult Literacy) *AutoTutor*, to gain reading comprehension skills. AutoTutor has conversational agentsth at teach literacy adults with low literacy skills comprehension strategies in 35 lessons. These comprehension strategies align with one or more of the following levels specified in the Graesser-McNamara theoretical framework of comprehension: *word, textbase, situation model* and *rhetorical structure*. We used the adult learners' average response times per question and performance across lessons to cluster the students' learning behavior. Performance was measured as the proportion of 3-alternative-response questions answered correctly. Lessons were coded on one of the four theoretical levels of comprehension. Results of the cluster analyses converged on four types of learners: proficient readers, struggling readers, conscientious readers and disengaged readers. Proficient readers were fast and accurate; struggling readers worked slowly but were not accurate; conscientious readers worked slowly and performed comparatively well; disengaged readers were fast but did not perform well. Interestingly, the behaviors of learners in different clusters varied across the four theoretical levels. Identifying types of readers can enhance the adaptivity of AutoTutor by allowing for more personalized feedback and interventions designed for particular learning behaviors.

## Keywords

CSAL; AutoTutor; Adult reader; Learner clustering; Intelligent Tutoring; Personalized Instruction

## 1. INTRODUCTION

### 1.1 AutoTutor

AutoTutor is a conversation-based intelligent tutoring system (ITS) that has promoted learning on a wide range of topics [9, 13, 22]. AutoTutor, on average, has shown learning gains of 0.8 σ [22] compared to various traditional teaching controls. AutoTutor holds a conversation with students following an expectation-misconception tailored (EMT) approach [11]. This is a tutoring dialogue made up of questions that assess a learner's understanding of the content by comparing it to expected answers or misconceptions in real time. Using this EMT approach, AutoTutor is constantly assessing the students by providing feedback, hints, pumps, prompts to guide learning of the content.

Traditional AutoTutor systems implement conversations called *dialogues* that model the interactions that occur between a single human tutor and human student. More recent versions of AutoTutor oftenemploy *trialogues* which are tutorial conversations between three actors: a teacher agent, a human learner, and a peer agent [10, 12]. Trialogues offer several affordances over dialogues. For example, in a trialogue setting, the human learner can model productive learning behaviors that are programmed into the peer agent. The peer agent may also express misconceptions that the human learner shares and the negative feedback received from the tutor agent can be directed to the peer agent instead of the human learner. This helps avoid many of the undesirable effects from receiving direct negative feedback. Trialogues help students master difficult material. For example, trialogues successfully helped students learn scientific reasoning skills in an AutoTutor offshoot called Operation ARA [20, 21].

Agent trialogues are implemented in *AutoTutor* for CSAL [9], an ITS developed at the Center for the Study of Adult Literacy (CSAL, http://csal.gsu.edu/content/homepage). The web-based system is designed to help adults with low literacy acquire strategies for comprehending text at multiple levels of language and discourse. The system includes two computer agents (a teacher agent and a peer agent) which have conversations with human learners and between themselves. The learners are guided through their learning process by the computer agents. These three-way conversations are designed to (a) provide instruction on reading comprehension strategies, (b) help the learner apply these strategies to particular texts, (c) assess the learner's performance on applying these strategies, and (d) guide the learner in using the digital facilities. While previous implementations of AutoTutor relied on written natural language input from the learner, the learners in AutoTutor for CSAL have difficulties with writing. Thus, this version of AutoTutor was designed so that students interact through point-and-click, answering multiple choice questions, or using drag-and-drop. The conversational feature of

AutoTutor still guides the learner, but the questions can be solved without typed input. The lessons typically start with a 2-3 minutes video that reviews a comprehension strategy. After the review, the computer agents scaffold students through the learning by asking questions, providing short feedback, explaining how the answers are right or wrong, and filling in gaps of information. Figure 1 is an example of the teacher agent (on the left) asking both the learner and the peer agent (on the right) to find the meaning of the word "bank" in the given context. The scores of both the human learner and peer agent are shown under their names. The learner chooses the answer by clicking while the peer agent gives his answer by talking.

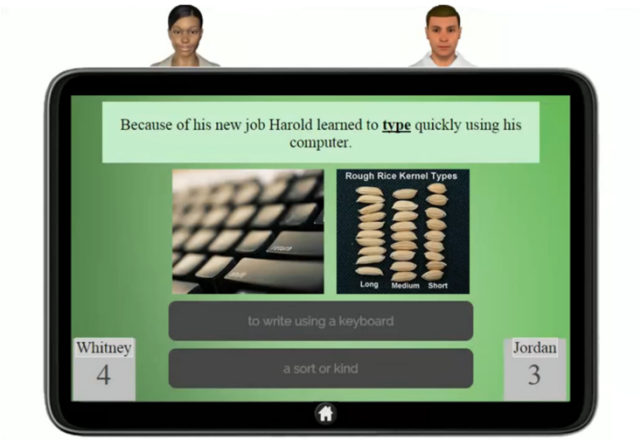


**Figure 1: Example trialogue with competition which focuses on the meaning of words from context.**

## 1.2 Theoretical Framework of Comprehension

The 35 lessons within AutoTutor align with the multilevel theoretical framework of comprehension proposed by Graesser and McNamara [13]. Six levels of comprehension were identified in Graesser and McNamara's framework. They are words, syntax, textbase, situation model, rhetorical structure/discourse genre, and pragmatic communication [13]. In this study, we focus on four of the six levels: *word, textbase, situation model and rhetorical structure.* The word level includes morphology, vocabulary and word decoding. The textbase level focuses on the explicit ideas in the text, but not the precise wording and syntax. The situation model refers to the subject matter content described in the text, including inferences activated by the explicit text. The model varies based on text type. In narrative text, the situation model includes the characters, objects, settings, events and other details of the story. In informational text, the model corresponds to substantive subject matter such as topics and domain knowledge. Rhetorical structure/discourse genre focuses on the category of text, such as narration, exposition, persuasion, and description. The word level represents the lower-level basic reading components, while the textbase, situation model and rhetorical structure level cover discourse components which were assumed to be more difficult to master [5, 24, 25].

## 1.3 Approaches of Categorizing Learners

A common goal of the learning sciences is to categorize learners based on their cognitive, motivational, and affective states. In the ITS domain, this is referred to as student modeling [23]. Student modeling is largely what enables ITS to be adaptive, with systems being designed to incorporate information pertaining to particular user characteristics. Specifically, ITS designers know that some specific cognitive states or behaviors are associated with learning and ensure the ITS can detect and respond appropriately to those features. Data mining approaches are often used to identify these attributes. For example, the ITS *Cognitive Tutor* employed a classifier to detect "gaming-the-system" behavior which occurred when users intentionally misused features of an ITS to progress through the content [1]. In another study that used data from students interacting with ALEKS (an ITS designed for math and science education), researchers were able to classify the learning persistence of a user as one of three distinct types [8]. Similarly, Del Valle and Duffy [7] clustered learners by their learning strategies in on online course and identified three types of learners: self-driven students, "get-it-done" students, and procrastinators. In another study, Wise et al. [27] clustered learners' online listening behaviors, and found three types of listeners with distinct behavioral patterns: superficial listeners, broad listeners and concentrated listeners.

In addition to categorizing or identifying learning behaviors from interacting with a system, researchers also categorize students based on individual differences in skill or knowledge gained a priori certain educational interventions. For example, in the ITS domain, students are often assessed on their prior knowledge of the domain material before interacting with an ITS, or at the early stages of the ITS content. They are commonly classified as having either high domain knowledge or low domain knowledge. There is evidence that high versus low domain knowledge students interact with ITS differently and require different pedagogical approaches to effectively learn from them. For example, an ITS using a vicarious learning design may benefit high domain knowledge students less than low-domain knowledge students [6]. This supports the idea that students with low-domain knowledge benefit more by observing peer agents or virtual tutees interacting with tutor agents. There is also evidence that high domain knowledge students sometimes suffer from an "expertise reversal effect" when presented with content they already understand [18]. When equipped with information about a learner's level of domain knowledge, ITS can leverage different pedagogical strategies to best cater to that student's capabilities.

The present study utilizes clustering analysis to achieve two goals. First, we characterize the behaviors of adults with low literacy skills who interacted with AutoTutor. Second, we examine whether adult readers' learning behaviors are associated with the different reading comprehension levels described above.

# 2. DATASETS AND DATA PROCESS

## 2.1 Data Sets

The data sets used for this study were taken from three waves of an intervention study consisting of 253 adult learners. The students participated in approximately 100 hours of hybrid classes which consisted of teacher-led sessions and AutoTutor sessions. The students took the Woodcock Johnson III Passage Comprehension subtest [28] before and after the intervention. While studying with AutoTutor, the logs of students' online learning activities were recorded by the system. The log file included learner information, class information, lesson and question information, response time and learning outcome.

In the intervention studies, 26 out of 35 AutoTutor lessons were assigned to the students; these 26 lessons were used for our analyses. We coded theoretical level of the lessons according to their primary theoretical levels. The classification of the primary theoretical levels is based on four discrete levels: Word, Textbase, Situation Model and Rhetorical Structure. The major components of Word lessons are word parts, word-meaning clues, learning new words and multiple meaning words. The Textbase lessons focus on pronouns, punctuation, key information, and main ideas. The Situation Model lessons mainly cover nonliteral language, connecting ideas, and inferences from text. The Rhetorical Structure lessons covered purpose of texts, steps in procedures, problems and solutions, cause and effect, compare and contrast, time and order, and other categories of rhetorical composition. In each lesson there are 10 to 35 questions. The questions in most lessons fall into two different difficulty levels. Normally a lesson starts with 10-15 medium level questions. Depending on the performance of the medium level questions, the learners are branched into hard or easy level questions in that lesson.

## 2.2 Data Process

First, we removed hard and easy questions so that only medium level questions were included in our statistical analyses. The reason for removing easy and hard questions was that response time was an important measure in the analysis, and response time could be confounded by using different question difficulty levels. Second, we removed motivational items; a motivational item was defined as any item that all the students answered correctly. These items could not be used for discriminating students and therefore they were removed from the analysis. Third, we examined the response time on each question and removed the outliers. According to the experimenters, the adult students infrequently took long breaks without logging out the system for various reasons, which led to some observations with extremely long response time. Following the rule of thumb about extreme outliers [19], we removed the response time which was three IQR (i.e. interquartile range) higher than the third quartile. For the lower end, the rule did not apply, so we replaced the bottom 1% of the observations with response times between 0 and 2 seconds with 3 seconds. The original log file had 102,519 observations. After data screening and cleaning, there were 42,289 observations from 253 students in dataset.

Next, we aggregated the data to student level and created variables for analyses. The aggregation was performed twice and two sets of features were created for analyses using the process described below.

In the log file, each observation represents an attempt that a student made on answering a question. All the students attempted multiple lessons, and within each lesson there were multiple questions, so each student had multiple observations in the log. The variables we used for the aggregations were the system-generated student ID, theoretical level of lessons, response time, and learning outcome. Each lesson was coded with a specific theoretical level (Word, Textbase, Situation Model or Rhetorical Structure) and the questions within the lesson were specific to the lesson's level. Response time was the time the learner spent working on the question, excluding the reading time. Learning outcome was either correct or incorrect. We aggregated the data based on these variables and calculated each student's average response time and accuracy at the four theoretical levels. After aggregation, the observations for each student were decreased to eight. The eight observations represented the average response time and accuracy at Word, Textbase, Situation Model and Rhetorical Structure levels. Response time was initially measured in seconds, which was a continuous variable. This measure remained the same after aggregation. Accuracy was a binary variable (i.e. 1 or 0) initially, but it became a continuous proportion correct variable after aggregation. Next, we changed the data format and combined the eight observations associated with one student into one observation with eight features. After this, there were 253 observations and each observation represented one student. The eight features were response time for Word, Textbase, Situation Model, and Rhetorical Structure level items, as well as the proportion correct for Word, Textbase, Situation Model, and Rhetorical Structure level items. This was how we created the first set of features. For the second feature set, we split response time into response time on correct answers and incorrect answers. Therefore, the response time features doubled from four to eight and the number of performance features remained four. Put together, we created two sets of features through aggregation. The first set had eight features and the second had twelve.

# 3. DATA EXPLORATION

Before data mining was carried out, we examined the student sample's response time and accuracy as a whole at the four theoretical levels to see whether response time was associated with theoretical level. The mean response time and accuracy at each level is shown in Table 1.

**Table 1: Means and standard deviations of response time and accuracy at four theoretical levels.**

| | Response Time | Response Time (Correct) | Response Time (Incorrect) | Accuracy |
|---|---|---|---|---|
| Word | 34.31 ($\sigma$ = 23.55) | 32.53 ($\sigma$ = 12.91) | 36.73 ($\sigma$ =16.71) | 0.67 ($\sigma$ = 0.47) |
| Textbase | 35.15 ($\sigma$ = 23.38) | 34.06 ($\sigma$ = 11.23) | 40.91 ($\sigma$ = 17.44) | 0.65 ($\sigma$ = 0.48) |
| Situation Model | 30.28 ($\sigma$ = 22.81) | 28.18 ($\sigma$ = 9.15) | 36.29 ($\sigma$ = 13.58) | 0.69 ($\sigma$ = 0.46) |
| Rhetoric Structure | 31.43 ($\sigma$ = 23.95) | 29.11 ($\sigma$ = 11.10) | 38.87 ($\sigma$ = 12.66) | 0.69 ($\sigma$ = 0.46) |

One-way ANOVAs were conducted to compare the means of response time, response time on correct items, response time on incorrect items and accuracy between the four theoretical levels. Results of the ANOVAs indicated that there were no significant differences between the four theoretical levels on response time or accuracy ($F(3, 996) = 1.90$, $p = 0.129$). However, we found theoretical level of the text affected both the time to give a correct response ($F(3, 996) = 17.75$, $p < 0.001$), and the time to give an incorrect response ($F(3, 996) = 6.02$, $p < 0.001$). Post hoc comparisons using the Tukey HSD test indicated that the average response time on correct attempts was longer at Word and Textbase levels than that of Situation Model and Rhetoric Structure levels. The average time on incorrect attempts at Textbase level was higher than that of Word and Situation Model levels. Since the differences found in response time on correct answers and incorrect were not consistent and did not show any pattern, we decided to group the students through clustering to investigate if theoretical levels influenced adult learners in a more nuanced way.

# 4. CLUSTER ANALYSES

Cluster analysis is a statistical exploratory tool used to find similar groups in an unsupervised fashion. It partitions objects into clusters so that the objects in the same cluster are more similar to each other than to those in other clusters. In educational settings, successful clustering has been achieved and the researchers identified learner groups with different behavioral patterns [3, 7, 27]. For example, Wise et al. [27] clustered learners' online listening behaviors and found three types of listeners with distinct behavioral patterns: superficial listeners, broad listeners and concentrated listeners. A similar goal can be transferred to our current context, with clustering possibly identifying groups with different learning behaviors across the four theoretical levels.

## 4.1 K-means Cluster Analysis

To carry out our clustering analysis, we applied a k-means clustering algorithm to our data. K-means clustering fits data points into clusters by iteratively reassigning and re-averaging the cluster centers until the points have reached convergence [15,16]. It is a common choice for clustering data since it is simple, effective and relatively efficient. We used R (version 3.3.3) to group students according to the k-means clustering algorithm of Hartigan and Wong [15].

**Table 2: Cluster means and standard deviations on the eight features.**

| | Cluster 1 (n = 64) | Cluster 2 (n = 45) | Cluster 3 (n = 88) | Cluster 4 (n = 53) |
|---|---|---|---|---|
| Time (Word) | 24.07 (σ = 7.02) | 46.21 (σ = 13.27) | 35.40 (σ =9.32) | 31.33 (σ = 8.37) |
| Time (Textbase) | 25.80 (σ = 5.33) | 51.31 (σ = 9.98) | 36.15 (σ =5.87) | 34.41 (σ = 7.76) |
| Time (Situation) | 22.40 (σ = 4.48) | 43.87 (σ = 7.51) | 29.76 (σ =4.90) | 31.57 (σ = 7.76) |
| Time (Rhetorical) | 22.10 (σ = 4.81) | 44.36 (σ = 7.77) | 31.86 (σ =5.06) | 32.72 (σ = 7.36) |
| Accuracy (Word) | 0.73 (σ = 0.13) | 0.69 (σ = 0.14) | 0.71 (σ = 0.13) | 0.54 (σ = 0.17) |
| Accuracy (Textbase) | 0.74 (σ = 0.13) | 0.70 (σ = 0.18) | 0.71 (σ = 0.12) | 0.54 (σ = 0.12) |
| Accuracy (Situation) | 0.73 (σ = 0.11) | 0.65 (σ = 0.08) | 0.74 (σ = 0.07) | 0.59 (σ = 0.11) |
| Accuracy (Rhetorical) | 0.75 (σ = 0.08) | 0.72 (σ = 0.09) | 0.71 (σ = 0.07) | 0.61 (σ = 0.10) |

Our choice to start with K=4 was guided by previous research. We assumed both engagement and disengagement existed while adult learners interacted with AutoTutor. For disengagement, a recent study on AutoTutor reported three types of behaviors associated with disengagement [14]. For engagement, another study used personalized time on item as a classifier, which was regard as a single type of behavior [21]. Put together, we assume there were four types of predominant behaviors that separate the learners into 4 clusters.We performed k-means clustering with k=4 twice: once with eight features and once with twelve features. As explained in section 2.2 (Data Process), the twelve features were developed from the eight features by dividing response time into response time on correct answers and incorrect answers. We also experimented with k = 3 and k = 5 and using the two feature sets. Compared with the 4-cluster solution, the 3-cluster solution lost some meaningful information. In the 5-cluster solution, two clusters had similar patterns. Therefore, we selected 4 as the optimum number of clusters. The results of the 4-cluster solution using eight features and twelve features are shown in Table 2 and Table 3, respectively.

We compared the 4-cluster solutions using 8 features with the one using 12 features. The four clusters showed similar patterns in the two solutions. The results of ANOVAs and post hoc tests comparing cluster differences on the grouping variables indicated similar between-cluster differences on response time and accuracy. We also tried k=3 and k=5 clustering, and compared the solutions from 8 features to that from 12 features. Both results indicated the consistency between solutions using 8 and 12 features. We further conducted Pearson correlation on the time variables (i.e. response time at different theoretical levels) with split time variables (i.e. response time on correct attempts and incorrect attempts at different theoretical levels). The results indicated significant moderate to strong correlations between these variables. The comparisons and statistical analyses suggested that splitting response time into two features did not contribute much to the discovery of the underlying structure. Following the principle of parsimony, we selected 8 features over 12 features for further analyses.

**Table 3: Cluster means and standard deviations on the twelve features**

| | Cluster 1 (n = 63) | Cluster 2 (n = 52) | Cluster 3 (n = 54) | Cluster 4 (n = 81) |
|---|---|---|---|---|
| Time on Correct (Word) | 24.08 (σ = 6.75) | 44.57 (σ = 16.03) | 35.47 (σ =8.11) | 30.05 (σ = 9.15) |
| Time on Incorrect (Word) | 23.78 (σ = 9.50) | 47.59 (σ = 15.87) | 48.03 (σ =14.67) | 31.55 (σ = 8.91) |
| Time on Correct (Textbase) | 24.98 (σ = 5.53) | 46.47 (σ = 10.36) | 32.95 (σ =6.41) | 33.90 (σ = 8.60) |
| Time on Incorrect (Textbase) | 31.23 (σ = 11.12) | 57.05 (σ = 19.39) | 43.25 (σ =15.59) | 36.79 (σ = 8.13) |
| Time on Correct (Situation) | 22.39 (σ = 3.84) | 39.78 (σ = 8.59) | 26.80 (σ =4.53) | 27.71 (σ = 5.46) |
| Time on Incorrect (Situation) | 27.35 (σ = 8.69) | 50.84 (σ = 10.84) | 34.91 (σ =11.17) | 34.82 (σ = 10.15) |
| Time on Correct (Rhetorical) | 20.04 (σ = 5.17) | 41.47 (σ = 10.24) | 27.55 (σ =4.76) | 28.58 (σ = 5.69) |
| Time on Incorrect (Rhetorical) | 29.11 (σ = 7.19) | 51.85 (σ = 12.47) | 40.21 (σ =8.83) | 37.22 (σ = 7.56) |
| Accuracy (Word) | 0.73 (σ = 0.12) | 0.69 (σ = 0.15) | 0.76 (σ = 0.12) | 0.57 (σ = 0.15) |
| Accuracy (Textbase) | 0.73 (σ = 0.12) | 0.69 (σ = 0.18) | 0.76 (σ = 0.12) | 0.58 (σ = 0.12) |
| Accuracy (Situation) | 0.73 (σ = 0.11) | 0.66 (σ = 0.09) | 0.76 (σ = 0.09) | 0.64 (σ = 0.10) |
| Accuracy (Rhetorical) | 0.73 (σ = 0.09) | 0.72 (σ = 0.11) | 0.73 (σ = 0.08) | 0.64 (σ = 0.09) |

## 4.2 Hierarchical Cluster Analysis

In addition to k-means clustering, we performed hierarchical cluster analysis, since k-means clustering is sensitive to the initial

centroids and also does not do well with clusters with non-spherical shape and different size [16]. Hierarchical clustering is different from k-means clustering that directly divides a dataset into a number of disjoint groups. Hierarchical clustering proceeds successively either by merging smaller clusters into larger ones (bottom-up), or by splitting larger clusters into smaller clusters (top-down) [17]. We performed hierarchical clustering using Ward's method [26], and compared 4-cluster solution with 3-cluster and 5-cluster solution. Similar to what we found with k-means clustering, 4-clustering solution was most meaningful.

With the help of an R package clVaid [4], we compared the 4-cluster solution based on k-means clustering algorithm with the 4-cluster solution based on hierarchical clustering algorithm. The scores of the two solutions on three measures were computed. The measures were connectivity, Silhouette Width, and Dunn Index. Connectivity measures the degree of connectedness of the clusters based on the k-nearest neighbors, and the better solution minimizes it. The connectivity scores for k-means and hierarchical solutions were 118.69 and 13.31. Silhouette Width and the Dunn Index measure compactness and separation of the clusters. A higher Silhouette value indicates higher degrees of confidence in a clustering solution and a higher Dunn score indicates a better separated clustering solution. The Silhouette value for k-means and hierarchical solutions were 0.17 and 0.33, and the Dunn scores for k-means and hierarchical clustering were 0.10 and 0.27. Hierarchical clustering outperformed k-means clustering on all three measures, so the final solution we selected was the 4-cluster solution based on hierarchical clustering algorithm. The response time and performance accuracy of the four clusters is shown in Figure 2.

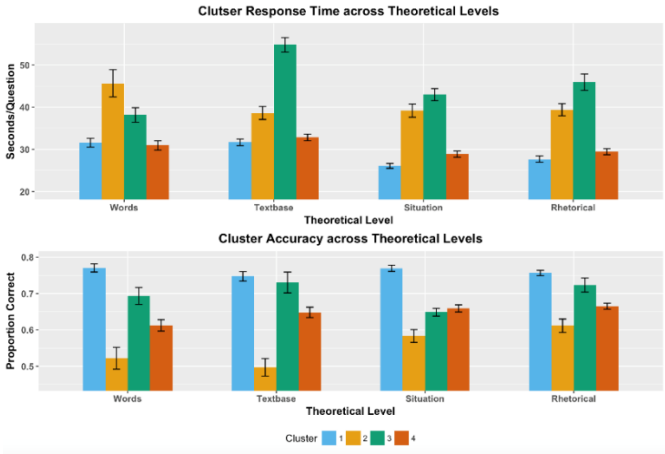


**Figure 2: Time and accuracy of four clusters at four different theoretical level.**

To compare the accuracy and time across clusters at different theoretical levels, we performed linear mixed-effects models using lme4 package in R [2]. We analyzed the effects of cluster and theoretical level on both proportion correct scores and time per question. In both models, subjects were specified as a random factor to control for the subject variance. For proportion correct scores, there was a statistically significant interaction between cluster and theoretical level, $F(9, 747) = 4.38$, $p<.001$, with the percentage of variance explained being 37.79%. For time per question, there was also a statistically significant interaction between cluster and theoretical level, $F(9, 747) = 11.45$, $p<.001$, vairance explained = 58.35%. However, time per question should be interpreted with caution since Situation Model and Rhetorical Structure lessons have multiple questions per text, which would shorten the expected time per question as learners had already built up their mental model for the text for most of the questions. Given these interactions, we will discuss the patterns of each cluster separately.

***Cluster 1: Proficient readers***

Cluster 1 is the biggest cluster with 39% (n = 98) of the study sample. These learners can be distinguished by their high speed and accuracy. As indicated by the results of mixed-effects models, the response time of Cluster 1 was shorter than the other three clusters at Situation Model and Rhetorical Structure level. At Word and Textbase level, there was no significant difference between the response time of Cluster 1 and Cluster 4, and Cluster 1 was faster than Cluster 2 and Cluster 3. Meanwhile, Cluster 1 achieved the highest proportion correct scores across all theoretical levels. Due to the students' high accuracy and short response time, we named this cluster "proficient readers." Proficient readers did not seem to be affected by theoretical level for accuracy, since they did equally well in lessons across different levels.

***Cluster 2: Struggling readers***

Cluster 2 is a smaller cluster with 12% of the study sample (n = 31). The response time of the learners in this cluster was comparatively long, and their accuracy was lower than the other clusters. According to the results of mixed-effects models, the response time of Cluster 2 on Word level questions was the longest, but their accuracy was the lowest. For Textbase, Situation Model and Rhetorical Structure level questions, the response time of Cluster 2 was the second longest, yet their accuracy remained the lowest among the four clusters. Due to the poor performance and long response time, we called this cluster "struggling readers." Unlike proficient readers who had stable performance across different theoretical levels, struggling readers did better in Situation Model and Rhetorical Structure lessons than Word and Textbase lessons.

***Cluster 3: Conscientious readers***

Cluster 3, like Cluster 2, contains 12% of the study sample (n = 31). The learners in Cluster 3 worked slowly and they achieved comparatively high performance accuracy. At Textbase, Situation model and Rhetorical Structure levels, the response time of Cluster 3 was the longest among the four clusters. Only at the Word level was the response time of Cluster 3 the second longest, trailing Cluster 2. Contrary to struggling readers who also worked slowly, Cluster 3 had the second highest accuracy. The proportion correct score differences between Cluster 1 and Cluster 3 ranged between 0.02 and 0.08 at different levels. We named this cluster "conscientious readers" because they achieved comparatively high accuracy through more effort. Similar to struggling readers, the conscientious readers' performance was associated with theoretical level. The results of mixed-effects models indicated that their performance at Textbase level was better than other levels.

***Cluster 4: Disengaged readers***

Cluster 4 is another large group representing 36% (n = 93) of the study sample. The learners in this cluster were almost as fast as the proficient readers, but their accuracy was comparatively low among the four groups. In particular, Cluster 4 learners were less

accurate than both proficient readers and conscientious readers. The response time of Cluster 4 was as short as Cluster 1 at Word, Situation Model and Rhetorical Structure level. At Textbase level, the response time of Cluster 4 was the second shortest. However, there was a large gap between the performance of Cluster 4 and Cluster 1. Results of mixed-effects models indicated learners in Cluster 1 and Cluster 4 differed in their proportion correct score, and this difference ranged between 0.09 to 0.16, depending on the theoretical level. We named learners in Cluster 4 "disengaged readers" because of their short response time and comparatively poor performance. Theoretical level also affected disengaged readers, since they performed worse on Word level lessons than Textbase, Situational Model and Rhetorical Structure level lessons.

# 5. DISCUSSION

We developed AutoTutor for CSAL to teach adults with low literacy skills reading comprehension strategies in 35 lessons [9]. The lessons align with one or more of the following levels specified in Graesser-McNamara theoretical framework of comprehension [13]: Word, Textbase, Situational Model and Rhetorical Structure. To better understand how low literacy adult students interact with AutoTutor, we analyzed the online learning log of 253 adult students who participated in three intervention studies. Our first goal was to classify adult learners' behavior patterns while they interacted with AutoTutor. Our second goal was to investigate whether adult learners' behaviors were associated with different reading components represented by the theoretical levels.

Regarding the first goal, we identified four clusters of adult learners with distinctive learning behaviors through cluster analysis. We named the four clusters proficient readers, struggling readers, conscientious readers and disengaged readers. Proficient readers worked fast and accurately. Among the four clusters, the response time of the proficient readers was the shortest, meanwhile, their accuracy was the highest at the four theoretical levels. Opposite to proficient readers, struggling readers worked slowly and inaccurately. Their response time was either the longest or the second longest at different theoretical levels, however, their accuracy remained the lowest overall. Conscientious readers also worked slowly, but unlike struggling readers, they achieved comparatively high accuracy. The response time of conscientious varied across the theoretical levels, but they achieved similar high accuracy at all the theoretical levels. This indicated their awareness of their skill level versus effort needed for mastery. Similar to proficient readers, disengaged readers worked fast. However, their performance was not as good. These readers might try to get through lessons quickly without paying much attention to the content.

With respect to the second goal, we found learning behaviors of individuals in the four clusters varied across theoretical levels in different ways. Proficient readers performed equally well at different theoretical levels, but they spent less time on Situation Model and Rhetorical Structure level questions. One possible explanation for the variation in time across theoretical levels is that Situation Model and Rhetorical Structure lessons have many questions in each text. This could shorten the expected time per question as learners built their mental models after the first a few questions. Struggling readers' behaviors indicated an obvious effect of theoretical level. Struggling readers performed worse on questions addressing Word and Textbase levels than Situation Model and Rhetorical Structure levels. Although struggling readers' performance was poor, they were comparatively better at lessons with discourse components. For conscientious readers, their behavior on Textbase level lessons stood out. These readers spent more time on Textbase level questions, and as a result, they achieved higher accuracy on these questions than for questions addressing other theoretical levels. The behavior of disengaged readers varied the most when data for questions that tapped basic reading components and those questions concerning discourse components. Despite spending a similar amount of time on questions addressing Word and discourse levels, disengaged readers performance was better for discourse level material.

According to previous research [5, 24, 25], Word items place lower loads on working memory than discourse items. We thus assumed discourse level items would be more difficult than word level items, leading to better performance for the latter item type. Yet our data indicated that this assumption did not apply to the adult readers. Among the four types of readers we identified, the behavior of proficient readers and conscientious readers was not affected by whether the items tapped basic reading level or the discourse level processes. We considered that disengaged readers and struggling readers might be influenced by the distinction in item type (basic versus discourse), but the trend in our data actually showed the opposite, with higher accuracy of discourse level items than word level items. In addition to finding that behavior differed across clusters when comparing word to discourse levels, we also found that behavior varied between the three discourse levels. For example, the performance of conscientious readers was best for Textbase level items, but the performance of struggling readers was best for Rhetorical Structure level items. Our finding that learner behavior varies by discourse level suggests these levels represent distinguishable components of comprehension, and supports previous work on AutoTutor, which found the three discourse levels were separable since they were not highly correlated [14].

Based on the findings of this study, we suggest that clustering methods can be used to enhance the adaptivity of ITS. In particular, assessments and feedback can be personalized to assist different groups of students that exhibit particular patterns of learning behaviors. Differences in time and accuracy on theoretical levels indicate that ITS implementations that provide feedback on accuracy alone or on time alone would be misguided. Feedback and assessment in ITS that take into account both student trends in accuracy and time and their interaction, or lack of interaction, with theoretical level should better target the student type and prove to be more appropriate.

Apart from separating learners according to their distinct behavior patterns, we could also identify the learners' strength and weakness with regards to specific types of learning material. For example, some readers may struggle with word level but excel at discourse level comprehension. These readers might benefit more if the instruction is tailored towards word level comprehension training. Feedback based on a generalization that contains only one of these levels may be ineffective and miss groups of students entirely.

# 6. ACKNOWLEDGMENTS

This research was supported by the National Center of Education Research (NCER) in the Institute of Education Sciences (IES) (R305C120001) and the National Science Foundation Data Infrastructure Building Blocks program under Grant No. (ACI-1443068). This work is partially supported by the National Science Foundation Data Infrastructure Building Blocks program under Grant No. (ACI-1443068).

## 7. REFERENCES


[1] Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004, April). Off-task behavior in the cognitive tutor classroom: when students game the system. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 383-390). ACM.

[2] . Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version, 1*(7), 1-23

[3] Bliuc, A. M., Ellis, R., Goodyear, P., & Piggott, L. (2010). Learning through face-to-face and online discussions: Associations between students' conceptions, approaches and academic performance in political science. *British Journal of Educational Technology, 41*(3), 512-524.

[4] Brock, G., Pihur, V., Datta, S., & Datta, S. (2011). clValid, an R package for cluster validation. *Journal of Statistical Software 25* (4).

[5] Cain, K. (2010). *Reading development and difficulties.* Oxford: Wiley-Blackwell.

[6] Craig, S. D., Gholson, B., Brittingham, J. K., Williams, J. L., & Shubeck, K. T. (2012). Promoting vicarious learning of physics using deep questions with explanations. *Computers & Education*, *58*(4), 1042-1048.

[7] Del Valle, R., & Duffy, T. M. (2007). Online learning: Learner characteristics and their approaches to managing learning. *Instructional Science*, *37*(2), 129-149.

[8] Fang, Y., Nye, B. D., Pavlik, P. I., Xu, Y. J, Graesser, A. C., & Hu, X. (2017, June). *Online learning persistence and academic achievement.* In Hu, X., Barnes, T., Hershkovitz, A., & Paquette, L. (Eds.), Proceedings of the 10th International Conference on Educational Data Mining (pp. 312-317). Wuhan, China: International Educational Data Mining Society.

[9] Graesser, A.C., Cai, Z., Baer, W.O., Olney, A.M., Hu, X., Reed, M., & Greenberg, D. (2016). Reading comprehension lessons in AutoTutor for the Center for the Study of Adult Literacy. In S.A. Crossley and D.S. McNamara (Eds.). *Adaptive educational technologies for literacy instruction* (pp. 288-293). New York: Taylor & Francis Routledge.

[10] Graesser, A. C., Forsyth, C. M., & Lehman, B. A. (2017). Two heads may be better than one: learning from computer agents in conversational trialogues. *Teachers College Record*, *119*(3), 1-20.

[11] Graesser, A. C., Hu, X., & McNamara, D. S. (2005). Computerized learning environments that incorporate research in discourse psychology, cognitive science, and computational linguistics. *Experimental Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer. Washington, DC: American Psychological Association*.

[12] Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science*, *23*(5), 374-380.

[13] Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in cognitive science*, *3*(2), 371-398.

[14] Greenberg, D., Graesser, A.C., Frijters, J.C., Lippert, A.M., & Talwar, A. (submitted). Using AutoTutor to track performance and engagement in a reading comprehension intervention for adult literacy students. Learning Disability Quarterly.

[15] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100-108.

[16] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651-666.

[17] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31*(3), 264-323.

[18] Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational psychologist, 38(1)*, 23-31.

[19] Levin, J., Fox, J. & Forde, D. (2010). *Elementary statistics in social research*. Boston: Allyn & Bacon Pearson.

[20] Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A. C., & Halpern, D. (2011). Operation ARIES! A Serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, & J. Lakhmi (Eds.), *Serious Games and Edutainment Applications* (pp.169-196). London: Springer-Verlag.

[21] Mills, C., Graesser, A., Risko, E. F., & D'Mello, S. K. (2017). Cognitive coupling during reading. *Journal of Experimental Psychology: General*, *146*(6), 872-908.

[22] Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, *24*(4), 427-469.

[23] Pavlik Jr, P. I., Brawner, K., Olney, A., & Mitrovic, A. (2013). Tutoring Systems. *Design Recommendations for Intelligent Tutoring Systems: Volume 1-Learner Modeling, 1*, 39. US Army Research Laboratory.

[24] Perfetti, C.A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*, 357-383.

[25] Van den Broek, P. W., White, M. J., Kendeou, P., & Carlson, S. (2009). Reading between the lines. Developmental and individual differences in cognitive processes in reading comprehension. In R. K. Wagner, C. Schatschneider & C. Phythian-Sence (Eds.), *Beyond decoding. The behavioral and biological foundations of reading comprehension* (pp. 107-123). New York, NY: The Guilford Press.

[26] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of* the American Statistical Association, 58(301), 236-244.

[27] Wise, A. F., Speer, J., Marbouti, F., & Hsiao, Y. T. (2013). Broadening the notion of participation in online discussions: examining patterns in learners' online listening behaviors. *Instructional Science*, *41*(2), 323-343.

[28] Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson tests of achievement.* Itasca, IL: Riverside Publishing.