International Journal of Artificial Intelligence in Education (2022) 32:297–322
https://doi.org/10.1007/s40593-021-00266-y

ARTICLE

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

Ying Fang[1,2] · Anne Lippert[3] · Zhiqiang Cai[4] · Su Chen[4] · Jan C. Frijters[5] · Daphne Greenberg[6] · Arthur C. Graesser[4]

Accepted: 5 July 2021 / Published online: 13 August 2021


## Abstract

A common goal of Intelligent Tutoring Systems (ITS) is to provide learning environments that adapt to the varying abilities and characteristics of users. This type of adaptivity is possible only if the ITS has information that characterizes the learning behaviors of its users and can adjust its pedagogy accordingly. This study investigated an intelligent tutoring system with computer agents (*AutoTutor*) designed to improve comprehension skills in adults with low reading literacy. One goal of this study was to classify adults into different clusters based on their behavioral patterns (accuracy and response time to answer questions) while they interacted with AutoTutor to help them improve their reading comprehension skills. A second goal was to investigate whether adults' behaviors were associated with different reading components. A third goal was to assess improvements in reading comprehension skills, based on psychometric tests, of different clusters of readers. Performance on AutoTutor was collected in a targeted 100-hour hybrid intervention for adult readers ($n$ = 252) that included both human teachers and the AutoTutor system. The adults' average accuracy and response time in AutoTutor were used to cluster the adults into four categories: higher performers (comparatively fast and accurate), conscientious readers (slow but accurate), under-engaged readers (fast at the expense of somewhat lower accuracy) and struggling readers (slow and inaccurate). Two psychometric tests of comprehension were used to assess comprehension. Gains in comprehension scores were highest for conscientious readers, lowest for struggling readers, with higher performing readers and under-engaged readers in between. The results provide guidance to enhance the adaptivity of AutoTutor.

**Keywords** Adult readers · AutoTutor · Comprehension strategy · Intelligent tutoring system · Personalized instruction

An earlier version of this paper was presented at the 11th International Conference for Educational Data Mining and was published in its proceedings (https://files.eric.ed.gov/fulltext/ED588062.pdf).

✉ Ying Fang
ying.fang07@gmail.com

Extended author information available on the last page of the article

## Introduction

Approximately one in five adults aged 16 or older in 33 OECD (Organization for Economic Cooperation and Development) countries have literacy skills at a low level of proficiency (OECD, 2016). Adults with low literacy skills are heterogeneous in characteristics, such as age, race/ethnicity, country of origin, educational level, literacy skills, interests, and goals (Elish-Piper, 2007). Population diversity makes it difficult for a single instructor to optimize learning in groups or classrooms of students, even when there are attempts to differentiate instruction among subgroups. A computer program, on the other hand, can offer personalized tutoring and adapt instruction to the individual learner based on the learner's responses to tasks (Fletcher, 2003; Graesser et al., 2017a; Woolf, 2009). In the United States, 1200 federally funded adult literacy programs were surveyed between 2001 and 2002, and results indicated that 80% of the programs used computers in some capacity with adult learners (Tamassia et al., 2007). The growth in computer usage makes it likely this number is even higher today. As part of the effort to increase computer-based instruction, an intelligent tutoring system (ITS), AutoTutor, was developed to help adult learners improve reading comprehension skills (Graesser et al., 2016, 2019).

The present study has three goals: (a) identify clusters of adult readers with low literacy skills who exhibited particular behavior profiles while using AutoTutor, (b) investigate whether adults' behaviors are associated with different reading components represented by different theoretical levels of reading comprehension, and (c) assess the extent to which each cluster of readers shows improvements in two psychometric measures of comprehension skill. We analyzed adult readers' performance data collected online in AutoTutor, which was part of a reading comprehension intervention where students learned through a combination of teacher-led and AutoTutor sessions. We conducted a clustering analysis that classified the adults on the basis of performance patterns, namely the accuracy and time to answer questions asked by AutoTutor. We next explored how the adult clusters' behaviors varied across different theoretical levels of reading comprehension. We also examined the association between these clusters of adults and gains on psychometrically validated comprehension tests that were administered before and after the intervention. The results of these analyses were expected to inform next steps in improving the adaptivity of AutoTutor for use by adult readers.

### AutoTutor

AutoTutor is a conversation-based intelligent tutoring system (ITS) that has promoted learning on a wide range of topics such as reading comprehension, computer literacy, physics and critical thinking in science (Graesser, 2016; Nye et al., 2014). The system has shown learning gains of 0.4 to 0.8 standard deviation units on average across topics compared to more traditional methods of learning and teaching (Graesser, 2016; Nye et al., 2014; VanLehn et al., 2007). Most of the AutoTutor systems implement *dialogue* conversations that model interactions occuring between a single human tutor and human student. More recent versions of AutoTutor often employ *trialogues,* which are tutorial conversations between three actors: a teacher agent, a peer agent, and the human student (Graesser et al., 2017b). Trialogues offer several affordances over

dialogues. For example, in a trialogue setting, the human student can observe productive interactions between the two agents and mimic this behavior. The peer agent may also express misconceptions that the human students often share. When these misconceptions of the agent are expressed, the tutor agent directs undesirable feedback to the peer agent instead of the human student. This serves to minimize the amount of negative feedback human students receive which could potentially lower their self-esteem. Struggling adult readers' self-esteem can also be bolstered in game-like scenarios that trialogues afford. These games or competitions occur between the human student and the peer agent and are programmed in ways that ensure the human student never loses.

Agent trialogues are implemented in *AutoTutor* for CSAL, an ITS developed in the Center for the Study of Adult Literacy (CSAL, Graesser et al., 2016; Graesser et al., 2019). This 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 implements trialogues in which two computer agents (a teacher agent and a peer agent) have conversations with adult learners and between themselves. The three-way conversations are designed to (a) provide instruction on reading comprehension strategies, (b) help the adults apply these strategies to particular texts and sentences, (c) assess the adults' performance on applying these strategies, and (d) guide the adults in using the computer.

The lessons in AutoTutor typically start with a 2–3 min video that reviews a comprehension strategy. After the review, the computer agents scaffold adult learning by (a) asking questions woven into the conversation about texts, sentence, words, or images, (b) providing short feedback, (c) explaining how the answers are right or wrong, and (d) providing correct responses to questions. Figure 1 is an example of a "game mode" lesson in AutoTutor where the human student and peer agent compete to earn points by correctly answering questions about lesson material. The teacher agent (on the left) is asking both the adult and the peer agent (on the right) to find out the correct affix for the word "check" in the given context. The scores of both the student and peer agent are shown under their names. The student chooses the answer by clicking whereas the peer agent gives his answer by talking.

(1) Cristina (Teacher Agent): Can you use the context of the sentence to figure out the correct affix? Sam, do you have an answer for this question?
(2) Sam (Human Student): [Click the answer "ing"]
(3) Cristina: Jordan, do you know which is the right answer?
(4) Jordan (Computer Peer Agent): I am not sure. I think the correct affix for this one is "able".
(5) Cristina: Jordan, your answer is incorrect. Sam, exactly, you are right. In this case, the correct affix is "ing".
(6) System: [Sam is correct and is given a point, Jordan is wrong and is awarded no points.]

AutoTutor lessons can be viewed on the general AutoTutor web site on multiple applications (sites.autotutor.org) and on the web site focusing on AutoTutor-ARC (Adult Reading Comprehension) for adult literacy instructors (adulted.autotutor.org).

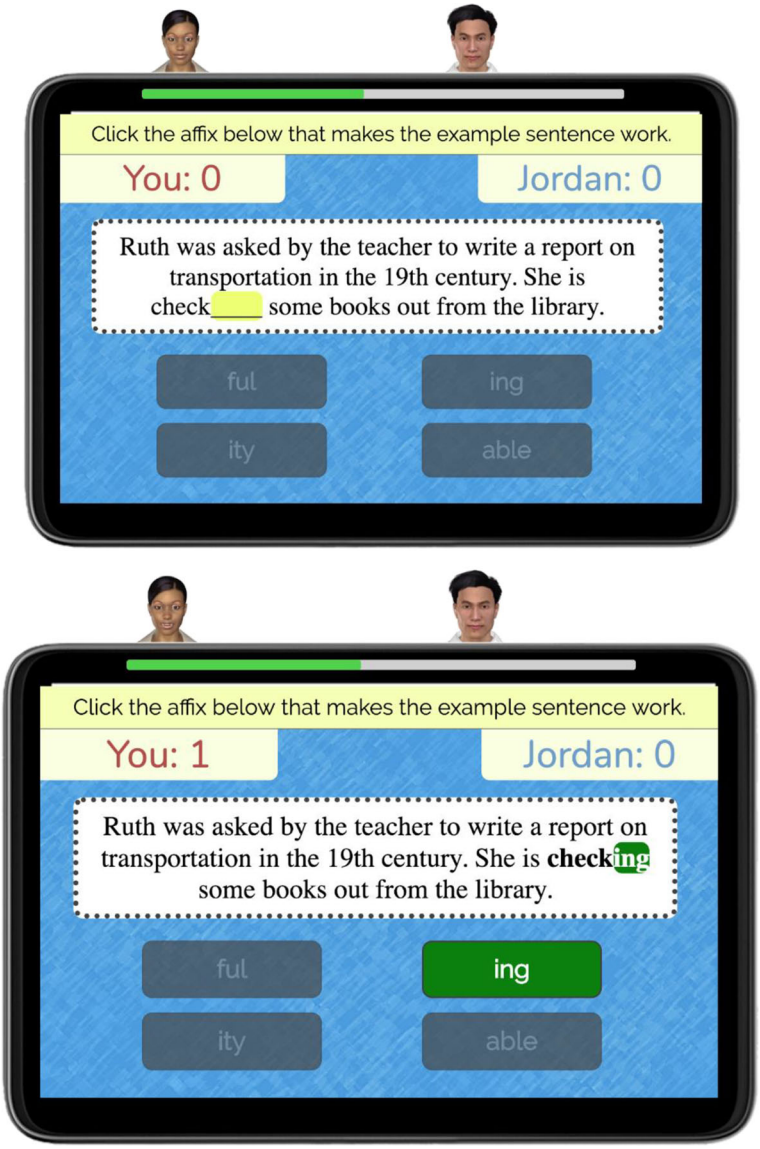


**Fig. 1** Example trialogue with competition that focuses on the affix of words in context

## Theoretical Framework of Comprehension

The 30 lessons within AutoTutor align with Graesser and McNamara's (2011) multi-level theoretical framework of comprehension. Six levels of comprehension were identified in this framework: *word, syntax*, the explicit *textbase*, the referential *situation model*, the *genre/rhetorical structure*, and the *pragmatic communication* level. The *word* and *syntax* levels represent the basic reading components that include morphology, word decoding, and syntax (Perfetti, 2007; Rayner et al., 2001; Sabatini et al., 2019). The *textbase* level focuses on the explicit ideas in the text, but not the precise

wording and syntax. These explicit idea units are often connected by co-reference (Halliday & Hasan, 1976; Kintsch, 1998). That is, referential cohesion occurs when a noun, pronoun or a noun-phrase refers to another constituent in the text. A referential cohesion gap occurs when the words in a sentence or clause do not connect to other sentences in the text. These cohesion gaps at the textbase level increase reading time (Haberlandt & Graesser, 1985; Just & Carpenter, 1987) and run the risk of disrupting comprehension (Kintsch, 1998).

The situation model refers to the subject matter content described in the text, including inferences activated by the explicit text (Kintsch, 1998; Zwaan & Radvansky, 1998). The situation model depends on the text genre. In narrative text, for example, the situation model includes the characters, objects, settings, goals, actions, and events that unfold over time in the plot. In informational text, such as an encyclopedia article, the situation model corresponds to the substantive subject matter described or explained. Zwaan and Radvansky (1998) proposed five dimensions of situation model, namely, causation, intentionality, time, space and people. The cohesion of situation model decreases when there is a discontinuity on one or more dimension (Zwaan et al., 1995; Zwaan & Radvansky, 1998). Cohesion breaks result in increased reading time and sometimes disrupt comprehension (O'brien et al., 1998; Rapp et al., 2007; Zwaan & Radvansky, 1998).

The *genre and rhetorical structure* focus on the category of text and its composition. Genre refers to category of text such as narration, exposition, persuasion, description, as well as the subcategories of these genres. The rhetorical structure provides the functional organization of paragraphs that involve various rhetorical frames, such as compare–contrast, cause–effect, claim–evidence, and problem–solution (Meyer et al., 2010). Readers without sufficient mastery of genres and rhetorical structures have more difficulties reading, comprehending, and recalling texts (Deane et al., 2006; Eason et al., 2012; Meyer et al., 2010; Williams et al., 2009).

The *Pragmatic communication* involves context-sensitive exchanges between speaker and listener, or writer and reader. Just as a speaker in a conversation has a purpose of conveying a message to the listener (Clark, 1996), the writer tries to convey a message to the reader (Rouet, 2006). Designers of AutoTutor did not directly implement the pragmatic communication level because this level is highly constrained by contexts that are difficult to stage reliably in the computer environment.

The *word* and *syntax* levels correspond to the lower-level basic reading components, whereas the *textbase*, *situation model* and *rhetorical structure* levels cover higher-level semantic and discourse components that presumably are more difficult to process (Millis et al., 2019). According to some theories, lack of mastery of basic components will have negative repercussions on deeper comprehension (Cain, 2010; Van den Broek et al., 2009; Vaughn et al., 2008).

## Approaches to Modeling Learners

The architecture of intelligent tutoring systems (ITS) has four major components (Graesser et al., 2017a; Woolf, 2009): the domain model, the student model, the pedagogical model, and the user interface. The domain model contains the knowledge, skills and strategies being tutored; it includes ideal expert knowledge as well as anticipated errors and misconceptions of students. The student model contains the

cognitive, motivational, affective and psychological states of students that are captured from their behaviors and performance during learning. The pedagogical model uses information from the domain and student models to select learning materials, tutoring strategies, actions and steps on what the tutor should do next. The user interface produces output in various media (e.g., texts, pictures, speech, sounds, animations, agents) after interpreting students' contributions through input media (e.g., typing, clicking, speech).

The four ITS components do not function in isolation. The student model enables the ITS to be adaptive by guiding the pedagogical model to select strategies, actions, and steps that are sensitive to the student's knowledge, skills, and abilities. VanLehn (2006) proposed two ways intelligent tutoring systems are configured to adapt to student model: the outer loop and the inner loop. The outer loop has macro-adaptation whereas the inner loop has micro-adaptation. The outer loop is the problem selection loop that decides what problem (e.g., task, main question, item) to present next for the student to work on. Problems are selected to fill gaps between the student model and the domain model. The inner loop refers to the tutoring actions and steps within a problem that are sensitive to the student performance. In AutoTutor, the outer loop consists of selecting the next sentence or text to work on and asking the adult a question about it. For example, in the first turn in Fig. 1, Cristina asks: "Can you use the context of the sentence to figure out the correct affix?" and the sentence is displayed on the screen. In contrast, the inner loop consists of the individual speech acts and feedback of Cristina in the conversation turns 1 through 6 in Fig. 1. For example, Cristina makes a request in turn 1 (Sam, do you have an answer for this question?) and gives feedback in turn 5 (Sam, exactly, you are right. In this case the correct affix is "ing."). The system interface also gives feedback with highlighting, color and points scored.

Data mining approaches are often used to compute the knowledge, skills, and abilities of individual students that get stored in the student model. For example, the ITS *Cognitive Tutor* employed a classifier to detect "gaming-the-system" behavior. This behavior occurs when students intentionally misuse features of an ITS (such as quickly asking for hints and help until the computer solves the problem) in order to progress through the content without learning or thinking about the material (Baker et al., 2008).

We conducted data mining in the present study by analyzing the log files of the AutoTutor student model. The student model data focused on the accuracy and time to answer questions during the conversation-based interaction with AutoTutor. Prior studies have suggested that the relationship between time/duration and performance is non-linear (Chounta & Carvalho, 2019; Daniel & Broida, 2004), and that time plus accuracy during training is found to provide a better classifier than time alone in predicting learning outcome measures (Carvalho et al., 2018). Therefore, we took both time and accuracy into consideration when we clustered the students. The accuracy and response time were recorded by AutoTutor for each question asked by Cristina to be answered by the adult student. The lessons and questions were categorized by whether they targeted the word, textbase, situation model, or genre/rhetorical structure levels of the multilevel theoretical framework. Note that we did not analyze syntax because there was only one lesson that focused on syntax. The two performance measures (accuracy and response time) for each of the four theoretical levels provided the data for clustering students.

### Reading Instruction in Current Study

The current study analyzed data from a reading intervention in which adult literacy students participated in blended classes consisting of two types of instruction: conventional teacher-led classroom instruction and computer-based AutoTutor instruction. The teacher-led instruction covered both basic level reading skills and deeper level comprehension skills. The basic level skills included morphology, word decoding, vocabulary, and syntax (Perfetti, 2007; Rayner et al., 2001). The instruction for basic level reading was called Phonological and Strategy Training (PHAST), which was a remedial reading program to help disabled readers achieve independent reading skills at the basic level. The focus of the program was on teaching word identification and independent decoding strategies (Lovett et al., 2000). The instruction for semantic and other deeper levels of comprehension was aligned with an instructional curriculum that was successful in helping struggling readers in middle and high school (Lovett et al., 2012). More specifically, there were both teacher-led and AutoTutor instruction on each of the five strategies that make up the curriculum named PACES: (1) **P**redicting topic and writer's purpose with text signals and key information, (2) **A**cquiring new vocabulary with context clues, (3) **C**larifying common sources of confusion about the text with clarifying questions, (4) **E**valuating, elaborating, and explaining through questioning, and (5) **S**ummarizing, identifying and constructing text structures. The PACES curriculum covered the word, textbase, situation model, and genre & rhetorical structure levels of the multilevel theoretical framework. The word level lessons in the PACES curriculum were different from a focus on basic level training; instead, they aimed to teach students how link the basic level components to deeper semantic comprehension. The AutoTutor sessions were mainly aligned with the scope and sequence of PACES, with modest variations to tailor the curriculum to the adult readers. Table 1 shows AutoTutor lessons and their alignment with PACES and the theoretical levels. All participants were also assessed before and after the instruction on standardized tests of comprehension that are described in the Methods section. These pretest and posttest scores are reported in order to see whether different clusters of readers obtained different posttest results.

## Method

### Participants

The blended intervention with teacher-led sessions and AutoTutor included 252 adults (188 female, 64 male) who were recruited from adult literacy classes in Atlanta ($n$ = 134) and Toronto ($n$ = 118). All participants were selected for the intervention if they read between the 3rd and 8th grade levels as determined by their literacy program. Participants ranged in age from 18 to 74, with a mean of 42.4 years (SD = 13.9). African American was the largest ethnic group (59.5%), followed by multiracial (17.5%), white (11.5%), and Asian (9.5%). The majority of the participants were native English speakers in both Atlanta (59.0%) and Toronto (50.8%) literacy classes.

**Table 1** AutoTutor Lessons and Alignment of Theoretical Levels and PACES Curriculum

| AutoTutor Lesson | Theoretical Level | PACES intervention |
|---|---|---|
| Digital Literacy Orientation | | |
| Text Signal | SM | P |
| Purpose of Texts | RS | P |
| Complex Texts | RS,SM | P |
| Word Parts | W | A |
| Punctuation | TB,SM | A |
| Word Meaning Clues | W | A |
| Learning New Words | W | A |
| Multiple Meaning Words | W, TB | C |
| Pronouns | TB,W | C |
| Non-Literal Language | SM | C |
| Review 1 | SM,W | P-A-C |
| Key Information | TB,SM | E |
| Main Ideas | TB, RS | E |
| Connecting Ideas | SM,TB, RS | E |
| Story Maps | SM,RS | S |
| Stories 1 | SM,TB,RS | E,S |
| Persuasion 1 | TB,RS | E |
| Review 2 | SM,TB, RS | P-A-C-E |
| Claims versus Support | RS,SM | S |
| Problems and Solutions | RS,TB,SM | S |
| Cause and Effect | RS,TB,SM | S |
| Describe Things | RS,TB,SM | S |
| Compare and Contrast | RS,TB,SM | S |
| Time and Order | RS,TB,SM | S |
| Steps in Procedures | RS,TB,SM | S |
| Review 3 | RS,TB,SM | P-A-C-E-S |
| Stories 2 | SM,TB | E |
| Inferences from Texts | SM,TB | E |
| Persuasion 2 | SM,TB | E |
| Forms and Documents | SM,TB | E |

*Note:* W = word, TB = textbase, SM = situation model, RS = rhetorical structure;
P = predicting, A = acquiring, C = clarifying, E = evaluating, S = summarizing

## Measures of Reading Comprehension

Reading skills assessments were administered one-on-one to participants in quiet environments at their adult literacy centers over multiple sessions. In total, adults were assessed on 37 measures that tested four categories of knowledge, skills, ability, and other psychological characteristics: (1) *basic reading* skills, including phonology, morphology, decoding, vocabulary, and fluency; (2) *comprehension and cognitive*

*skills and abilities,* such as reading comprehension, general knowledge, reasoning, oral and written communication, short-, working- and long-term memory; (3) *motivation,* such as self-reported motivation for-, breadth of-, and depth-of-reading, intrinsic motivation toward reading, expectancy for reading success, and prior experiences with reading; and (4) *self-report data*, including student demographics, computer familiarity, and frequency of reading different types of print. In this study, we focused on two measures from the comprehension category that were collected both before intervention (pretest) and after intervention (posttest): Woodcock-Johnson III Passage Comprehension subtest (Woodcock et al., 2007) and the Reading Assessment for Prescriptive Instructional Data (RAPID) Passage Comprehension subtest developed by Lexia Learning (Foorman et al., 2017). Each battery consists of multiple subtests. For this study, we focused on subtests that measure overall reading comprehension skill. RAPID was added after an initial wave of data collection so the number of participants who completed it is smaller than WJ-III.

The Passage Comprehension subtest of the Woodcock-Johnson III was administered by a human tester. The items were texts with one or two sentences with missing words indicated by blanks. The participant silently read each item and filled in the blank by speaking the missing word out loud. Items were administered by completing pages in the testing booklet until the participant provided incorrect responses to six consecutive items. Participants' performance on Woodcock-Johnson III was analyzed using the raw scores (ranging from 0 to 47).

The Reading Comprehension subtest of the RAPID was a web-based test administered on a computer. All items had a multiple-choice format and participants selected answers by mouse clicks. The participant saw passages ranging from approximately 200 to 1300 words. With the passage still in view, the participant then answered questions by selecting one out of four choices for each question. The performance scores (ranging from 0 to 1000), which provided an estimate of a student's development in reading comprehension, was analyzed in this study as an outcome measure.

### Measures Used for Clustering Analysis

When students interacted with AutoTutor, the computer recorded both the time to answer the question and accuracy of the response to each question. Time was measured from the onset of the question (i.e., the question was shown on the computer screen) to the click on an option indicating the participant's answer. Accuracy was measured as the answer being correct or incorrect. Time and accuracy are referred to as performance measures. The lessons and questions were categorized by whether they targeted the word, textbase, situation model, or genre/rhetorical structure levels of the multilevel theoretical framework. The averaged performance measures (time and accuracy) for each of the four theoretical levels of a student were the inputs for the cluster analysis.

### AutoTutor Lessons

There were 29 AutoTutor lessons assigned to the adult participants in the intervention study. Many of the lessons had two texts with several sentences and 10–12 multiple-choice questions associated with each text. Other lessons focused on sentences or words, with 10–35 questions per lesson. There were two types of lessons in AutoTutor:

(1) lessons with a fixed sequence of questions where every student had the opportunity to answer all the questions, and (2) lessons adaptive to the performance of the adult students based on the early phase of the lesson. For example, there were lessons that started with medium-level difficulty texts and corresponding questions. Halfway through the lesson, the system would branch to easier or more difficult texts depending on the adult's proportion of correct responses for the initial text of medium difficulty. When an adult's proportion of correct responses met or exceeded a threshold (i.e., 0.67 in most lessons, where 0.33 was chance performance), the student would be assigned the more difficult text or reading material (i.e., "the difficult path"). Otherwise, the student received an easier text (i.e., "the easy path").

The difficulty of texts was scaled on objective measures computed by Coh-Metrix (Graesser et al., 2014). Coh-Metrix is a system that scales texts on difficulty with respect to multiple levels of language and discourse by analyzing characteristics of words, syntax, discourse cohesion and text genre. There are five major dimensions that Coh-Metrix uses to scale a text: word concreteness, syntactic simplicity, referential cohesion, deep cohesion, and narrativity. The dimensions are aligned with the word, syntax, textbase, situation model, and genre/rhetorical structure levels of the multilevel theoretical framework (Graesser & McNamara, 2011). A composite measure called *formality* is based on the five dimensions and was used as a single approximate index of text difficulty. Graesser et al. (2014) reported that text formality score from Coh-Metrix correlated highly (0.66 to 0.72) with other standard metrics of text difficulty, such as Flesch-Kincaid grade levels scores (Klare, 1974) and Lexile scores (Stenner, 1996). Coh-Metrix formality score was used as the measure of text difficulty to scale the texts; the difficulty level (i.e., easy, medium, difficult) of the texts within each lesson was coded based on the formality scores.

In some lessons, there is a fixed sequence of outer-loop questions that all students receive. However, in most of the lessons, the set of outer-loop AutoTutor questions is adaptive to the performance of the adults in the early phase of a lesson. Consequently, students receive a somewhat different sample of texts, sentences and questions in the later phase of these adaptive lessons. For these adaptive lessons, all adults start out on materials that are medium in difficulty. These items were the focus in the present study when we classified the participants in clustering analyses. Specifically, we limited our analysis to the items in the lessons with fixed sequences of questions, and items associated with the medium-level texts. We restricted our clustering and follow-up analyses to these constant items, because all the participants who were present for the lesson would be certain to receive these questions during AutoTutor sessions. The contingent items in the later phases were excluded because not every adult had the opportunity to work on them.

## Design and Procedure

The data covered three waves of an intervention that had modest changes between data collection cycles in order to make adjustments in the intervention. Modifications are customary in interventions and cycles were annotated in data analyses. The interventions covered the time span of January 2015 to December 2016, and each intervention lasted for approximately 4 months. All waves followed a similar procedure. Before beginning the intervention, participants took pretests that measured their prior reading

skills as well as other cognitive and psychological characteristics (see Section 2.2). During the intervention, participants attended adult literacy classes that consisted of both a teacher-led classroom component and an AutoTutor component. Teachers selected an assignment for each AutoTutor session which specified the lesson to complete on a specific day. The AutoTutor lesson was aligned with the human-led PACES component on most days, but sometimes the group ran out of time, so AutoTutor was completed at the beginning of the next session. Session duration varied from 1.5 to 3 h, with 2–3 sessions per week. A typical session was composed of approximately 25% AutoTutor interactions, with the remainder human-led instruction.

After the 4-month intervention, the participants completed posttests that assessed their reading comprehension skills. In all three waves, Woodcock-Johnson (WJ) III Passage Comprehension measures were used in the pretest and posttest. In waves 2 and 3, the RAPID Comprehension subtest was used together with Woodcock-Johnson (WJ) Passage Comprehension III measures to evaluate students' reading comprehension skills.

### Data Coding and Data Preprocessing

Performance measures on the 29 AutoTutor lessons were collected. To explore how learning behaviors varied by theoretical levels, we first coded each lesson in terms of the one or more theoretical levels it covered. Each lesson tapped 1–3 of the four theoretical levels (i.e., word, textbase, situation model, and rhetorical structure). A group of researchers collaboratively constructed a Q-matrix that assigned a measure of the relevance of each of the four theoretical components to each of the lessons when the AutoTutor system was being designed. The relevance of a theoretical level to a lesson was the extent to which the level was tapped in the lesson. The expert-assigned codes were primary, secondary, tertiary or no relevance of a component to a lesson. In the analyses reported in this article, the theoretical levels of the lessons were based on their primary theoretical level.

The AutoTutor data was a log file that included 252 adults' learning records. We removed the incomplete and non-discriminating items (i.e., items that all the students answered correctly). We also removed 3% of the observations with times that are more than 3 times the interquartile range below the first quartile or above the third quartile. The log file ended up having 42,288 observations, with each observation consisting of an attempt a student made at answering a question. All students attempted multiple lessons. Within each lesson there were 10–35 questions, so each student had multiple observations in the log file. As described, each lesson was coded with a specific theoretical level and each question within a lesson received the same coding. For example, we coded questions as *textbase* if they had the question and alternative answers directed at textbase considerations. It was important to examine how adult performance varied with theoretical level, so we aggregated the data and calculated each student's time and accuracy for each theoretical level when averaged across lessons. As a consequence of this aggregation, the observations for each adult consisted of eight values, namely the average time and accuracy at the word, textbase, situation model and rhetorical structure levels. Time was measured in seconds, which was a continuous variable. Accuracy was measured as being correct or incorrect, which was a binary variable (i.e., 1 or 0). After aggregation, accuracy represented proportion of

correct answers, which was a continuous variable. The cluster analysis was based on this aggregated data for each subject.

## Data Analysis

### Cluster Analysis

We performed clustering analyses to address our first goal. Namely, we wanted to know whether AutoTutor data would reveal distinct behavioral patterns in adults with low literacy skills. Cluster analysis partitions objects into clusters so that the objects in the same cluster are more similar to each other than to those in other clusters. We used an R package clValid (Brock et al., 2011) to compare the solutions based on k-means clustering algorithm with those solutions based on a hierarchical clustering algorithm using Ward's method (Ward Jr, 1963). We computed the scores of different solutions on three measures, namely connectivity, Silhouette Width, and Dunn Index. Connectivity measures the degree of connectedness of the clusters. Silhouette Width and the Dunn Index measure the compactness and separation of the clusters.

### Mixed-Effects Modelling

Our second goal was to determine whether adult readers' learning behaviors are associated with different reading comprehension levels. To this end, we performed linear mixed-effects regressions to analyze the effect of clusters and theoretical level on both time per question and proportion correct scores using the lme4 package in R (Bates et al., 2014). We added clusters into the linear mixed models as an independent variable to assess its association with time per question and proportion correct scores for confirmatory purposes. In both models, we specified subjects as a random factor to adjust for the subject variance.

### General Linear Model

We conducted general linear regression analyses to address the third goal, which is to examine whether the learning gains vary for different adult clusters. More specifically, we tried to predict the posttest scores as a function of pretest scores and adult clusters. We performed this procedure on the two reading comprehension measures: Woodcock Johnson III Passage Comprehension subtest scores and the RAPID Reading Comprehension subtest scores, separately.

# Results

## Time and Accuracy at Four Theoretical Levels

Table 2 shows the means and standard deviations of time and accuracy for each of the four theoretical levels. We conducted a mixed-effects linear regression analysis to predict response time as a function of theoretical level to test the association between

**Table 2** Means and standard deviations of response time in seconds and accuracy (proportion correct) at four theoretical levels

| | Time Mean (SD) | Accuracy Mean (SD) |
|---|---|---|
| Word | 33.8 (12.6) | 0.67 (0.16) |
| Textbase | 35.8 (10.9) | 0.68 (0.16) |
| Situation Model | 30.8 (9.21) | 0.69 (0.11) |
| Rhetorical Structure | 32.0 (10.0) | 0.70 (0.10) |

response time and theoretical levels. The subjects were specified as a random-effect factor to adjust for the subject variance. The fixed effect of theoretical level was statistically significant ($F$ (3, 666) = 19.56, $p < .001$), indicating that time varied across the four levels. The average response time for questions at textbase level was longer than word ($t$ (669) = 2.51, $p < .01$), situation model ($t$ (660) = 6.92, $p < .001$), and rhetorical structure ($t$ (661) = 5.24, $p < .001$). The average time adults spend on answering word level questions was longer than that of situation model ($t$ (671) = 4.46, $p < .001$) and rhetorical structure questions ($t$ (672) = 3.19, $p < .01$). Similarly, we performed a mixed-effects linear regression to predict accuracy as the function of theoretical level to test the association between accuracy and theoretical levels. The results were statistically significant for the fixed effect of theoretical level ($F$ (3, 673) = 2.88, $p = .04$). The average accuracy on questions at the rhetorical structure level was higher than word ($t$ (681) = 2.57, $p = .01$) and textbase ($t$ (667) = 2.30, $p = .02$). The accuracy differences between other theoretical levels were not statistically significant. Since the differences found in response time and accuracy were small and did not show any meaningful patterns, we decided to group the adults through clustering to investigate whether theoretical levels influenced adults in a more nuanced way.

### Four Clusters with Distinct Patterns

For the number of clusters, we started with k = 4 guided by previous research. That is, a recent study using a smaller dataset from the first wave of the same intervention reported four types of performance profiles for the items: correctly and quickly completed, incorrectly and quickly completed, correctly and slowly completed, and incorrectly and slowly completed (Graesser et al., 2018). Our assumption is that there are students whose dominant behavioral pattern falls into the four categories. Therefore, we started with k = 4. We also experimented with k = 3 and k = 5 and made comparisons between these solutions. Compared to the 4-cluster solution, the 3-cluster solution combined two distinctive clusters (i.e., under-engaged readers and struggling readers) into one cluster and lost some meaningful information about time and accuracy. In the 5-cluster solution, Cluster 1 (i.e., higher performers) was further split into two clusters. One cluster was the most accurate among all the clusters; their mean response time was longer than the other cluster but was still short compared to the rest of the clusters. The other cluster spent the least amount of time answering questions, and their accuracy was the second highest across the theoretical levels. These two clusters were both fast and accurate compared to the other three clusters, so we decided not to split them.

Therefore, the 4-cluster solution was selected as the (locally) optimum solution. The 4-cluster, 3-cluster and 5-cluster solutions are shown in Tables 3, 4 and 5, respectively.

Next, we compared 4-cluster solutions using k-means clustering with hierarchical clustering. Hierarchical clustering outperformed k-means clustering on connectivity, Silhouette Width, and Dunn Index. Therefore, the final solution we selected was the 4-cluster solution based on the hierarchical clustering algorithm. The means and standard deviations of each cluster on the eight clustering variables (the labels of the clusters are introduced in section 3.2), and the demographics of the four clusters are shown in Table 3 and Table 6, respectively. Overall, there is no statistically significant difference between clusters regarding age, gender, ethnicity, or first language.

We applied linear mixed-effects models to compare accuracy and time across clusters at different theoretical levels. In both models, the predictors were cluster, theoretical level and their interaction. Subjects were specified as a random-effect factor to adjust for the subject variance. For proportion correct scores, there was a statistically significant interaction between cluster and theoretical level, $F(9, 65) = 4.04$, $p < 0.001$. For time per question, there also was a statistically significant interaction between cluster and theoretical level, $F(9, 651) = 11.41$, $p < 0.001$. Given these interactions, we will discuss the patterns of each cluster separately. The average time per question and proportion correct for the four clusters are shown in Figs. 2 and 3, respectively. The pairwise contrasts (i.e., multiple comparison tests adjusted by Tukey's method) between clusters at each categorical level are shown in Table 7.

### Cluster 1: Higher Performers

Cluster 1 is the largest cluster with 39% ($n = 97$) of the study sample. These adults can be distinguished by their comparatively short response times and higher accuracy. The response time of Cluster 1 was shorter than the other three clusters for situation model questions. At the other three theoretical levels, there was no significant difference between the response time of Cluster 1 and Cluster 3, whereas Cluster 1 was faster

**Table 3** Means and standard deviations of response time in seconds and proportion correct for four clusters

| | Cluster 1 (n=97) (Higher Performer) M (SD) | Cluster 2 ($n$=31) (Conscientious Reader) M (SD) | Cluster 3 ($n$=93) (Under-engaged Reader) M (SD) | Cluster 4 ($n$=31) (Struggling Reader) M (SD) |
|---|---|---|---|---|
| Time (W) | 31.46 (10.52) | 38.16 (9.59) | 30.94 (10.58) | 45.65 (17.95) |
| Time (T) | 31.63 (7.90) | 54.79 (9.54) | 32.85 (7.47) | 38.61 (8.52) |
| Time (SM) | 26.04 (5.70) | 42.98 (7.93) | 28.90 (7.17) | 39.18 (8.67) |
| Time (RS) | 27.65 (7.36) | 45.92 (10.83) | 29.44 (6.89) | 39.41 (7.98) |
| Accuracy (W) | 0.77 (0.11) | 0.69 (0.13) | 0.61 (0.15) | 0.52 (0.17) |
| Accuracy (T) | 0.75 (0.13) | 0.73 (0.16) | 0.65 (0.14) | 0.50 (0.13) |
| Accuracy (SM) | 0.77 (0.08) | 0.65 (0.06) | 0.66 (0.10) | 0.58 (0.10) |
| Accuracy (RS) | 0.76 (0.07) | 0.72 (0.11) | 0.67 (0.08) | 0.61 (0.10) |

*Note.* W = word, T = textbase, SM = situation model, RS = rhetorical structure

**Table 4** Means and standard deviations of response time in seconds and proportion correct for three clusters

| | Cluster 1 M (SD) (n=97) | Cluster 2 M (SD) (*n*=62) | Cluster 3 M (SD) (n=93) |
|---|---|---|---|
| Time (W) | 31.46 (10.52) | 41.90 (14.76) | 30.94 (10.58) |
| Time (T) | 31.63 (7.90) | 46.70 (12.12) | 32.85 (7.47) |
| Time (SM) | 26.04 (5.70) | 41.08 (8.46) | 28.90 (7.17) |
| Time (RS) | 27.65 (7.36) | 42.66 (9.99) | 29.44 (6.89) |
| Accuracy (W) | 0.77 (0.11) | 0.61 (0.17) | 0.61 (0.15) |
| Accuracy (T) | 0.75 (0.13) | 0.61 (0.19) | 0.65 (0.14) |
| Accuracy (SM) | 0.77 (0.08) | 0.62 (0.09) | 0.66 (0.10) |
| Accuracy (RS) | 0.76 (0.07) | 0.67 (0.12) | 0.67 (0.08) |

*Note.* W = word, T = textbase, SM = situation model, RS = rhetorical structure

than Cluster 2 and Cluster 4. Meanwhile, Cluster 1 achieved the highest proportion correct scores across all theoretical levels. Because of the students' high accuracy and short response time, we named this cluster "higher performers." The accuracy of higher performers did not seem to be affected by theoretical level, since they did equally well in lessons across different levels.

### Cluster 2: Conscientious Readers

Cluster 2 had 12% of the study sample ($n = 31$). The adults in Cluster 2 worked slowly and they achieved comparatively high accuracy. The response times of Cluster 2 were the longest among the four clusters for three of the four comprehension levels. Contrary to struggling readers who also worked slowly, Cluster 2 had the second highest accuracy. We named this cluster "conscientious readers" because they put in the most effort and achieved comparatively high accuracy. Similar to struggling readers, the conscientious readers' performance was associated with theoretical level. The results of

**Table 5** Means and standard deviations of response time in seconds and proportion correct for five clusters

| | Cluster 1 M (SD) (*n*=60) | Cluster 2 M (SD) (n=31) | Cluster 3 M (SD) (n=93) | Cluster 4 M (SD) (n=31) | Cluster 5 M (SD) (*n*=37) |
|---|---|---|---|---|---|
| Time (W) | 27.70 (7.91) | 38.16 (9.59) | 30.94 (10.58) | 45.65 (17.95) | 37.55 (11.43) |
| Time (T) | 27.85 (6.08) | 54.79 (9.54) | 32.85 (7.47) | 38.81 (8.79) | 37.76 (6.58) |
| Time (SM) | 23.67 (4.12) | 42.98 (7.93) | 28.90 (7.17) | 39.46 (8.00) | 29.89 (5.85) |
| Time (RS) | 24.39 (5.86) | 45.92 (10.83) | 29.44 (6.89) | 39.65 (7.98) | 32.93(6.46) |
| Accuracy (W) | 0.72 (0.10) | 0.69 (0.13) | 0.61 (0.15) | 0.52 (0.17) | 0.85 (0.08) |
| Accuracy (T) | 0.70 (0.12) | 0.73 (0.16) | 0.65 (0.14) | 0.49 (0.13) | 0.82 (0.11) |
| Accuracy (SM) | 0.75 (0.09) | 0.65 (0.06) | 0.66 (0.10) | 0.58 (0.10) | 0.80 (0.07) |
| Accuracy (RS) | 0.75 (0.07) | 0.72 (0.11) | 0.67 (0.08) | 0.61 (0.10) | 0.76 (0.08) |

*Note.* W = word, T = textbase, SM = situation model, RS = rhetorical structure

**Table 6** Age, gender, race and first language status of the four clusters

| | Cluster 1 (n=97) Higher Performer | Cluster 2 (n=31) Conscientious Reader | Cluster 3 (n=93) Under-engaged Reader | Cluster 4 (n=31) Struggling Reader |
|---|---|---|---|---|
| Age (in years) | 40.20 (SD=13.40) | 45.00 (SD=13.27) | 41.87 (SD=14.08) | 48.61 (SD=13.61) |
| Gender | | | | |
| Female | 72 | 23 | 70 | 24 |
| Male | 25 | 8 | 23 | 7 |
| Race/Ethnicity | | | | |
| African American | 43 | 24 | 59 | 25 |
| White | 16 | 1 | 10 | 2 |
| Multiracial | 25 | 3 | 12 | 3 |
| Other | 13 | 3 | 12 | 1 |
| First Language | | | | |
| English | 57 | 14 | 48 | 20 |
| Non English | 40 | 17 | 45 | 11 |

mixed-effects models indicated that their performance at the textbase level was better than other levels. This result suggests they had a close reading of the explicit text.

### Cluster 3: Under-engaged Readers

Cluster 3 is another large group representing 37% ($n = 93$) of the study sample. The adults in this cluster were almost as fast as the higher performers, but their accuracy was lower than higher performers and conscientious readers. The response times of Cluster 3 were as short as higher performers at the word, textbase and rhetorical structure levels. At the situation model level, the response time of Cluster 3 was the second shortest. However, there was a large gap between the performance of Cluster 3 and Cluster 1. The adults in Cluster 1 and Cluster 3 differed in their proportion correct scores, and this difference ranged from 0.11 to 0.16 depending on the theoretical level. We named the adults in Cluster 3 "under-engaged readers" because of their short

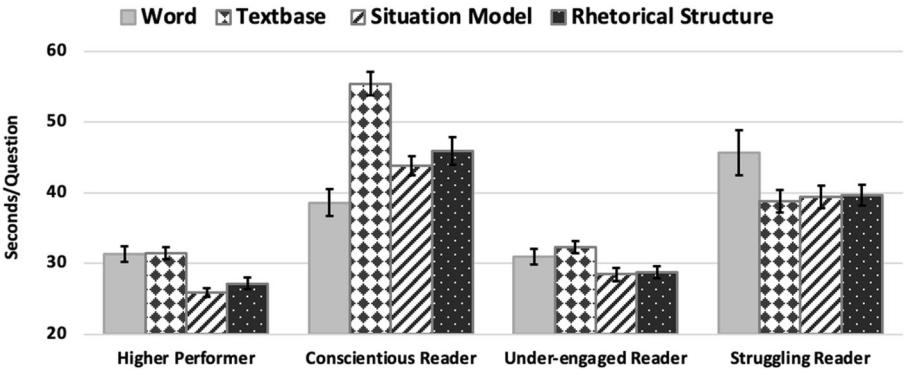


**Fig. 2** Mean and standard errors of time per question for four clusters at four theoretical levels

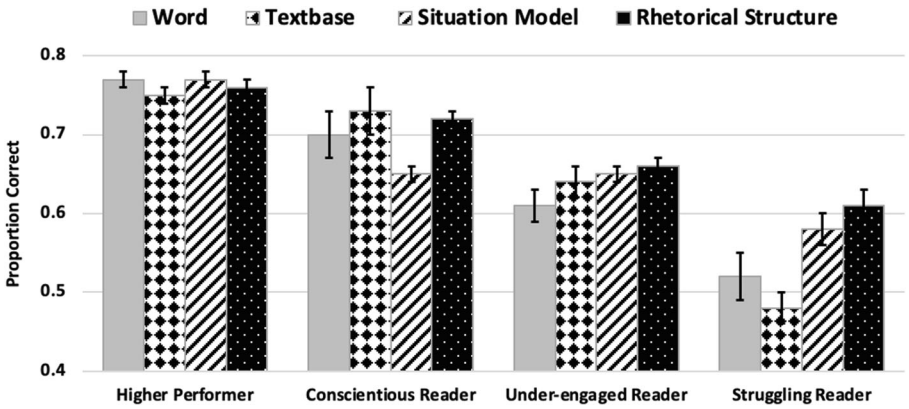


**Fig. 3** Mean and standard errors of proportion correct for four clusters at four theoretical levels

response time and comparatively poor performance. Theoretical level also affected under-engaged readers, in that they performed worse on the word level lessons compared to textbase, situational model and rhetorical structure level lessons.

### Cluster 4: Struggling Readers

Cluster 4 is a smaller cluster with 12% of the study sample ($n = 31$). The response times of the adults in this cluster were comparatively long, and their accuracy was lower than the other clusters. The response time of Cluster 4 on word level questions was the longest, but the accuracy was the lowest among the four clusters. For textbase, situation model and rhetorical structure level questions, the response time of Cluster 4 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 higher performers who had stable performance across different

**Table 7** Pairwise contrast of time in seconds and proportion correct between clusters

| | | Word | Textbase | Situation Model | Rhetorical Structure |
|---|---|---|---|---|---|
| Cluster 1-Clutser 2 | Time | −8.11* | −24.75*** | −18.77*** | −18.76*** |
| | Accuracy | 0.07** | 0.02 | 0.12*** | 0.04 |
| Cluster 1-Clutser 3 | Time | 0.30 | −1.13 | −2.88* | −1.67 |
| | Accuracy | 0.16*** | 0.11*** | 0.12*** | 0.11*** |
| Cluster 1-Clutser 4 | Time | −14.41*** | −7.01*** | −13.99*** | −12.83*** |
| | Accuracy | 0.25*** | 0.27*** | 0.19*** | 0.16*** |
| Cluster 2-Clutser 3 | Time | 8.41* | 23.62*** | 15.89*** | 17.10*** |
| | Accuracy | 0.09*** | 0.09*** | 0.00 | 0.07** |
| Cluster 2-Clutser 4 | Time | −6.29* | 17.73*** | 4.78* | 5.94* |
| | Accuracy | 0.17*** | 0.25*** | 0.07* | 0.12*** |
| Cluster 3-Clutser 4 | Time | −14.71*** | −5.89** | −11.10*** | −11.16*** |
| | Accuracy | 0.09*** | 0.16*** | 0.07** | 0.05* |

* $p < .05$, ** $p < .01$, *** $p < .001$

theoretical levels, struggling readers did better in situation model and rhetorical structure lessons than word and textbase lessons. Perhaps they were using world knowledge and experience to compensate for their deficits at the word and textbase levels.

### Reading Comprehension Skill Improvement of the Four Clusters of Readers

To compare the improvement of reading skills between the clusters, we examined their pretest and posttest performance for the two reading comprehension measures. Table 8 shows the pretest and posttest scores of the four clusters on the two reading comprehension measures. We further computed the effect sizes (Cohen's *d*) of the four clusters using the two measures, which are shown in Fig. 4.

A general linear regression analysis was conducted to predict the posttest scores of Woodcock Johnson III as a function of the pretest scores, clusters and their interaction. The assumption of homogeneity of variance was met according to Levene's Test ($p = 0.54$), and the residuals were normally distributed based on the Shapiro-Wilk test ($p = 0.40$). The results of the model indicated a statistically nonsignificant interaction between clusters and pretest scores, so we removed the interaction term from the model. The model with the two main effects was statistically significant ($F$ (4, 207) = 113.8, $p < 0.001$), with $R^2$ of 0.68. The main effect of pretest was statistically significant ($p < 0.001$). The main effect of cluster was also statistically significant ($p = 0.04$). The $R^2$ accounted for by pretest and cluster were .67 and .01, respectively. The pairwise comparison with Tukey's method indicated that higher performers performed better than struggling readers ($t$ (207) = 2.84, $p = 0.02$). The differences between other groups were not statistically significant. A similar procedure was conducted using RAPID reading comprehension subtest scores. The pretest and posttest scores were both right-skewed, so we used log-transformed data for the model. The assumption of homogeneity of variance was met according to Levene's Test ($p = 0.49$), and the residuals were normally distributed based on Shapiro-Wilk test ($p = 0.14$). The general linear model that predicts posttest scores of RAPID with pretest scores and cluster was found to be statistically significant ($F$ (4, 155) = 65.59, $p < 0.001$), with $R^2$ of 0.62. The main effect of pretest and cluster were both statistically significant ($p_{pretest} < 0.001$; $p_{cluster} = 0.01$). The $R^2$ explained by pretest and cluster were .59 and .03, respectively. The pairwise comparison using Tukey's method indicated that higher performers scored significantly better than struggling readers ($t$ (155) = 3.31, $p = 0.01$) and under-engaged readers ($t$ (155) = 2.75, $p = 0.03$) on the posttest. The contrasts among other groups were not statistically significant. We also computed the effect sizes (Cohen's *d*) of the four clusters using the two measures, which are shown in Fig. 4.

**Table 8** Means and standard deviations of the four clusters' scores in reading comprehension tests

| | Higher Performer | Conscientious Reader | Under-engaged Reader | Struggling Reader |
|---|---|---|---|---|
| WJ Pretest | 28.06 (4.20) | 23.63 (4.87) | 25.29 (3.55) | 23.43 (4.04) |
| WJ Posttest | 29.37 (4.48) | 25.78 (4.81) | 26.92 (4.14) | 24.32 (3.78) |
| RAPID Pretest | 470.60 (102.62) | 387.63 (71.23) | 401.24 (78.33) | 363.89 (74.90) |
| RAPID Posttest | 487.78 (106.14) | 411.12 (66.72) | 402.41 (70.42) | 361.26 (70.98) |

*Note.* WJ = Woodcock-Johnson III

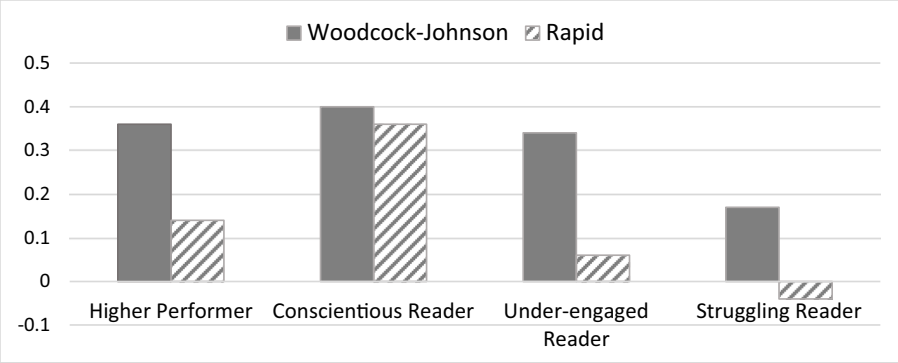


**Fig. 4** Effect sizes of four clusters with two reading comprehension tests

## Discussion

There were three major goals in the analyses. Our first goal was to classify adults' behavior patterns while they interacted with AutoTutor. Our second goal was to investigate whether adults' behaviors were associated with different reading components represented by the four theoretical levels of reading comprehension. Our third goal was to explore the extent to which the clusters showed improvement from pretest to posttest using well-established psychometric measures.

With respect to the first goal, the cluster analyses unveiled four clusters that capture distinctive behavioral patterns of adults with low literacy skills using AutoTutor. We named the four clusters higher performers, conscientious readers, under-engaged readers, and struggling readers. Higher performers worked fast and accurately. Their response time was the shortest and their accuracy was the highest across theoretical levels. On the opposite end of the spectrum were struggling readers who worked slowly and inaccurately. Their response time was either the longest or the second longest over different theoretical levels, and they had the lowest overall accuracy on questions. Conscientious readers also worked slowly, but unlike struggling readers, they were much more accurate. The response time of conscientious readers varied across the theoretical levels, but they consistently achieved comparatively high accuracy across all the theoretical levels. These adults might be sensitive to their own skill level and extended the effort they would need to master each lesson. Similar to higher performers, under-engaged readers answered questions quickly. However, they were less accurate than both high performers and conscientious readers. It appears these adults tried to get through lessons quickly without paying sufficient attention to the content. Prior research has identified various clusters with different reading competencies (Strucker et al., 2007; Talwar et al., 2020). These studies classified adults based upon their scores on reading measures such as vocabulary, word reading and spelling, which need to be collected through reading tests. In contrast, the behavioral data can be collected automatically when adults interact with AutoTutor during their learning. As such, our approach provides a way to classify adult readers without requiring them to take extra tests.

Regarding the second goal, we found learning behaviors of adults in the four clusters varied across theoretical levels in different ways. *Higher performers* performed equally

well at different theoretical levels, but they spent less time on situation model and rhetorical structure level questions than the questions at other theoretical levels. One possible explanation for the variation in time across theoretical levels is that situation model and rhetorical structure lessons have many questions associated with each text. We expected the time per question to decrease after the first few questions associated with a text, since it was during these initial questions that adults built their mental models of the subject matter being discussed in the text. Once the mental model is in place, it can be more readily accessed and lead to shorter response times. We performed a follow-up mixed-effects linear regression analysis to predict higher performers' response time as a function of question number in situation model and rhetorical structure lessons. The subjects were specified as a random-effect factor to adjust for the subject variance. The results indicated that the response time decreased with the increase of question number ($B = -.31$, $p < 0.001$), which confirmed our assumption. In word lessons, however, each question is associated with its particular single sentence or short text, so there was no cumulative effect as in the situation model and rhetorical structure lessons. Textbase lessons are similar to situation model and rhetorical structure lessons in that one text is associated with multiple questions. However, the textbase questions focus on explicit ideas in the text, so the questions are associated with details that are independent from others in different parts of a text. In such cases, learners are accessing unique information that is less dependent on an integrated global representation of the text compared with the situation model and rhetorical structure text level questions. Another possible explanation for the comparatively longer response times of higher performers at the textbase level compared to other levels is that they had a standard of comprehension to focus carefully on the explicit text (Baker, 1989; Van den Broek et al., 2009). They achieved high accuracy across all lessons, but they spent more time on textbase lessons because the questions were based on a close reading of the explicit text.

*Conscientious readers* exhibited distinctive behavior on textbase lessons. These adults spent much more time on textbase level questions than other levels, and as a result, they achieved higher accuracy on these questions than questions addressing other theoretical levels. Similar to higher performers, conscientious readers may also have a standard of comprehension to focus on the explicit text. They spent time reading the texts closely and achieved high accuracy. The performance of conscientious readers on situation model and rhetorical structure questions was not as good as on textbase lessons but still quite impressive. Overall, the conscientious readers spent longer time on the texts and questions at the discourse comprehension components (i.e., textbase, situational model, rhetorical structure) and achieved comparatively higher accuracy than two of the other clusters, thereby suggesting better comprehension.

*Under-engaged readers* performed better on discourse level questions than on word level questions, although the average time they spent on word level questions was longer than that on situation model and rhetorical structure. Similar to struggling readers, under-engaged readers showed asymmetry between low-level and high-level reading skills. Landi (2005) found two patterns of asymmetry between lexical and comprehension components (i.e., high-lexical/ low-comprehension pattern, low-lexical/ high-comprehension pattern), but the under-engaged readers' performance mainly fell into the low-lexical/high comprehension category. It is possible the background knowledge of these adults helped them build coherent mental models in lessons tapping high-

level comprehension components (Kendeou et al., 2014; O'Reilly et al., 2019), but such knowledge was not as useful for lower level components such as word decoding. Overall, the learning behaviors of the four clusters varied across theoretical levels, which suggests that these levels represent distinguishable components of comprehension. This finding also supports previous studies of AutoTutor that reported the three discourse levels were separable since they were not highly correlated (Graesser et al., 2019).

*Struggling readers'* performance was poor overall, but they performed better on situation model and rhetorical structure level items than on word and textbase level items. They also spent longer time on word lessons than they did on lessons at the other levels. These results seem counterintuitive since word level items correspond to basic reading components and presumably are easier to process than textbase, situation model and rhetorical structure items that cover discourse components (Millis et al., 2019). There is some evidence for a general dissociation between comprehension and low-level basic reading skills (Landi, 2010; Perfetti, 2007), which could help explain why adults' performance on these two skill types would dissociate. Adults could potentially do better on higher level components because they can draw upon their world knowledge to answer questions, but this strategy is not as useful for lower level components such as word decoding.

Regarding the third goal, namely, exploring the extent to which the clusters showed improvement from pretest to posttest, we found a relationship between adult learners' improvement in reading comprehension and their behavior patterns in AutoTutor. The higher performers achieved the highest learning gains among the four clusters. The higher learning gains of this cluster were expected, given that this group showed the highest accuracy in the AutoTutor lessons. In contrast, struggling readers showed the least improvement among all groups, and this was also reflected in their poor accuracy and longer time spent on AutoTutor lessons. For struggling readers, the lessons might be too difficult, so they spent considerable time in AutoTutor making efforts to comprehend but not making progress, a signal of "wheel spinning" (Beck & Gong, 2013; Fang et al., 2017). Surprisingly, the performance of under-engaged readers was not significantly different from that of conscientious readers for the two reading comprehension tests. One possibility is that the 100-h intervention was not long enough for the conscientious readers to improve significantly more than the under-engaged readers. Given the trend shown in the results, the difference between conscientious readers and under-engaged readers might be shown with a longer intervention duration. The under-engaged readers' scores were significantly lower in RAPID, but not Woodcock-Johnson III when compared to the higher performers. One possibility of the difference might be attributed to the different samples included in RAPID and Woodcock-Johnson III. RAPID was only used in waves two and three, but Woodcock-Johnson III was used in all three waves. Therefore, we conducted a follow-up general linear model analysis to compare the improvement in Woodcock-Johnson III scores of the four clusters limiting the data to waves two and three. The results indicated nonsignificant differences between clusters. As such, the finding that under-engaged readers' scores were significantly lower than higher performers only in RAPID was consistent using different samples. One explanation of these results is that only RAPID was sensitive enough to pick under-engagement and penalized it accordingly.

## Implications and Limitations

This study suggests that clustering methods can be used to enhance the adaptivity of ITS in future studies. Differences in time and accuracy on theoretical levels indicate that an ITS that provides feedback on accuracy alone or on response time alone would be misguided. Instead, feedback and assessment that take into account trends in accuracy, time, and their interaction can better target the adult profile of reading comprehension. Assessments and feedback can be personalized to assist different groups of adults who exhibit particular patterns of learning behavior.

When the readers are classified into one of the four clusters, AutoTutor can be designed to select materials, items and trialogue moves that are sensitive to the characteristics of different clusters. For instance, struggling readers can be given more practice on basic level reading skills (i.e., word decoding and vocabulary), given that the intervention was apparently too difficult for them. Under-engaged readers need to be encouraged to spend more time concentrating on the explicit text and concentrating on the instructional interventions. High performers may be encouraged to increase reading activities on topics that interest them so they read more, or they can be assigned challenging texts that are useful to advance their careers and life. Conscientious readers can be granted freedom to proceed at their own pace. As such, students' strengths and weaknesses detected by an ITS can help the ITS improve adaptivity, as well as provide suggestions to human instructors who assist adult readers using the technology.

Although clustering analyses of the students' learning behavioral data can benefit both ITS and human instruction, the analysis is contingent on a sufficient amount of data being collected. That is, the system needs to collect students' behavioral data (i.e., response time and accuracy) for a sufficient amount of time or over a certain number of lessons in order to cluster the students into different groups reliably. We are not sure about the minimum time or number of lessons required for reliable clustering without validating the results with other datasets. Currently, we are proceeding with AutoTutor data collection on larger samples of adults with low literacy skills. The next step is to validate the findings with the new data and to estimate how many observations are needed for a reliable assignment of an adult to a particular cluster. Another limitation of this study is that Woodcock-Johnson III was developed for individuals of all ages, and RAPID was developed for K-12 students. Neither test has been specifically validated on adult literacy students. The differences between the two tests might be related to the discrepancy in students' reading improvement shown by the two reading measures.

**Code Availability** Not applicable.

**Authors' Contributions** Ying Fang: Writing, revision, and data analysis.
Anne Lippert: Data collection and editing.
Zhiqiang Cai: Software development and data collection.
Su Chen: Data analysis.
Jan Frijters: Experimental design and reviewing.
Daphne Greenberg: Experimental design and reviewing.
Arthur Graesser: Software implementation, reviewing and editing.

**Funding** The research reported here was supported by the Institute of Education Sciences, US Department of Education, through grants R305C120001 and R305A200413, and the National Science Foundation under the award The Learner Data Institute (award #1934745).

**Data Availability** The data for this study can be obtained from DataShop at https://datashop.memphis.edu/Project?id=24 upon request. Once the request is approved, the data can be exported from DataShop. The researchers interested in working with this data can also directly communicate with the corresponding author, who may be able to help facilitate any requests.

## Declarations

**Conflict of Interest** None.



## References

Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review, 1*(1), 3–38.

Baker, R. S. J. D., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18*(3), 287–314.

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.

Beck, J. E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th international conference on artificial intelligence in education* (pp. 431–440). Springer.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2011). clValid, an R package for cluster validation. *Journal of Statistical Software, 25*(4). https://doi.org/10.18637/jss.v025.i04.

Cain, K. (2010). *Reading development and difficulties*. Wiley-Blackwell.

Carvalho, P. F., Gao, M., Motz, B. A., & Koedinger, K. R. (2018). Analyzing the relative learning benefits of completing required activities and optional readings in online courses. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings of the 11th international conference on educational data mining* (pp. 418–423). International Educational Data Mining Society.

Chounta, I. A., & Carvalho, P. F. (2019). Square it up! How to model step duration when predicting student performance. In D. Azcona & R. Chung (Eds.), *Proceedings of the 9th international conference on Learning Analytics & Knowledge* (pp. 330–334). Association for Computing Machinery.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology, 31*(3), 207–208.

Deane, P., Sheehan, K. M., Sabatini, J., Futagi, Y., & Kostin, I. (2006). Differences in text structure and its implications for assessment of struggling readers. *Scientific Studies of Reading, 10*(3), 257–275.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader–text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*(3), 515–528.

Elish-Piper, L. (2007). Defining adult literacy. In B. J. Guzzetti (Ed.), *Literacy for the new millennium*, *Adult literacy* (Vol. 4, pp. 3–16). Praeger.

Fang, Y., Nye, B. D., Pavlik, P. I., Xu, Y. J, Graesser, A. C., & Hu, X. (2017). 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). International Educational Data Mining Society.

Fletcher, J. D. (2003). Evidence for learning from technology-assisted instruction. In H. F. O'Neil & R. S. Perez (Eds.), *Technology applications in education: A learning view* (pp. 79–99). Erlbaum.

Foorman, B. R., Petscher, Y., & Schatschneider, C. (2017). *Technical manual for Lexia RAPID assessment version 3.0: Grades 3-12.* Lexia learning. Retrieved from: http://www.lexialearningresources.com/RAPID/RAPID_TechnicalK2.pdf

Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education, 26*, 124–132.

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 & D. S. McNamara (Eds.), *Adaptive educational technologies for literacy instruction* (pp. 288–293). Taylor & Francis Routledge.

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

Graesser, A.C., Greenberg, D., Frijters, J.C., & Talwar, A. (2018). *Using computer agents to track performance and engagement in a reading comprehension intervention for adult literacy students.* Manuscript submitted for publication.

Graesser, A. C., Greenberg, D., Olney, A. M., & Lovett, M. W. (2019). Educational technologies that support reading comprehension for adults who have low literacy skills. In D. Perin (Ed.), *Wiley adult literacy handbook* (pp. 471–493). Wiley.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3*(2), 371–398.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115*(2), 210–229.

Graesser, A. C., Rus, V., & Hu, X. (2017b). Instruction based on tutoring. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 460–482). Routledge Press.

Haberlandt, K. F., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General, 114*(3), 357–374.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman Group Ltd..

Just, M. A., & Carpenter, P. A. (1987). The psychology of reading and language on English competence. *Language Research, 39*, 441–471.

Kendeou, P., Van Den Broek, P., Helder, A., & Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research & Practice, 29*(1), 10–16.

Kintsch, W. A. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly, 10*, 62–102.

Landi, N. (2005). *Behavioral and electrophysiological investigations of semantic processing in skilled and less-skilled comprehenders* (Doctoral dissertation, University of Pittsburgh).

Landi, N. (2010). An examination of the relationship between reading comprehension, higher-level and lower-level reading sub-skills in adults. *Reading and Writing, 23*(6), 701–717.

Lovett, M. W., Lacerenza, L., & Borden, S. L. (2000). Putting struggling readers on the PHAST track: A program to integrate phonological and strategy-based remedial reading instruction and maximize outcomes. *Journal of Learning Disabilities, 33*(5), 458–476.

Lovett, M. W., Lacerenza, L., De Palma, M., & Frijters, J. C. (2012). Evaluating the efficacy of remediation for struggling readers in high school. *Journal of Learning Disabilities, 45*, 151–169.

Meyer, B. J., Wijekumar, K., Middlemiss, W., Higley, K., Lei, P. W., Meier, C., & Spielvogel, J. (2010). Web-based tutoring of the structure strategy with or without elaborated feedback or choice for fifth-and seventh-grade readers. *Reading Research Quarterly, 45*(1), 62–92.

Millis, K., Long, D.L., Magliano, J.P., & Wiemer, K. (2019) (Eds.). *Deep comprehension: Multi-disciplinary approaches to understanding, enhancing, and measuring comprehension*. New York: Routledge.

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.

O'brien, E. J., Rizzella, M. L., Albrecht, J. E., & Halleran, J. G. (1998). Updating a situation model: A memory-based text processing view. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(5), 1200–1210.

OECD. (2016). Skills matter: Further results from the survey of adult skills. *OECD Skills Studies*. Paris, France: OECD Publishing.

O'Reilly, T., Wang, Z., & Sabatini, J. (2019). How much knowledge is too little? When a lack of knowledge becomes a barrier to comprehension. *Psychological Science, 30*(9), 1344–1351.

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

Rapp, D. N., Broek, P. V. D., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11*(4), 289–312.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*(2), 31–74.

Rouet, J. (2006). *The skills of document use: From text comprehension to web-based learning*. Erlbaum.

Sabatini, J., O'Reilly, T., Dreier, K., & Wang, Z. (2019). Cognitive processing challenges associated with low literacy in adults. In D. Perin (Ed.), *Wiley handbook of adult literacy*, 15-39.

Stenner, A. J. (1996). *Measuring reading comprehension with the Lexile framework*. Durham, NC: MetaMetrics, Inc. http://files.eric.ed.gov/fulltext/ED435977.pdf.

Strucker, J., Yamamoto, K., & Kirsch, I. (2007). *The relationship of the component skills of reading to IALS performance: Tipping points and five classes of adult literacy learners* (NCSALL report no. 29). Cambridge, MA: National Center for the Study of Adult Learning and Literacy.

Talwar, A., Greenberg, D., & Li, H. (2020). Identifying profiles of struggling adult readers: Relative strengths and weaknesses in lower-level and higher-level competencies. *Reading and Writing*, 1–17.

Tamassia, C., Lennon, M., Yamamoto, K., & Kirsch, I. (2007). *Adult education in America: A first look at result from the adult education program and learner surveys*. Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/ETSLITERACY_AEPS_Report.pdf.

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). The Guilford Press.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227–265.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31*(1), 3–62.

Vaughn, S., Fletcher, J. M., Francis, D. J., Denton, C. A., Wanzek, J., Wexler, J., Cirino, P. T., Barth, A. E., & Romain, M. A. (2008). Response to intervention with older students with reading difficulties. *Learning and Individual Differences, 18*, 338–345.

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

Williams, J. P., Stafford, K. B., Lauer, K. D., Hall, K. M., & Pollini, S. (2009). Embedding reading comprehension training in content-area instruction. *Journal of Educational Psychology, 101*(1), 1–20.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). *Woodcock-Johnson III normative update complete*. Riverside Publishing.

Woolf, B. P. (2009). *Building intelligent tutoring systems*. Morgan Kaufman.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 386–397.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162–185.

## Affiliations

**Ying Fang[1,2] · Anne Lippert[3] · Zhiqiang Cai[4] · Su Chen[4] · Jan C. Frijters[5] · Daphne Greenberg[6] · Arthur C. Graesser[4]**

[1] Central China Normal University, Wuhan, China

[2] Arizona State University, Tempe, AZ, USA

[3] Prairie View A & M University, Prairie View, TX, USA

[4] University of Memphis, Memphis, TN, USA

[5] Brock University, St. Catharines, ON, Canada

[6] Georgia State University, Atlanta, GA, USA