IES GRANTS: R305C120001; R305A200413
Graesser, A.C., Greenberg, D., Frijters, J.C., & Talwar, A. (2021). Using AutoTutor to track performance and engagement in reading comprehension. *Revista Signos, 54(107), 1089-1113.*

Using AutoTutor to Track Performance and Engagement in a Reading Comprehension Intervention for Adult Literacy Students

Arthur C. Graesser
University of Memphis, Memphis, TN USA
graesser@memphis.edu

Daphne Greenberg
Georgia State University, Atlanta, GA USA
dgreenberg@gsu.edu

Jan C. Frijters
Brock University, St. Catharines, Ontario, Canada
Jan.frijter@brocku.ca

Amani Talwar
Georgia State University
Atlanta, GA USA
Atalwar1@gsu.edu

Abstract: A large percentage of adults throughout the world have low reading skills. Computer technologies can potentially help these adults improve their literacy in addition to instructors at literacy centers. AutoTutor was designed to teach comprehension strategies by implementing conversational *trialogues* in which two computer agents (tutor and peer) hold spoken interactions with the adult about words, sentences, and text in digital lessons. The agents model comprehension strategies, ask questions, and give feedback on adult answers. AutoTutor records in log files the adults' performance, namely the time and accuracy of answering questions in the conversation. We assessed the value of AutoTutor in a study with 52 adult literacy students in the United States and Canada who interacted with AutoTutor as part of a 4-month intervention with human instructors. Performance in AutoTutor was tracked at four theoretical discourse levels (words, explicit textbase, conceptual situation model, rhetorical structure) and also engagement, with an objective psychometric measure of comprehension skill both before and after the intervention. The results showed that AutoTutor provides nuanced performance and engagement measures that predicted comprehension improvements and can be used to guide formative assessment for instructors.

Key Words: adult literacy, AutoTutor, conversational agents, comprehension training discourse levels

INTRODUCTION

Giovanni Parodi had many visions and missions throughout his career that involved many countries and many disciplines to help the world understand reading and writing in diverse populations. The first author of this manuscript visited Valparaíso, Chile three times over the last two decades. Giovanni was interested in the relations between reading and writing at an early conference he hosted. He published an article in a major international journal, *Reading and Writing* (Parodi, 2007), that investigated reading-writing relationships, with a distinctive emphasis that multiple levels of discourse need to be considered rather than only decoding, vocabulary and syntax. At a recent conference he hosted in 2018, the emphasis was on understanding reading and writing literacy with different media. One example explored by his research team was whether printed text or computer delivery was most preferred and used by researchers who were desired a deep analysis of subject matter. The case was made that there are times when printed texts have advantages. These are just a few examples of Giovanni's contributions that span multiple countries, disciplines, methodologies (e.g., eye tracking, linguistic analyses, think aloud protocols), and levels of discourse processing. This paper builds on this vision by exploring a digital technology on the web that helps struggling adult readers improve their comprehension at multiple levels of discourse. The digital technology, *AutoTutor*, has conversational agents that hold conversations with the learner and each other in natural language. Interestingly, one of the early evaluations of AutoTutor on a science subject matter appeared in the journal that Giovanni launched, namely *Regista Signos* (Jackson & Graesser, 2007).

Computer use is ubiquitous in today's society and woven into the lives of the majority of adults. Many adults who struggle with reading comprehension are also exposed to computer resources but do not necessarily use technologies to advance their reading skills. A recent adult literacy assessment (Programme for International Assessment of Adult Competencies, PIAAC; OECD, 2013) reports that 81% of people at the two lowest levels of reading have used a computer, with 76% reporting that they use a computer in daily life. These rates of access to technology for struggling adult readers are respectable, but hardly guarantees that the adults are using digital resources to promote reading comprehension. Promoting reading comprehension skill development in adults is important, as it is a major requirement for employment in higher paying jobs in the 21st century (Autor, Levy, & Murnane, 2003). The purpose of this article is to describe a web-based tool designed to increase the reading comprehension skills of adults who struggle with reading, as well as describe results of a study with 52 adult literacy students who interacted with the tool as part of their class with human instructors in a 4-month intervention. Unlike human instruction, AutoTutor can track performance of students during learning by the correctness and time of answers to questions in the AutoTutor conversations. We document how well these performance data can predict gains in comprehension skills and also engagement in the learning process.

1. Theoretical framework

Comprehension is a complex skill that involves multiple levels of discourse and particular strategies that are affiliated with particular levels of discourse (Crossley & McNamara, 2016; McNamara, 2007). This Introduction discusses these levels and affiliated strategies. We do so in the context of adaptive learning technologies, notably

AutoTutor, that trains adults in improving comprehension strategies. We start with the technology and then branch to the theoretical components that AutoTutor implements. AutoTutor was developed as part of a larger intervention study to help adult literacy students learn comprehension strategies (Graesser et al., 2016; Graesser, Greenberg, Olney, & Lovett, 2019; Graesser, Li, & Forsyth, 2014). It was designed for use in group instruction, tutorial sessions, and/or by adult students working independently outside of an instructional setting.

1.1 The value of adaptive computer technologies in the web in adult reading instruction

Poor attendance is a frequently cited problem in adult literacy programs (e.g., Greenberg et al., 2011) so a web-based component to reading instruction can help adults in a number of ways. The adults can work on the program when they cannot attend class on a regular basis including their home. Individualized instruction, as in the case of AutoTutor, is generally considered better than instruction in which all students follow the same scope and sequence at the same pace (Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007). This personalized adaptation allegedly keeps the adult engaged and minimizes drop out.

AutoTutor follows suggested guidelines for adult education computer programs by including motivating and animated agents, a simple and intuitive design, abundant scaffolds, and modules that are short in duration (National Research Council, NRC, 2011; Newnan, 2015). It follows many of the universal design for learning (UDL) premises, by including whenever possible "… multiple means of representation-…anticipating and addressing in advance any physical, perceptual, and cognitive barriers that might interfere with students' learning" (Hall, Cohen, Vue, & Ganley, 2015, p. 72). Most importantly, it follows the critical requirement of motivating adult literacy students.

Motivation has been proven critically important to reading development (e.g., Fulmer & Frijters, 2011), and the nature of motivation for reading changes across the lifespan (Ryan & Moller, 2017; Wlodkowski & Ginsberg, 2017). All adult literacy students need to feel that the instructional experience meets their needs, and is a worthwhile investment of personal time and energy, which are often limited by competing demands in their lives. Since most adults who attend adult literacy programs do so voluntarily, if the instruction is not adult-oriented, engaging, and pertinent to adult daily life, adult literacy students will stop attending. Therefore, AutoTutor incorporated content that was interesting and useful to adults.

Learner engagement is obviously important in intelligent tutoring systems as well as a broad array of learning environments (e.g., Graesser & D'Mello, 2012), including adult literacy (Chen et al., 2021; Windisch, 2016). Engagement is a core dimension of motivation, which is known to predict learning. Across many skill domains, when comparing a person with high motivation and low skills to an individual with low motivation and high skills, the person with the high motivation can often outperform the one with low motivation, even though the lower motivated individual has higher skills (Ginsberg & Wlodkowski, 2015). While empirical research on adult reading performance is scant, this dynamic has been observed for adolescent reading (e.g., Wolters, Denton, York, & Francis, 2014).

In addition to including easy-access, individualized self-paced instruction, and intuitive design, AutoTutor was designed to optimize engagement by including a number of other features. First, it has two computer agents (a teacher and a peer student) that hold a conversation with the student in a *trialogue* (Graesser et al., 2014). The agents guide the adult literacy student on what to do next, model activities and strategies, provide immediate feedback on why the adult is correct or incorrect when completing an activity, express positive encouraging messages when the adult is not performing well, and sometimes stage game-like competitions between the adult literacy student and a peer student agent (with the adult always winning, thereby enhancing self-esteem). AutoTutor includes lessons with texts that have adult-oriented practical value and/or interest (such as rental agreements, job applications, recipes, health information). Texts are selected by AutoTutor to be at a reading level that the student can handle (not too hard or too easy). Finally, AutoTutor was designed to be intelligently adaptive to the student's performance rather than rigidly scripted.

1.2 A typical AutoTutor lesson

Thirty-five AutoTutor lessons have been developed for comprehension instruction during the course of AutoTutor development (Graesser et al., 2016). An introductory video on digital skills is presented to help adults learn skills such as keyboard input, scrolling, and login details. Most lessons begin with a 2-3 minute video that reviews a comprehension strategy. The introductory video presents a didactic lecture of the lesson with visual images and example materials to refer to. Each lesson is unlocked in a sequential order, with all previous lessons unlocked for students to retake lessons if they desire. Every lesson has between 12 and 35 questions and the computer records their performance on these questions, which consists of the correctness of their answers and the time taken to answer the questions. Example lessons can be viewed online (www.arcweb.us, www.sites.org).

The AutoTutor trialogues include two computer agents, a teacher (Cristina) and a peer student (Jordan). These "talking heads" help adults learn by interacting with the adult literacy students in natural language and by frequently referring to texts and multimedia. They scaffold students through different types of reading comprehension strategies by questioning, hinting, eliciting information, giving short feedback, explaining how answers are right or wrong, and filling in gaps of information. Jordan often presents difficulties he is having, which form the basis of the comprehension strategy lesson Cristina presents to Jordan and the adult literacy student. Each lesson takes between 10 and 50 minutes to complete and involves adult authentic activities such as filling out job application forms and instructions on how to change a flat tire. Trialogue conversations are crafted so that the adult literacy students do not feel they are failing. For example, when Cristina asks the human student and Jordan a question, the human student is often asked the question first, with Jordan typically agreeing with the human student or getting the answer wrong. Most of the negative feedback is directed at Jordan so that the human student does not perceive constant failure from negative feedback. Another adaptive feature of AutoTutor is all lessons begin at an intermediate difficulty level, with the level being changed to being more or less difficult based on the adult literacy student's performance. To decrease the amount of required digital and typing skills, students respond to all of the questions with only a few functions, such as clicking a response, dragging and dropping, or highlighting (although in three of the lessons, there is a small number of questions that require the student to type an answer with a word or phrase).

AutoTutor has an audio function, which the students can use if they find specific continuous text difficult to read (the audio function is not available for question items). Finally, there is a *home* button in case the student wants to restart a lesson and a *repeat* button if the student wants the recent conversational turn to be repeated.

An example of an AutoTutor interaction will illustrate how the two agents scaffold reading comprehension strategies and enhance motivation. Figure 1 is a screenshot from a lesson that provides a practical example about changing a tire. As the adult literacy students read the text on changing a tire, they need to consider the order of events, a task that requires deep understanding. Figure 1 shows the text and a small portion of the conversation that establishes this understanding.

Insert Figure 1 here

In this particular lesson, the text has the rhetorical structure of a procedure, where a sequence of actions must be performed in a certain order. The order of mention in the text may not be the same as the chronological order in which the events unfold in the world. Signal words (i.e., *before, after, during, first, and then*) help clarify when order of mention is different than the sequence of events in the world. The teacher agent instructs the student agent and adult literacy student to take turns identifying the next step in the procedure by either clicking on a sentence or a highlighted word in the text that acts as a signal word (in Figure 1, Jordan, the student agent, has already had his turn and has selected the word “before”).

Sometimes the chronological order does not match the presentation order in the passage, and in these cases, adult literacy students must reconstruct the mental model of the chronological order. That is, the order of actions requires that the reader pay attention to both the order of information presented in the text and the signal words. This example reflects the rich literature in discourse processing on how *situation (mental) models* are constructed in the in the mind of the reader from the language and discourse in the explicit text (Kendeou & O'Brien, 2018; Kintsch, 1998; Van den Broek, White, Kendeou, & Carlson, 2009). AutoTutor instructs the learner that the order of mention may deviate from the order of events in the world.

Another aspect of this trialogue conversation is worth noting. The conversation regarding tire changing steps is in the *testing mode* because the teacher agent is testing the adult or peer agent on their understanding by asking questions or soliciting actions (turn 1 in Figure 1), giving short feedback ("you are right" in turn 3), and also providing content that repeats, elaborates, or explains the correct answer (the second sentence in turn 3). During AutoTutor lesson development, testing mode was not uniformly used because it has a "schoolish" pragmatic aspect that may be demotivating for adults. Another type of trialogue conversation is *game mode*, which is presumably more motivating, and involves the adult competing for points with the student agent in a game. In *help mode*, the adult literacy student helps the struggling student agent with a task. By successfully assisting the student agent, the self-esteem of the adult literacy student can increase. These pragmatic conversational modes illustrate how the agent conversation in AutoTutor enhance both motivation and cognitive comprehension strategies.

1.3 Comprehension strategies and theoretical structure of lessons

AutoTutor lessons were aligned with a successful teacher-led strategy intervention that was designed for high school struggling readers (Lovett, Lacerenza, De Palma, & Frijters, 2012). The strategies (*PACES*) included in the AutoTutor lessons cover: (1) ***P***redicting topic and writer's purpose; (2) ***A***cquiring new vocabulary; (3) ***C***larifying common sources of confusion; (4) ***E***valuating, elaborating, and explaining; (5) ***S***ummarizing, identifying, and constructing text structures.

The theoretical model underlying the AutoTutor lessons consisted of the multilevel framework of Graesser and McNamara (2011), as well as other discourse processing models (Kintsch, 1998; Perfetti, 1999). The Graesser and McNamara (2011) framework identifies six levels: *words, syntax*, the explicit *textbase*, the referential *situation model*, the *genre/rhetorical structure*, and the *pragmatic communication* level. This study focused on four of the levels: words, textbase, situation model, and genre/rhetorical structure. Words represent the lower-level *basic reading* components that include morphology, word decoding, and vocabulary (Perfetti, 2007; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). According to some theories, if the basic components are not mastered, there will be negative consequences on the development of deeper comprehension (Connor et al., 2007; Van den Broek et al., 2009; Vaughn et al., 2008). The other components represent the *discourse* components, which are allegedly more difficult to master. The *textbase* level focuses on the meaning of explicit ideas. The *situation model* (also referred to as a mental model) represents the subject matter, including inferences generated by background knowledge. This model differs based on text type. For example, narrative text includes information about characters, settings, actions, and emotions while informational text covers specific content (e.g., knowledge and inferences about cooking and food when reading a recipe). *Genre/rhetorical structure* (also referred to as the type of discourse and its composition) focuses on the type of text, such as narrative, persuasion, and informational genres. Each genre includes a different type of structure. For example, fiction may have a conflict and resolution whereas persuasive essays have arguments and supporting information. There are different informational texts, such as compare-contrast and problem-solution. We refer to this component as rhetorical structure, noting that these structures differ among the genres. The distribution of AutoTutor lessons covered a broad distribution of comprehension skills representing words, sentence comprehension, narrative text, persuasive text, informational text of different subcategories, and also digital media.

The texts in AutoTutor were scaled on difficulty using automated measures of Cohmetrix (cohmetrix.com, Graesser et al., 2014; McNamara et al., 2014). Cohmetrix scales texts on difficulty by considering all levels of discourse (words, syntax, textbase situation model, rhetorical structure, genre) rather than the traditional measures that merely measure word length, sentence length and word frequency. Cohmetrix has a composite measure of *formality* that considers all of these discourse levels (Graesser et al., 2014); oral conversational language is at the low end and stilted academic discourse is at the high end.

1.4 Introduction to current study

This article investigated adult literacy students who participated in a 100-hour reading intervention that was blended between teacher-led sessions and AutoTutor. This article reports results of data that were recorded from the AutoTutor log files that were collected throughout the 100-hour intervention over approximately 4 months. It should be noted that AutoTutor was approximately 25% of the intervention so it is inappropriate to make any claims that AutoTutor was the primary cause of improvements in comprehension skills and engagement. However, AutoTutor was able to extract samples of performance and engagement of the adults while they were learning with AutoTutor. These data were expected to be diagnostic of the adults' experiences during the intervention with respect to learning and engagement.

The first goal of this study was to determine whether there are performance differences as a function of the four theoretical levels (Words, Textbase, Situation Model, and Rhetorical Structure) with respect to the accuracy and time of answers to questions that adults were asked during the intervention. There are different hypotheses about how these theoretical levels may differ. One hypothesis is that that time will increase and accuracy decrease as the levels go from words and explicit information to higher levels of discourse and meaning, generating the following prediction for time (and reverse for accuracy): Words < Textbase < Situation Model < Rhetorical Structure. A modified form of this hypothesis only considers the split between Words and the other three levels that address multi-sentence discourse: Words < Textbase = Situation Model = Rhetorical Structure. The instruction of words could have shorter time cycles of interaction because the text and task have a small information load that spans a word or a single sentence. In contrast, the discourse levels would take more time (and have lower accuracy) because there is a text with several sentences and higher information load to process. It is also conceivable that the discourse levels would take so much time and concentration that the adults would disengage and perform poorly. There is a competing hypothesis, however, that addresses the number of new contexts that need to be constructed to answer the questions. For lessons that address words, each question addresses a new context; for those that address discourse (the other three levels), several questions are asked about a single text so there are fewer contexts that need to be constructed. The prediction for time to answer a question is: Words > Textbase = Situation Model = Rhetorical Structure. Accuracy may also follow this pattern, but there is a trade-off between item difficulty and performance so the prediction is not clear-cut.

The second goal addresses the readers' engagement. The accuracy and time profiles collected by AutoTutor served as a diagnostic measure of whether a student is engaged. A participant is not likely to be engaged when the performance is very low or near chance levels. Allocation of reading time is another example of a behavior that has been used to measure reading motivation (Mills, Graesser, Risko, & D'Mello, 2017). In essence, extremely short times or extremely long times are probabilistic signals of disengagement (Baker et al., 2008). Extremely short times are a signal of the reader quickly perusing the material or "gaming the system" in route to getting a correct answer without comprehending or learning. Extremely long times are a signal of mind wandering (e.g., Feng, D'Mello, & Graesser, 2013) or simply taking a break and leaving the learning environment for a span of time. An engaged reader stays within the *zone of engagement* whereas disengaged readers exceed the boundaries of these fast and slow times.

The accuracy and performance profiles can theoretically define different categories of students. A *proficient* reader is accurate and comparatively fast, although not too fast to the point of disengagement. A *disengaged* reader is inaccurate and extremely slow or fast, depending on how the reader handles the frustration of underperforming. A *conscientious* reader is accurate and comparatively slow, although not too slow to the point of mind wandering and disengagement. The present study investigated the relationship between the adults' time and accuracy profiles and whether the combination of these two factors predicted improvement in comprehension skills, as measured by an independent objective test of comprehension.

Research is conspicuously absent in the role of digital technologies in improving comprehension training in adults with low literacy. This project with AutoTutor is the first to investigate the role of conversational agents in conversation-based training and assessment for this population. The accuracy and time of answers to conversation-based questions in AutoTutor allow us to track performance in a blended intervention with human instructors and explore whether the performance profiles predict improvements in comprehension skills and engagement in the intervention. The AutoTutor project is also the first to differentiate performance on lessens that tap different levels of discourse processing for adult learners.

2. Methodological framework

2.1 Participants

As part of a larger intervention study (grant number to be revealed after peer review process) 52 students in Metro-Atlanta (n = 20) and Metro-Toronto (n = 32) were recruited from adult literacy classes. The majority of the participants were female (73.1%) and native English speakers (71.2%). Race and ethnicity were reported as follows: majority was Black/Caribbean/African American (61.5%), followed by White (23.1%), Multiracial (9.6%) and Asian (5.8%). Participants ranged in age from 16-69 years with a mean age of 40 (SD = 14.97). Although 26.9% of the participants claimed that they had a high school diploma, all adults were enrolled in classes targeting those who read between the 3.0 and 7.9 grade levels. Close to 30% reported that they had attended at least one special education class as a child, with 26.9% reporting that they had been tested as a child for a learning disability.

2.2 AutoTutor lessons and measures

AutoTutor lessons are available on the web and can be accessed and used at no cost (www.arcweb.us, www.sites.autotutor.org). The lessons are undergoing changes with research and development, but are very similar to the lessons developed for this intervention.

AutoTutor lessons were conducted during class time. Students worked independently on the lessons while the teacher was present in the room to aid with any technological difficulties (for example, if the computer "froze"). There were 29 lessons in AutoTutor in which performance measures were collected. Each lesson was scaled on the primary theoretical level and any secondary or tertiary theoretical levels. Regarding the classification on the primary theoretical levels, the 29 lessons were classified according to four discrete levels. The Word measures covered the following strategies: word parts, word-meaning clues, learning new words, multiple meaning words. The Textbase lessons covered punctuation, pronouns, key information, main ideas, and persuasion 1. The Situation Model lessons covered nonliteral language, text signals, connecting ideas, stories (1, 2, and 3), persuasion 2, inferences from text, forms and documents, and 2 review lessons. The Rhetorical Structure lessons included purpose of texts, steps in procedures, problems and solutions, compare and contrast, cause and effect, describing things, time and order, and one review. We also measured lessons on a continuous measure with respect to theoretical level. Each lesson received a score of

1.00 on a theoretical level if it was the primary assignment, .67 if it was secondary, .33 if it was tertiary, and .00 if it was not included.

Most of the lessons in AutoTutor were adaptive to the adult's accuracy over the course of a lesson. The words, sentences, or texts started out being medium in difficulty in the beginning phase of the lesson (1/4 to 1/2), as scaled on objective measures of word or text difficulty that are specified in Graesser, Feng, and Cai (2017). Accuracy was measured in this early phase of a lesson. When the accuracy met or exceeded some threshold (i.e., .67 in most lessons, whereas .33 was approximately chance accuracy), the subsequent assigned materials were more difficult. When accuracy failed to meet threshold, the subsequent assigned materials were easier. The assignment of accuracy-contingent materials was an important feature of the AutoTutor intelligent tutoring system. The analyses reported in the Results section focused entirely on the medium items. These were questions asked about words, sentences, or texts that were scaled on a medium level of difficulty. The observations that involved branching to easier or more difficult materials were collected but not reported here because we wanted to make some comparisons on materials that all adults received. For these medium-level difficulty observations, we collected the accuracy of performance and the time to answer the question. Time was measured from the onset of the question to the onset of the participant's answer, which was always a click on an option on the computer screen. Accuracy and time to respond per item (question) scores can be calculated with different units of analysis, namely participants, lessons, and questions within a lesson. There were 13,556 observations altogether when considering number of completed participants, lessons, and questions.

As expected, the response times per question were positively skewed when we observed the distribution of times. That is, some of the times were extremely long, possibly because the participants left the lesson and returned or tuned out for a long duration. We handled these outliers in two ways. First, there was a truncated time measure that computed a personalized distribution of times for each participant and replaced observations of greater than 3 standard deviation z-score units with the score they would receive at 3 z-score units above the mean. Second, the log time measure consisted of a log transformation on the raw response times. The log transformation is commonly computed on response times in order to convert a positive skewed distribution to a normal distribution (Tabachnick & Fidell, 2019).

In addition to the data collected by AutoTutor, as part of the larger intervention study, 43 of the 52 participants completed both a pretest and a post-test on the Passage Comprehension subtest of the Woodcock-Johnson III Normative Update battery (Woodcock, McGrew, & Mather, 2007). Form A of the assessment was administered at both time points. On average, the post-test was administered after 4.07 months of instruction.

3. Results

3.1 Goal 1: Analysis of theoretical levels

The participants completed all of the presented items in 78.1% of the lessons that were assigned (between 42.3% and 94.2% among the 29 lessons). Accuracy and time scores were scored for the questions in the 29 lessons for the 52 adults. Table 1 presents the mean times and accuracy of answers as a function of the four theoretical levels. As can be seen in Table 1, accuracy is highest and the answer times are shortest for the Word level compared to the three discourse levels (Textbase, Situation Model, and Rhetorical Structure). To confirm this trend, mixed-effect models on accuracy and time were performed to test the difference among four theoretical levels. The model included by-subject (participant), by-item (question), by-lesson random intercepts; there also are by-subject random slopes on different theoretical levels and random intercepts for the interaction between lesson and item for the nesting relationships. When the unit of analysis is the item, we applied the logistic mixed effect model on the accuracy because the item response was either correct or incorrect. Type II Wald Chi-square test showed that there was significant difference (χ^2 (3) = 8.34, p = 0.040) among four theoretical levels. We found the natural log odds ratio for Words was significantly higher than each of the three discourse levels, which in turn did not differ from each other. The coefficient value of the adults' accuracy on Word level was greater than the coefficients on the three discourse levels. The estimates of accuracy from the logistic model are presented in Table 1 and mirror the raw accuracy scores.

In contrast, time did not significantly vary among theoretical levels in a mixed effect model, an Analysis of Variance of type III with Satterthwaite, F(3,25.8)= 0.058, p = 0.981. It appeared that the Word items were the fastest and the Rhetorical Structure items were the slowest in Table 1, but that relationship did not prove to be statistically significant. Therefore, analyses of the times among the theoretical levels can be put on the same playing field.

Insert Table 1 here

We performed a follow up analyses on the continuous measures of theoretical level, with the 29 lessons as a unit of analysis. There was a significant positive correlation between accuracy and word level, $r = .386$, $p < .05$, but not between any of the discourse levels. The times showed no significant correlations among theoretical levels. The continuous theoretical measures showed an informative pattern of intercorrelations. Whereas Word level had a significant negative correlation with each of the three discourse levels (textbase, situation model, rhetorical structure, *r*s = -.365, -.485, and -.567, respectively), the correlations among the three discourse levels were all nonsignificant (between -.318 and .236). These correlations underscore the difference between the Word level, which involve basic reading processes, and the three discourse level theoretical constructs which are quantitatively separable.

3.2 Goal 2: Engagement

The next phase of our analysis had two sub-goals. The first sub-goal was to probe whether there was significant within participant and/or within lesson*participant variation in mean scores and time spent per lesson. The second was to investigate whether patterns of scores combined with time spent could characterize different types of engagement with the AutoTutor material. In the final phase of the analysis, our indices of engagement were used as predictors of reading comprehension gains.

Intra-individual variability. Performance on AutoTutor items was nested within the lessons, and lessons were nested within the individual, so a multilevel model was used to answer the first question. All models were formulated using SAS/STAT Version 14.2, Version 9.4 of the SAS System for Windows, and PROC MIXED (SAS Institute, 2013). Two models were formed, one with item accuracy as the outcome and one with log-transformed time per item (i.e., to normalize the overall distribution) as the outcome. Both models were null models, with no fixed effect predictors, only person and lessons nested within person as random effects. This model allowed for the calculation of the intra-class correlation for each of these random effects. In the context of this model, these coefficients index the proportion of variance in accuracy or duration accounted for by the person, or by the lesson nested within person (Hox, 2010).

In the model for time per item, the variance component for person was significant ($\sigma_{\text{person}} = 0.025$, *se* = .005, $p < .001$). The intra-class correlation was also moderately large ($\rho = .17$) indicating that time spent per item was more similar within, rather than across, persons. Similarly, the random effect for lesson within person was significant ($\sigma_{\text{person}} = 0.032$, *se* = .002, $p < .001$), with the intra-class correlation ($\rho = .40$) substantial, indicating that time spent per lesson was substantially similar within, compared to across individuals. A similar pattern was observed when the accuracy on each item was the outcome, with significant random effects for both participant ($\sigma_{\text{person*lesson}} = 0.005$, *se* = .001, $p < .001$, $\rho = .022$) and lesson nested within participant ($\sigma_{\text{person}} = 0.007$, *se* = .001, $p < .001$, $\rho = .056$). Note that the intraclass correlations were much larger when time per item was the outcome, compared to item accuracy, which suggests that the time is a more important determinant of individual performance patterns.

The violin plots within Figures 2 and 3 illustrate these dynamics. In Figure 2, the light grey density plots on the left represent the distribution of time spent per item across all participants; whereas dark grey density plots on the right represent time for a single example participant, chosen at random. Figure 3 is similar, but with accuracy as the outcome. On rhetorical structure items, the example participant was faster than the group as a whole, but slower on the other three theoretical dimensions; however, the example participant was less accurate on rhetorical structure, situation model, and textbase items, but similar to the rest of the group on word items.

Insert Figures 2 and 3 here

Types of engagement. The previous analysis indicated that significant and meaningful variance existed for both item time and accuracy when considering participants and lessons nested within participants. With these data at hand, we generated an index representing each participant's *zone of engagement*, which is defined solely by response time as items completed (whether completed and scored as correct or incorrect) neither too slowly nor too quickly relative to a participant's personal average speed. For each lesson nested within each participant, a distribution of (log-transformed) time spent per item was calculated. We calculated the mean score for items completed within +/- 0.5 SD of the mean lesson time. This partitioning resulted in one overall accuracy score that only included items within each participant's unique *personal speed zone*. These were the items that we considered the adult to be engaged in the learning experience. On average, this constituted 37.9% of items completed. The remainder of items was assumed to be completed *outside* of the zone of engagement.

Thus, when combined with participant accuracy, items varied along two dimensions: correctly versus incorrectly completed, and completed quickly versus slowly. The performance profiles per item were distributed as follows among the remaining items that were outside of the engagement zone: items representing proficient performance which were completed correctly and quickly (25.8%); disengaged performance which were items completed incorrectly and quickly (6.9%) or slowly (12.0%), and conscientious performance which were items completed correctly and slowly (17.5%). Since these categories were formed per item within the lessons, each participant had a profile of engagement along these dimensions. For example, the example participant in Figures 2 and 3 had the following profile: a mean score of .58, with 42.4% items completed within the zone of engagement; 21.4% proficiently completed items; 7.9% and 13.8% items characterized by a disengaged style, being completed quick and slowly, respectively; finally, 14.5% of items were completed conscientiously.

3.3 Relationship with reading comprehension

On the Woodcock-Johnson Passage Comprehension subtest (Woodcock et al., 2007), post-test scores ($M = 28.81$, $SD = 3.87$) were significantly higher than pre-test scores ($M = 27.02$, $SD = 4.25$), $t = -4.23$, $p < .001$. A participant-level model was formed to test whether scores within the zone of engagement and profiles of engagement were related to reading comprehension skills assessed at the end of the intervention. Prior to this stage of the analysis, an assessment of regression assumptions (i.e., linearity, normality, distribution of residuals, and outlying/influential scores) was conducted, indicating that our models met all critical assumptions. Reading comprehension at post-test was first regressed onto the following covariates: pre-test reading comprehension, average time per item, and overall number of items completed. This step accounted for 58.6% of the variance [F (3, 41) = 21.80, $p < .001$] in post-intervention reading comprehension, with only pretest reading comprehension significantly predicting posttest scores (standardized $\beta = .774$, $p < .001$).

In the second step, mean item score for items within each participant's zone of engagement, and proportion of items in each of the four performance profiles were entered as predictors. This step accounted for an additional 10.4% of the variance [F (5, 36) = 2.67, p = .037] in post-intervention reading comprehension, with accuracy on zone of engagement items (standardized β = .597, p = .005) and conscientiously-completed items predicting posttest scores (standardized β = -.352, p < .001). The fact that the coefficient was significantly positive is consistent with the expectation that the adults learned best when they were in the zone of engagement. Interestingly, the coefficient for conscientiously-completed items was negative, indicating that the greater proportion of items completed this way, the lower the posttest reading comprehension score, controlling for pretest. Apparently, these items that they answered correctly but took a long time to answer did not positively contribute to learning. The proportion of proficient and/or disengaged items did not predict post-test reading comprehension. The proficient items perhaps reflected what they had already mastered whereas the disengaged items did not reflect productive concentration on the material. As a note, this regression analysis was also completed using change in reading comprehension scores (via post-test minus pre-test difference scores) as the outcome. The results were identical, whether or not pre-test was also included as an additional covariate. These results support the conclusion that our engagement indicators predict gains or change in reading comprehension.

4. Discussion

This study explored whether the conversation-based training and assessment in AutoTutor lessons with conversational trialogues could help us understand improvements in comprehension skills and engagement in lessons for struggling adult readers. We used AutoTutor (Graesser et al., 2016, 2019) to track the performance of 52 adult literacy students who were part of a 100-hour intervention to improve their reading comprehension skills. Our first goal was to explore whether the adult students' time and accuracy of answers to conversation-based questions in AutoTutor varied as a function of the lessons' targeting particular discourse levels. The results revealed that accuracy of lessons targeting the word level was more accurate than the multi-sentence discourse levels (textbase, situation model, rhetorical structure), but the time was statistically equivalent among all four levels. Our second goal was to explore whether AutoTutor's time and accuracy profiles reflected engagement and could predict improvements in comprehension skills, as measured by Woodcock-Johnson III. The results supported the conclusion that comprehension skill improvement on this psychometric test of reading comprehension was best predicted by the extent to which the student was in the personalized zone of engagement with AutoTutor (with reasonable accuracy and times that were not too slow or fast). Consequently, the performance profiles of AutoTutor have shown promise in understanding adult students' reading challenges and improving their engagement and comprehension.

Regarding the first goal, we confirmed one of our hypotheses that accuracy would be higher for the word level than the three discourse levels (textbase, situation model, and rhetorical structure). The word level captures basic reading components of morphology, decoding, and vocabulary. These items place comparatively low loads on working memory because there is a focus on individual words and/or single sentences. In contrast, the discourse levels involve multiple sentences and deeper levels of comprehension requiring reasoning and inferences (Millis, Long, Magliano, & Wiemer, 2019) that would be expected to be more challenging. The deeper reading components at the discourse levels are more time-consuming, strategic, and taxing on cognitive resources. However, differences did not emerge among the three discourse levels.

Interestingly, comparisons among the lessons indicated that the three discourse levels were separable because they were not highly correlated with each other. This makes it feasible to discriminately track progress and performance on these different discourse levels that have been adopted by contemporary theoretical frameworks in discourse processing (Graesser & McNamara, 2011; Kintsch, 1998; Perfetti, 1999). The distinction between the word level and the three discourse levels was apparent in these analyses, with words representing basic reading processes.

Surprisingly, no significant relationships were found between time spent on questions and the four theoretical levels. One reason may be some tradeoffs between the factors that contribute to processing time for the lessons at the word level versus the three discourse levels. On the one hand, the questions in the word lessons should be comparatively fast because there were fewer words in the focal question and answer whereas the questions in discourse lessons referred to lengthier texts. On the other hand, the word lessons had a large number of independent items with new situations to construe whereas the discourse lessons had texts with 8-12 question items about the text. Another reason is more interesting and nuanced, namely that of engagement, the essence of the second goal of this study. It is important to separate learning experiences when the adult is engaged in the lesson versus disengaged.

Under the second goal, we used the time to answer individual questions (i.e., items) to help assess the adults' engagement in the lesson when they answer the questions. Sometimes the adults were very engaged, as defined by their personalized time on an item for a lesson. This is when they are reading and concentrating at their own pace at their personalized zone of engagement (see Mills et al., 2017).   Once items outside a participant's *zone of engagement* were filtered out, AutoTutor performance was strongly and significantly related to post-intervention reading comprehension scores. This effect was also observed with pre-post intervention change in reading comprehension.

The results also uncovered interesting findings on those observations that were outside of the zone of engagement. *Proficient* observations of adults were accurate and fast, but apparently they did not predict learning of comprehension skills over a 4-month intervention. These items were presumably those items that captured mastered skills so there was no important new learning. *Disengaged* observations of adults were the incorrect items that had very short times that reflected either (1) mechanically pushing buttons, "gaming" the system (Baker et al., 2008) or mind wandering (e.g, Feng et al., 2013). These observations are worthless because the student is not concentrating on the material. *Conscientious* observations of adults are those that are completed correctly but slowly. We had expected that participants with conscientious observations would show learning, but that was not supported. Apparently, it takes more scaffolding to support progress from these learning experiences.

Findings from this study signify that effective intelligent tutoring systems should not only consist of quality content, but should also provide teachers and researchers with nuanced performance and engagement data. Such data can be used in summative, formative, and stealth assessment. The goal is to advance the student at their zone of proximal development through these forms of assessment. Summative assessment is simply a score for each measure that is tracked, such as accuracy, answer time, and engagement percentage. Formative assessment has such scores, ideally segregated by comprehension level, presented to the instructor during the course of training. This information can be used to help the instructor make changes to the tasks, lessons, and materials from day to day. Stealth assessment uses the scores to guide the computer to adaptively making changes, without the instructor or student knowing about the scores and adaptation. For example, if the adult is performing poorly on a lesson and has low engagement, the computer would present easier and more interesting material to the adult learner.

Total correct scores may provide sufficient information about students who are competently mastering all of the material, but total correct scores do not provide enough information about struggling students. This study provides an example of an intelligent tutoring system that provides engagement profiles of items, as well as data specific to theoretical levels of comprehension material taught. This study illustrates that some students, for example, may correctly answer an item by mechanically pushing buttons (lucky guesses), others correctly answer the item because they are proficient on that item, and others correctly answer after conscientious deliberation. A total correct score alone does not differentiate these processes.

Students who are struggling with a specific level of comprehension may benefit from instruction specifically tailored to that level. For example, a student may need help with the textbase level of comprehension, but not with the word level. The only way that instructors and researchers can obtain this type of information is if the tutoring system provides a way to track both time and accuracy performance on different types of items, and a way to systematically conduct a nuanced assessment of engagement. We believe that it is well worth the effort for developers of intelligent tutoring systems to take these steps and a number of researchers have done so (e.g., Baker et al., 2008; Graesser, & D'Mello, 2012). This direction is particularly important for adults with low literacy skills, as the present study has shown. Far too many of these students are struggling with current interventions so this problem can hopefully be mitigated with programs like AutoTutor that track accuracy, time, and engagement during the process of learning.

A web-based instructional program is not for every struggling adult reader. A successful experience on AutoTutor requires the adult to have access to a computer that can handle a web-based program, and to have the motivation and self-regulation to work on a computer. Nevertheless, an increasing number of adult literacy advocates are encouraging the infusion of technology into adult reading interventions (e.g., NRC, 2011; Newnan, 2015), so it is critical to understand the role that web-based software can play in the adult literacy curriculum. Computers can collect the time, accuracy, and other performance measures on-line while adult literacy students interact with the learning environments and hopefully benefit from the intervention.

The present study used AutoTutor (Graesser et al., 2016) to track the performance of 52 adult literacy students who were part of a 100-hour intervention to improve their reading comprehension skills. Our first goal was to investigate the extent to which accuracy, and time varied as a function of theoretical level of comprehension (Word, Textbase, Situation Model, and Rhetorical Structure). Our second goal was to investigate the extent to which the accuracy of the adults' performance is predicted by their time profiles, as well as whether their engagement significantly predicted improvements in their comprehension, as measured by the Woodcock-Johnson III Normative Update battery. This study has shown how AutoTutor can be woven into an intervention with human teachers and tutors to help adults with lower levels of literacy better comprehend text.

BIBLIOLOGICAL REFERENCES

Autor, D., Levy, F., & Murnane, R.J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118, 1279-1334.

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

Chen, S, Fang, Y., Shi, G., Sabatini, J., Greenberg, D., Frijters, J., & Graesser, A.C. (2021). Automated disengagement tracking within an intelligent tutoring system. *Frontiers in Artificial Intelligence, 3, 1-16.*

Connor, C.M., Morrison, F.J., Fishman, B.J., Schatschneider, C., & Underwood, P. (2007). The early years: Algorithm-guided individualized reading instruction. *Science, 315,* 464-465.

Crossley, S.A., & McNamara, D.S. (2016)(Eds.). *Adaptive educational technologies for literacy instruction*. New York: Taylor & Francis Routledge.

Feng, S., D'Mello, S, & Graesser, A., (2013). Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review, 20,* 586-592.

Fulmer, S. M., & Frijters, J. C. (2011). Motivation during an excessively challenging reading task: The buffering role of relative topic Interest. *Journal of Experimental Education, 79*(2), 185-208.

Ginsberg, M. B., & Wlodkowski, R. J. (2015). Motivation and culture. In J. M. Bennett (Ed.). *The Sage encyclopedia of intercultural competence* (pp. 634-637). Los Angeles: Sage.

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

Graesser, A.C., & D'Mello, S. (2012). Emotions during the learning of difficult material. In. B. Ross (Eds.), *The Psychology of Learning and Motivation*, vol. 57 (pp. 183-225). Elsevier.

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

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.

Graesser, A.C., Feng, S., & Cai, Z. (2017). Two technologies to help adults with reading difficulties improve their comprehension. In E. Segers and P. Van den Broek (Eds.), *Developmental perspectives in written language and literacy. In honor of Ludo Verhoeven* (pp. 295-313). John Benjamin Publishing Company.

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. *Elementary School Journal, 115*, 210-229.

Greenberg, D., Wise, J. C., Morris, R., Fredrick, L. D., Rodrigo, V., Nanda, A. O., & Pae, H. K. (2011). A randomized control study of instructional approaches for struggling adult readers. *Journal of Research on Educational Effectiveness, 4*, 101-117. doi: 10.1080/19435747.2011.555288.

Hall, T.E., Cohen, N., Vue, G., & Ganley, P. (2015). Addressing learning disabilities with UDL and technology: Strategic Reader. *Learning Disability Quarterly, 38,* 72-83.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications.* 2nd Edition. New York: Routledge.

Jackson, G. T., & Graesser, A. C. (2006). Applications of human tutorial dialog in AutoTutor: An intelligent tutoring system. *Revista Signos*, *39*, 31–48.

Kendeou, P., & O'Brien, E. J. (2018). Theories of text processing: A view from the top-down. In M. Schober, D. N. Rapp, & M. A. Britt (Eds.), Handbook of discourse processes (2nd, pp. 7–21). New York: Routledge.

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

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*(2), 151-169.

McNamara, D.S. (2007)(Ed.). *Theories of text comprehension: The importance of reading strategies to theoretical foundations of reading comprehension*. Mahwah, NJ: Erlbaum.

McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, MA: Cambridge University Press.

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.

Mills, C., Graesser, A.C., Risko, E.F., & D'Mello, S.K. (2017). Cognitive coupling during reading. *Journal of Experimental Psychology: General, 146,* 872-883.

National Research Council [NRC]. (2011). *Improving Adult Literacy Instruction: Options for Practice and Research.* Committee on Learning Sciences: Foundations and Applications to Adolescent and Adult Literacy, Alan M. Lesgold and Melissa Welch-Ross, Editors. Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press.

Newnan, A. (2015). *Learning for Life: The Opportunity for Technology to Transform Adult Education*. http://tytonpartners.com/library/learning-for-life-the-opportunity-for-technology-to-transform-adult-education/.

OECD (2013), *Time for the U.S. to Reskill?: What the Survey of Adult Skills Says*. Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264204904-en

Parodi, G. (2007). Reading–writing connections: Discourse-oriented research. *Reading and Writing, 20,* 225–250.

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. Brown & P. Hagoort (Eds.), *The Neurocognition of Language* (pp. 167-210). New York: Oxford University Press.

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

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

Ryan, R. M., & Moller, A. C. (2017). Competence as central, but not sufficient, for high-quality motivation: A self-determination theory perspective. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.). *Handbook of competence and motivation, second edition: Theory and application* (pp. 214-231). New York:

SAS Institute Inc. 2013. SAS/STAT. Cary, NC: SAS Institute Inc.

Tabachnick, B. G, & Fidell, L. S. (2019). Using Multivariate Statistics, 7th Edition. New York, NY: Pearson.

Van den Broek, P., 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: The Guilford Press.

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.

Windisch, H. C. (2016). How to motivate adults with low literacy and numeracy skills to engage and persist in learning: A literature review of policy interventions.

International Review of Education, 62(3), 279-297.

Wlodkowski, R. J., & Ginsberg, M. B. (2017). *Enhancing adult motivation to learn: A comprehensive guide for teaching all adults*. New York: John Wiley & Sons.

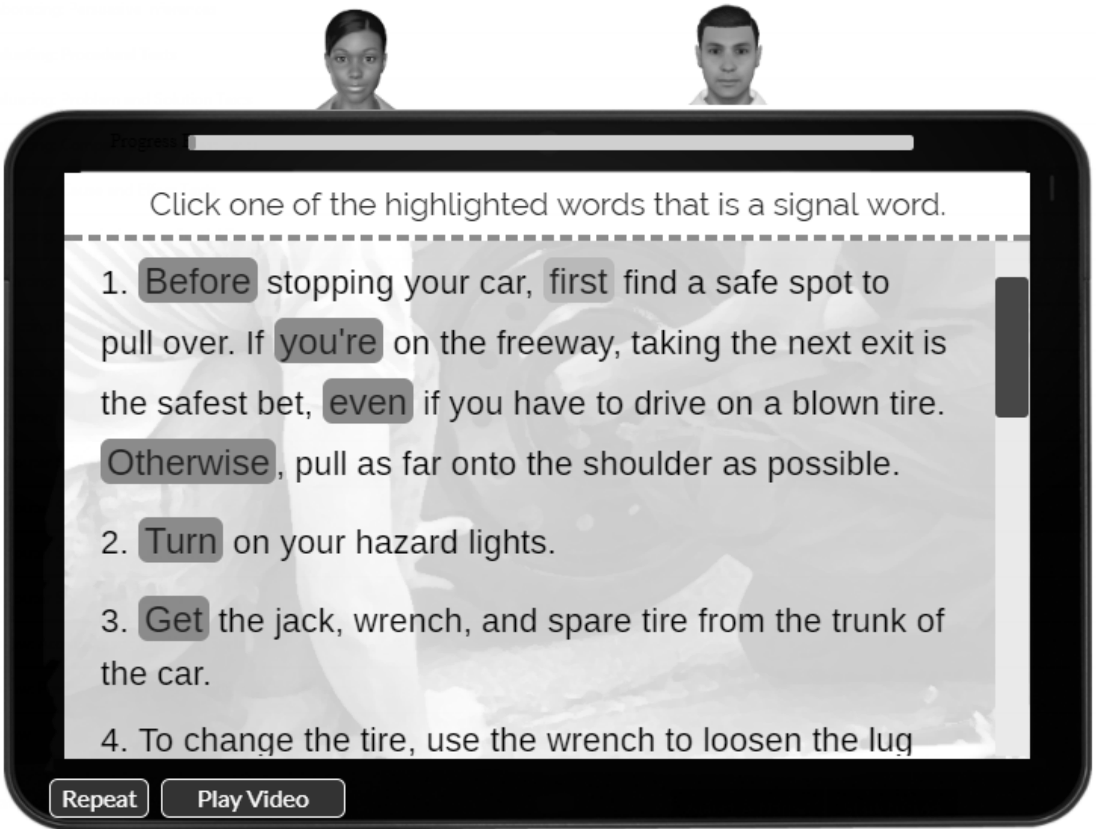
Wolters, C. A., Denton, C. A., York, M. J., & Francis, D. J. (2014). Adolescents' motivation for reading: Group differences and relation to standardized achievement. *Reading and Writing, 27*(3), 503-533.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). *Woodcock-Johnson III Normative Update Complete*. Rolling Meadows, IL: Riverside Publishing.

Table 1.

Descriptive Analysis and Mixed Effect Model on Accuracy and Time

	Word	Textbase	Situation Model	Rhetorical Structure
No. of Observations	1455	1981	5049	5071
Accuracy				
Mean Across Items	0.795	0.694	0.671	0.685
SD, Items	0.404	0.461	0.470	0.464
Model Parameter	1.66	-0.588	-0.763	-0.584
p Value	0.000	0.058	0.004	0.028
Estimated Odds	1.66	1.07	0.894	1.07
Predicted Accuracy	0.840	0.744	0.710	0.745
Time				
Mean Across Items	31.7	35.1	35.2	37.1
SD, Items	30.4	30.2	31.6	38.1
Model Parameter	2.87	2.23	2.84	3.15
p Value	--	0.804	0.716	0.694
Predicted Time	34.3	36.5	37.1	37.7



Click one of the highlighted words that is a signal word.

1. Before stopping your car, first find a safe spot to pull over. If you're on the freeway, taking the next exit is the safest bet, even if you have to drive on a blown tire. Otherwise, pull as far onto the shoulder as possible.

2. Turn on your hazard lights.

3. Get the jack, wrench, and spare tire from the trunk of the car.

4. To change the tire, use the wrench to loosen the lug

(1) Cristina (teacher agent): Tiffany, click on the words that signal to us the order in which a procedure is done.

(2) Tiffany (adult literacy student): [selects the word *first (*correct answer)]

(3) Cristina: Yeah! That is right! *Before* and *first* are signal words. These words signal to us that we must do something prior to doing something else.

Figure 1. Example screenshot and conversation for a lesson on changing a tire.

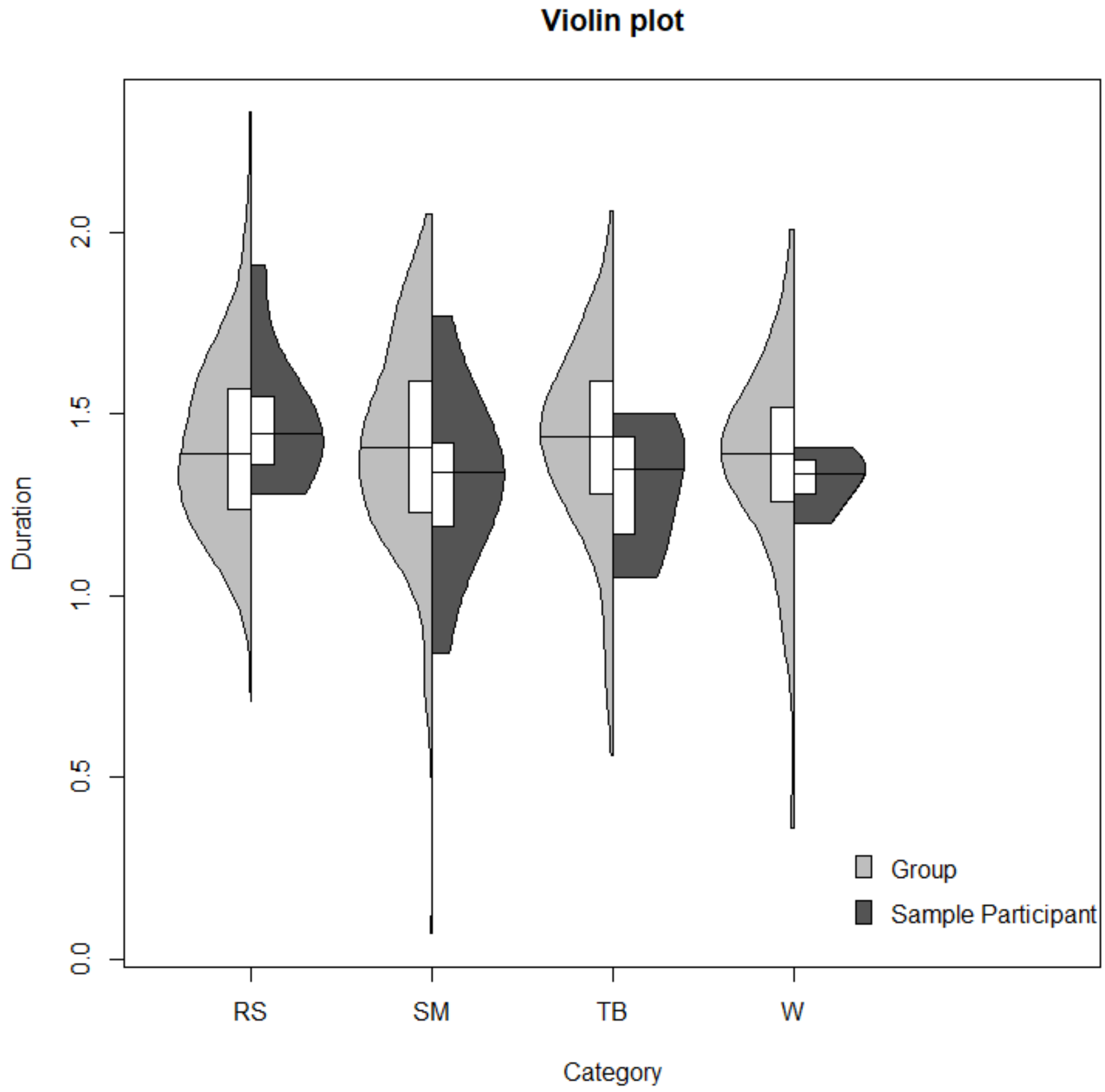


Figure 2. Violin plots depicting distributions of log10 transformed average time (higher numbers represent more time spent) per AutoTutor item across the four theoretical levels: Rhetorical Structure (RS), Situation Model (SM), Textbase (TB), and Word (W). The midline represents the median score, while the white box circumscribes the 25th and 75th percentiles. The distribution of the group as a whole is represented on the left of each plot in light grey, while the right in darker grey represents our example participant.

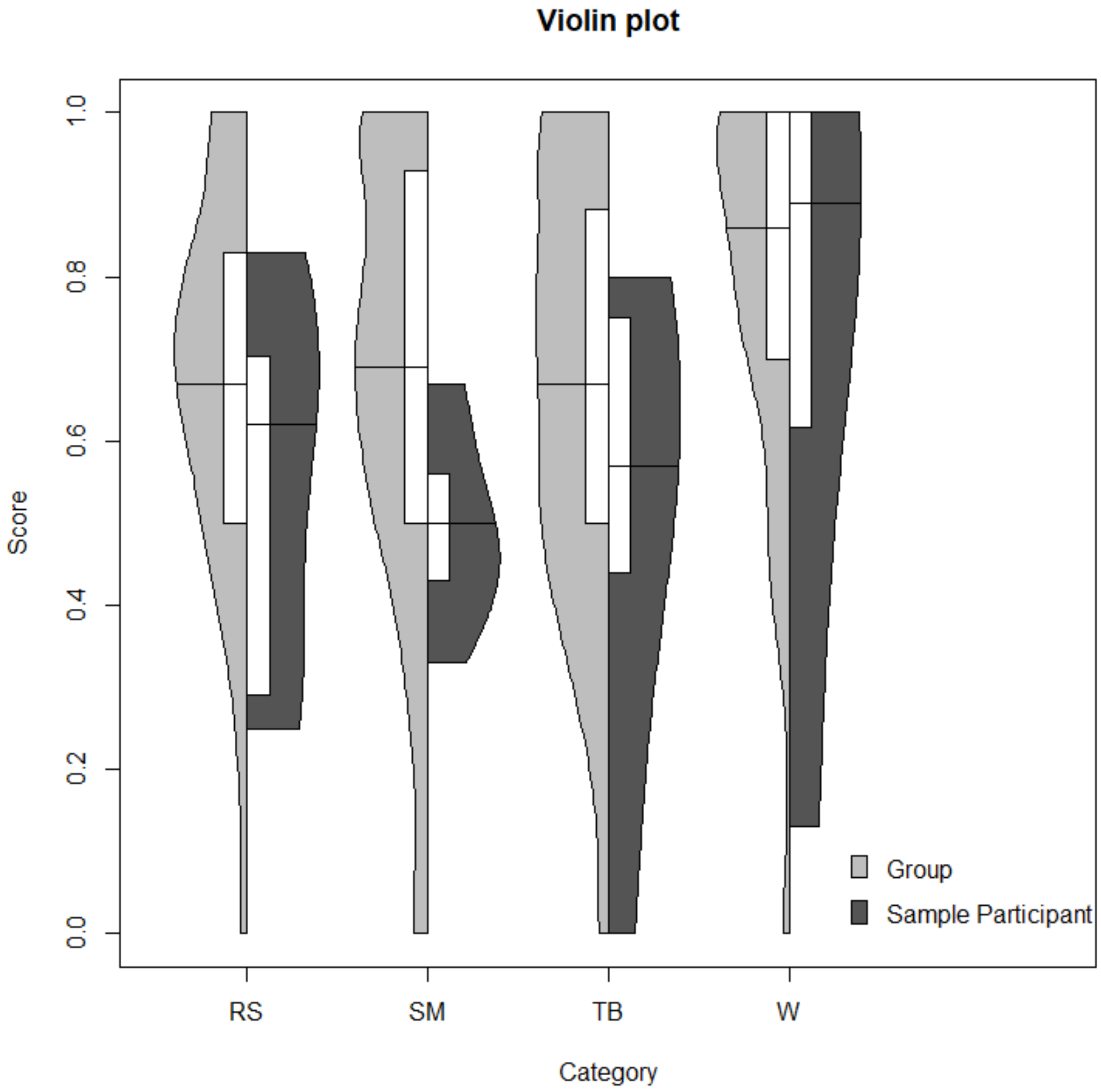


Figure 3. Violin plots depicting distributions of AutoTutor scores (Y-axis; mean proportion correct across lessons) across the four theoretical levels: Rhetorical Structure (RS), Situation Model (SM), Textbase (TB), and Word (W). The midline represents the median score, while the white box circumscribes the 25th and 75th percentiles. The distribution of the group as a whole is represented on the left of each plot in light grey, while the right in darker grey represents our example participant.