Behaviormetrika
https://doi.org/10.1007/s41237-018-0065-9

ORIGINAL PAPER

Exploring an intelligent tutoring system as a conversation-based assessment tool for reading comprehension

Genghu Shi[1] · Anne M. Lippert[1] · Keith Shubeck[1] · Ying Fang[1] · Su Chen[1] · Philip Pavlik Jr.[1] · Daphne Greenberg[2] · Arthur C. Graesser[1]

Received: 23 March 2018 / Accepted: 27 August 2018

Abstract

Reading comprehension is often assessed by having students read passages and administering a test that assesses their understanding of the text. Shorter assessments may fail to give a full picture of comprehension ability while more thorough ones can be time consuming and costly. This study used data from a conversational intelligent tutoring system (AutoTutor) to assess reading comprehension ability in 52 low-literacy adults who interacted with the system. We analyzed participants' accuracy and time spent answering questions in conversations in lessons that targeted four theoretical components of comprehension: Word, Textbase, Situation Model, and Rhetorical Structure. Accuracy and answer response time were analyzed to track adults' proficiency for comprehension components, and we analyzed whether the four components predicted reading grade level. We discuss the results with respect to the advantages that a conversational intelligent tutoring system assessment may provide over traditional assessment tools and the linking of theory to practice in adult literacy.

Keywords Adult readers · Assessment · AutoTutor · Intelligent tutoring systems · Reading comprehension

Communicated by Ronny Scherer and Marie Wiberg.

✉ Genghu Shi
gshi@memphis.edu

[1] University of Memphis, Memphis, TN 38152, USA

[2] Georgia State University Atlanta, Atlanta, GA 30302, USA

Published online: 31 August 2018

1 Introduction

Tests of reading comprehension are widely used both in and out of the classroom for a variety of reasons. Educators, practitioners, and others may desire a way to monitor progress, to detect and diagnose reading difficulties, or to test cognitive skills that underlie reading development and disorders. Each of these reasons requires an accurate assessment of comprehension. Standardized reading comprehension tests, such as Woodcock–Johnson (Woodcock et al. 2001), have historically provided a measure of comprehension ability using a series of passages and conventional assessment techniques, such as multiple choice, cloze, or spoken retelling (Fletcher 2006). These techniques offer many advantages to the extent that they are validated psychometrically, can be easily and quickly administered, and have a long history of being administered to large numbers of students. In contrast, the disadvantages of the traditional summative assessments are that they may not be particularly motivating to the test takers, which threatens the validity of the assessment, and it has been difficult to assess deeper levels of comprehension (see Sabatini et al. 2012). The ideal assessment would be motivating, efficient, and comprehensive.

Researchers in the assessment world have recently explored conversation-based assessment (CBA) that weaves test questions within conversations with computer agents to assess literacy, science, and other competencies in summative assessments (Zapata-Rivera et al. 2015). The test takers answer questions in natural language in the context of a conversation with two or more agents. When the response is incomplete or indicative of a misunderstanding, the agents ask follow-up questions to find out more what the test takers know, how deeply they know it, and sometimes why they have failed to provide a complete answer. From the standpoint of instruction and pedagogy, as opposed to summative assessment tests, a formative assessment with agents provides teachers with insights about a student's strengths and weaknesses during the process of learning. Student learning is expected to improve with instruction that is sensitive to the students' individual knowledge, skills, and abilities (Graesser et al. 2017; Woolf 2009).

Advances in the learning sciences, measurement, and electronic technologies have paved the way for a new generation of reading assessment and instruction using Intelligent Tutoring Systems (ITS). These systems have the potential of being very rich and fine grained as the actions and conversational moves are collected online during the process of learning and assessment. In particular, *AutoTutor* (Graesser 2016; Nye et al. 2014) is an ITS that helps students learn by holding a conversation in natural language. More specifically, there is a continuous multiturn tutorial conversation that occurs between an AutoTutor agent (or 2 agents) and the student during the course of answering a main question or solving a problem. If the student is incomplete or wrong, an AutoTutor agent asks additional questions or provides hints to get the student to say more or do more. The goal is to encourage students to articulate answers or perform actions that exhibit mastery, reasoning, and correct thinking rather than merely presenting information to student to comprehend. AutoTutor has been successfully implemented and

tested in university populations with adult topics. There has yet to be a systematic evaluation of its value as an assessment and instruction environment for comprehension in struggling adult readers. The present study attempts to fill this gap. In particular, we consider whether users' responses to questions in AutoTutor reflect their mastery of comprehension components that are theorized to be critical for understanding a text. In doing so, we explore how adult learners' proficiency in these comprehension components can be used as an assessment of reading difficulties and predictive of grade reading level.

1.1 AutoTutor

Intelligent tutoring systems are computerized learning environments that model a student's psychological states to provide instruction that is adaptive to these states and that advances the educational agenda (Graesser et al. 2017; Woolf 2009). When ITSs are designed with care, they promote learning gains ranging from 0.4 to 1.1 standard deviations higher than traditional classroom environments, such as reading textbooks or lecture-based instruction (Kulik and Fletcher 2016; VanLehn 2011).

AutoTutor is an example of a particular class of ITSs that use natural language conversations to model learners' knowledge. The structure of the conversations in both AutoTutor and human tutoring follows an expectation and misconception tailored (EMT, Graesser 2016) dialogue. The EMT dialogue is the primary pedagogical method of scaffolding good student answers. Each task or problem in the lesson is associated with a list of expectations (anticipated good answers, steps in a procedure) and a list of anticipated misconceptions (bad answers, incorrect beliefs, errors, bugs). As students express their answers over multiple conversational turns, the information they provide is compared with the expectations and misconceptions. Students sometimes give incorrect answers, so the tutor follows up with a hint, so they have another chance at a good answer. In addition to the main questions asked by AutoTutor, two of the common conversational moves by AutoTutor are short feedback and hints:

Feedback Feedback is a tutor agent's response to the human student's last answer which indicates the quality of the answers. The answer quality can be labeled as positive (e.g., "Excellent!", "Great answer"), negative (e.g., "Not really", "Not quite") or neutral (e.g., "I see", "Uh huh!").

Hints Hints are leading questions or statements that direct the human student to answer the main question.

Empirical evidence for AutoTutor shows learning gains of approximately 0.80 sigma (standard deviation units) when compared to non-interactive learning environments such as reading a textbook (Graesser 2016; Nye et al. 2014).

A recent version of AutoTutor has been developed by researchers in the Center for the Study of Adult Literacy (CSAL; Graesser et al. 2016a) to help struggling adult readers improve their comprehension skills. This version of AutoTutor has 30 lessons designed to help adults with low literacy improve their reading comprehension strategies.

1.2 Theoretical model underlying AutoTutor lessons

The theoretical model underlying AutoTutor lessons is based on the multilevel framework of comprehension developed by Graesser and McNamara (2011). This framework identifies six levels of reading comprehension components: *Words, Syntax*, the explicit *Textbase*, the referential *Situation Model*, the *genre/Rhetorical Structure*, and the *pragmatic communication* level. *Words* and *Syntax* represent the lower level basic reading components that include morphology, word decoding, word order and vocabulary (Perfetti 2007; Rayner et al. 2001). The other components represent the discourse components, which may be more difficult to learn. The *Textbase* level focuses on the meaning of explicit ideas in the text, but not necessarily the exact wording and syntax. The *Situation Model* (sometimes called the mental model) is a representation of the subject matter and requires inferences to be made that rely on world knowledge (Zwaan et al. 1995; Zwaan and Radvansky 1998). This model differs by text type. For example, the Situation Model corresponding to narrative text would include information about characters, settings, actions, and emotions while for informational text it would contain more technical content (e.g., knowledge and inferences about automobiles when reading a maintenance document on a truck). *Genre and Rhetorical Structure* focus on the type of discourse and its composition. *Genre* refers to the type of discourse, such as narrative, persuasive, and informational genres, as well as the subcategories of these genres. For instance, narrative encompasses folktales and novels, whereas persuasive texts include newspaper editorials and religious sermons. The *Rhetorical Structure* of a text provides the differentiated functional organization of paragraphs. In addition, there are different rhetorical frames, such as compare–contrast, cause–effect, claim–evidence, and problem–solution (Meyer et al. 2010). *Pragmatic communication* involves context-sensitive exchanges between speaker and listener, or writer and reader. The 30 AutoTutor lessons contain at least one lesson from these six theoretical levels with the exception of syntax and pragmatic communication.

AutoTutor lessons were also constructed to promote deeper metacognitive standards of comprehension in users. Metacognition refers to a person's knowledge (conscious or implicit) about his or her own cognitive processes (Hacker et al. 1998). Metacognition is thought to be an important process in improving reading comprehension (Baker 1989; Graesser 2015), yet many readers do not have adequate standards of metacognition, especially when navigating twenty-first century materials (Graesser 2015). For example, extremely shallow readers view reading comprehension as simply recognizing words in the texts; they believe they are sufficiently comprehending the text if they know the meanings of the words. Readers at the sentence level or Textbase standard of comprehension regard reading as interpreting the meaning of individual sentences in a text (Graesser 2015; van den Broek et al. 2011); they believe they are comprehending the text if the individual sentences make sense to them. In contrast, deeper readers have a discourse coherence standard that attempts to establish discourse cohesion, generate inferences to fill in cohesion gaps, and follow the structure of the discourse genre (Bohn-Gettler 2014; Van den Broek et al. 2011). The conversations embedded in AutoTutor lessons tap into these

different standards of comprehension and try to guide students toward deeper metacognitive awareness.

1.3 Characteristics of adult learners

To better understand the potential of AutoTutor as a formative assessment for reading comprehension, the present study examined the use of AutoTutor in a sample of heterogeneous adult learners (aged 16–65) with low literacy skills. Adults with low literacy skills tend to vary not only in demographic variables (age, gender, and race/ethnicity), but also in terms of educational backgrounds, learning disabilities, and primary languages (English or other) as well as their motivation for taking part in adult literacy courses (National Research Council 2011). For instance, adults with low literacy may want to improve their reading skills out of a wish to become proficient in English, as a pre-requisite for pursuing a college degree, to get promoted in career or job training, because of encouragement from family, or a desire to improve their quality of life (Malicky and Norman 1994; Tighe et al. 2013). Due to the heterogeneity of the population, more research is necessary to better understand how to tailor instruction and materials to meet the various needs of the individuals in this group.

1.4 The current study

This article reports analyses of data from adult literacy students who participated in a 100-h reading intervention that was blended between teacher-led sessions and AutoTutor. The data were recorded in AutoTutor log files throughout the 100-h intervention over approximately 4 months. There are two major goals of the current study that relate the adults' interactions with the AutoTutor environment, performance among four theoretical levels of comprehension, and reading ability in terms of grade level. First, the AutoTutor system can provide a more nuanced assessment of reading problems than a single overall performance score by relating adults' behaviors within AutoTutor to the four theoretical comprehension components. Second, by determining how the different theoretical components contribute to adults' grade level of reading, we can link theory to application and influence real-world teaching practices.

Our first goal is to determine whether accuracy and time spent on conversation-based questions in AutoTutor are indicative of comprehension. Specifically, we examined whether the adults' accuracy and time spent answering questions would vary among the four theoretical levels of comprehension. We hypothesized that adults with lower literacy skills would have higher accuracy and spend less time on questions related to the word level compared to the three discourse levels.

Our second goal is to relate adults' proficiency in different comprehension components to their overall reading ability. To this end, we mapped adults' performance within each of the four theoretical levels of comprehension onto the grade level at which they read. We defined performance for a given theoretical level as the proportion of questions within that level that were correctly answered. We measured grade

level using the Woodcock–Johnson III Passage Comprehension (Woodcock et al. 2001). The adults' accuracy within each of the four theoretical levels were expected to be systematically related to their reading grade levels.

2 Methods

2.1 Participants

The 52 participants were recruited from literacy classes of the Center for the Study of Adult Literacy (CSAL) both in Metro-Atlanta ($n=20$) and Metro-Toronto ($n=32$). The ages of participants varied from 16 to 69 with a mean of 40.0 (SD = 15.0). The majority (73%) of the participants were female. All participants read from 1.9 to 8.9 grade levels ($M=3.9$, SD = 1.6) measured by Woodcock–Johnson III Passage Comprehension (Woodcock et al. 2001). Among the participants, 30% reported they were diagnosed as learning disabled or attended special education programs when they were children. The participants consisted of 70% native English speakers. Additionally, 69% of participants received public assistance at some point. During the intervention, the participants completed an average of 71% of the 30 lessons.

2.2 Procedure

After giving informed consent to participate, participants answered demographic questions which investigated their age, gender, race/ethnicity, educational background, native language, age of English language acquisition, and whether they had ever received public assistance (e.g., food stamps).

Before the first day of the literacy classes, the Woodcock–Johnson III Passage Comprehension test (Woodcock et al. 2001) was administered to the participants to obtain their pre-test score in reading comprehension. During a period of 4 months, the participants were offered 100 h of intervention. In the AutoTutor intervention, participants first received evidence-based, teacher-led classroom instruction to decode or comprehend material presented by the computer agent. Afterwards, participants engaged in solving word or reading comprehension problems. Finally, participants were provided independent reading time.

The AutoTutor component of the intervention covered 30 curriculum lessons, and each lesson took 20–50 min to complete. One concern was that the teacher-led instruction could impact, either positively or negatively, the participants' performance within AutoTutor. However, our previous analysis (Shi et al. 2017) showed that participants made no significant learning gains within each of the four theoretical levels during the period of the intervention. We thus felt confident that the participants' performance on the four theoretical levels reflected their pre-existing reading abilities. All participants received the tests of their reading abilities as the compensation for being part of the study and the study was approved by the Institutional Review Board of Georgia State, University of Memphis, Brock University, and University of Toronto.

2.3 Measures

The version of AutoTutor presented to participants consisted of 30 lessons designed to help adults with low literacy improve their reading skills, especially skills required for the comprehension of text. The system is adaptive in the sense that many lessons have easy, medium, versus difficult materials (words and texts), which were measured by Coh-Metrix (Graesser et al. 2014b), and students are assigned different levels of materials based on previous responses. Coh-Metrix is a computer tool that analyzes many different linguistic features of words, sentences, and multi-sentence texts. For the texts in the CSAL lessons, for example, Coh-Metrix computes text formality, which is a comprehensive score of the difficulty level of a text.

Within most of the AutoTutor lessons, the students first received materials at a medium level of difficulty and answer 8–12 questions that were embedded in conversation with the computer agents. Depending on the student's performance on these questions, the student received either the easier or harder material next. That is, higher performance on the medium materials would lead to more difficult material whereas lower performance on the medium material would lead to easier material. Since writing is generally problematic for adults with low literacy skills (Olney et al. 2017), the interactions in AutoTutor for CSAL are largely point-and-click, multiple-choice questions, or drag-and-drop. However, it is allegedly the conversational component that drives learning, scaffolded in AutoTutor through EMT-structured conversations.

Instead of a simple dialogue between the tutor agent and the human learner, the conversation in AutoTutor for CSAL has two agents and one human learner who participate in *trialogues* (Graesser et al. 2014a; see also Johnson et al. 2017). Trialogues offer several affordances appropriate for adult learners with low domain ability that are not available to dialogues. For example, in a trialogue, adult learners can learn vicariously by observing interactions between a student and tutor agent so that learning is possible even with minimal skills. A peer agent and human student may share a misconception that can be presented to the tutor by the peer agent. The tutor agent, in turn, gives negative and corrective feedback to the peer agent which prevents or reduces the potentially negative motivational impact upon the human learner with negative feedback. Trialogues also facilitate competition between the adult learner and a peer agent in a game setting, which can be highly motivating (Graesser et al. 2016b). For example, Fig. 1 depicts a game scenario in AutoTutor where the adult learner competes with the peer agent to correctly answer questions and a cumulative score is kept. The game is designed to always allow the adult learner to win to promote self-esteem and self-efficacy. The goal in Fig. 1 is to figure out the meaning of the target word, which has multiple meanings, in the sentence of a "Word" lesson. When a learner answers incorrectly or does not provide a complete answer, the EMT trialogue kicks in and the learner receives a hint from one of the two agents, providing another chance with somewhat more guidance.

The AutoTutor data collected in the log files of each participant included participant ID, the number of times a lesson was attempted, the number of times a question within a lesson was attempted, the accuracy of answering a question when first attempted (0 or 1), the time in seconds to answer a question in the lesson when first

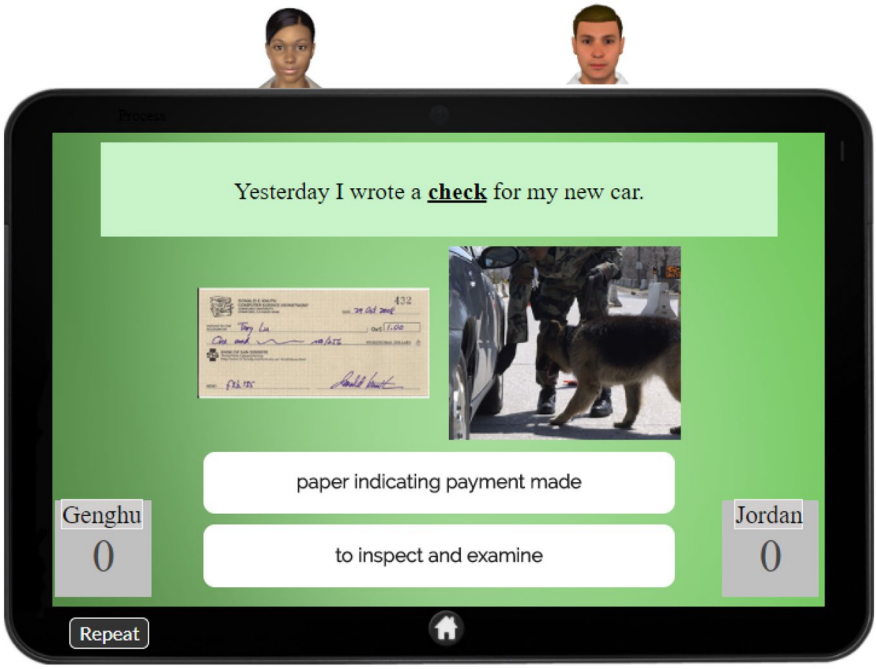


Fig. 1 AutoTutor interface depicting a friendly game between the adult learner and a peer agent

attempted, the difficulty of the materials (easy, medium, hard), and the theoretical levels that a lesson addressed. There were other measures collected but they are not relevant to the present article. The data were stored in a data management system on a central computer sever that handled all of the participants' data that were collected on personal computers on the web.

The present study analyzed a subset of the data that were collected. We extracted data for each participant for each of the questions within the 29 lessons that were completed; we eliminated the single lesson that focused on Syntax because there was only one lesson in that category. If the participant did not complete a particular lesson, all of the observations of the lesson were deleted from the analyses. We further reduced the dataset by considering accuracy and times for questions pertaining only to medium-level words or texts. All participants would at minimum receive questions at a medium difficulty level at the beginning of the lesson (and for some lessons the complete lessons) to ensure that all of the participants had the opportunity to contribute data to the same set of questions. Thus, we used only data corresponding to questions on medium levels of difficulty. On average, the participants completed 23 lessons (ranging from 2 to 29 lessons), and each lesson contained 14.6 questions (medium level), with a range from 6 to 30 questions.

The medium level of the lessons varied in the number of texts the user interacted with and the length of the texts. The medium level for lessons whose primary comprehension component focused on discourse (i.e., not Words) typically consisted of a single text, such as an informational article or fictional story, around 250–300

words in length. The data from these lessons included the responses to a set of sequentially presented questions regarding the text. For instance, the medium level text for the lesson *Compare and Contrast* discusses differences and similarities in the athletic careers of Kobe Bryant and Michael Jordan. Each question addressed some aspect of this single text. The medium level for lessons whose primary comprehension component was *Words* contained multiple stimuli in which a single question was asked per stimulus. For example, in the medium level of the lesson *Multiple Meaning Words,* participants were presented a series of questions in a fixed order. For each question (as illustrated in Fig. 1), the participants were shown a sentence (around 8–20 words in length) containing a word with multiple meanings (e.g., bank, check, etc.) and were asked to choose the correct definition of the word based on the context of the sentence. In this way, data for this lesson corresponded to a single response generated by a single stimulus.

When we examined the distributions of the resulting data, we found that response time per question was positively skewed, which is typical for response time data. To reduce the bias brought by the potential outliers, we truncated the time by replacing the observations beyond three standard deviations above the mean for the subject with the value at three standard deviations above the mean for the subject; this truncation was performed for each participant separately.

We defined accuracy as the score (1 as correct, 0 as incorrect) the participant received on the first attempt on each question of a lesson.

We defined time on a question as the number of seconds it took for a participant to answer a particular question, from the onset of the question to the participants' click on an answer. Time was assumed to be a relevant indicator of reading proficiencies. The participants were unaware that the time they spent on answering questions would be assessed.

Each lesson tapped 1–3 of the four theoretical levels, i.e., Word, Textbase, Situation Model, and Rhetorical Structure. We assigned a measure of the relevance of each of the four theoretical components to each of the lessons. We defined the relevance of a theoretical level on a lesson as the extent to which the level was tapped in the lesson. The assigned codes were primary, secondary, tertiary or no relevance of a component to a lesson. We quantified the orderings so that components with primary relevance for a lesson received a value of 1.00, secondary relevance received a value of 0.67, tertiary relevance received a value of 0.33, and no relevance received a value of 0.00. Table 1 shows the 29 AutoTutor lessons and the relevance of each theoretical component for each lesson. The columns named *W*, TB, SM, and RS designate the measure of relevance for *Word*, *Textbase*, *Situation Model*, and *Rhetorical Structure*, respectively, for each lesson. The *levels* column summarizes this information, listing the components that are relevant for each lesson in order of relevance (i.e., the first component listed is the most relevant). For example, *Stories 1* addresses aspects of comprehension primarily at the level of *Situation Model* (1.00), then *Textbase* (0.67), and lastly *Rhetorical Structure* (0.33), but not *Word* (0.00).

We also categorized questions on one of the four theoretical levels. The category for a particular question within a particular lesson was simply the primary theoretical level that characterized the lesson (Table 1). For example, we see in Table 1 that questions in *Compare and Contrast* belonged to the category *RS* since this lesson

Table 1 Relevance of comprehension components, average accuracy, and average time spent per question for each of the 29 AutoTutor lessons

							Accuracy		Time (s)	
Lesson name	*N*	Theoretical levels	*W*	TB	SM	RS	Mean	SD	Mean	SD
Word parts	46	*W*	1.00	0.00	0.00	0.00	0.763	0.297	30.0	19.7
Word meaning clues	34	*W*	1.00	0.00	0.00	0.00	0.690	0.180	46.8	25.7
Learning new words	46	*W*	1.00	0.00	0.00	0.00	0.808	0.196	31.1	14.6
Multiple meaning words	47	*W*, TB	1.00	0.67	0.00	0.00	0.822	0.119	22.0	12.2
Pronouns	48	TB, *W*	0.67	1.00	0.00	0.00	0.749	0.223	38.6	23.1
Punctuation	47	TB, SM	0.00	1.00	0.67	0.00	0.606	0.217	29.7	10.1
Non-literal language	43	SM	0.00	0.00	1.00	0.00	0.726	0.120	27.2	10.4
Text signals	49	SM	0.00	0.00	1.00	0.00	0.737	0.216	31.3	13.4
Purpose of texts	49	RS	0.00	0.00	0.00	1.00	0.581	0.143	30.8	15.0
Key information	38	TB, SM	0.00	1.00	0.67	0.00	0.761	0.122	21.3	12.0
Main ideas	39	TB, RS	0.00	1.00	0.00	0.67	0.654	0.200	47.1	23.1
Claims versus support	42	RS, SM	0.00	0.00	0.67	1.00	0.620	0.184	27.4	19.7
Connecting ideas	43	SM, TB, RS	0.00	0.67	1.00	0.33	0.587	0.200	74.3	17.1
Stories 1	46	SM, TB, RS	0.00	0.67	1.00	0.33	0.691	0.289	16.5	13.2
Stories 2	35	SM, TB	0.00	0.67	1.00	0.00	0.824	0.189	37.6	18.3
Story maps	44	SM, RS	0.00	0.00	1.00	0.67	0.566	0.164	35.5	14.2
Persuasion 1	41	TB, RS	0.00	1.00	0.00	0.67	0.651	0.144	43.6	19.1
Persuasion 2	32	SM, TB	0.00	0.67	1.00	0.00	0.691	0.182	38.4	17.4
Steps in procedures	39	RS, TB, SM	0.00	0.67	0.33	1.00	0.789	0.107	38.8	20.3
Problems and solutions	40	RS, TB, SM	0.00	0.67	0.33	1.00	0.676	0.168	25.0	12.6
Compare and contrast	33	RS, TB, SM	0.00	0.67	0.33	1.00	0.815	0.178	73.1	42.6
Cause and effect	43	RS, TB, SM	0.00	0.67	0.33	1.00	0.571	0.166	56.6	25.5
Describing things	43	RS, TB, SM	0.00	0.67	0.33	1.00	0.758	0.123	44.4	17.4
Time and order	37	RS, TB, SM	0.00	0.67	0.33	1.00	0.806	0.163	28.9	16.7
Inferences from texts	36	SM, TB	0.00	0.67	1.00	0.00	0.680	0.198	43.2	24.8
Forms and documents	31	SM, TB	0.00	0.67	1.00	0.00	0.487	0.136	42.5	17.8
Review 1	22	SM, *W*	0.67	0.00	1.00	0.00	0.734	0.192	25.0	12.9
Review 2	44	SM, TB, RS	0.00	0.67	1.00	0.33	0.664	0.154	26.1	13.2
Review 3	40	RS, TB, SM	0.00	0.67	0.33	1.00	0.686	0.215	33.9	18.8

The right four columns give means and standard deviations on accuracy and time of adult learners per lesson

W, *TB*, *SM*, and *RS* stand for word, textbase, situation model, and rhetorical structure, respectively. *N* is the number of adults who completed each lesson

primarily focused on aspects of Rhetorical Structure. We then defined performance for each theoretical level as the average accuracy on questions within that theoretical level.

Our measure of grade reading level of a participant was the Woodcock–Johnson III Passage Comprehension subtest (Woodcock et al. 2001), which had been administered prior to the AutoTutor intervention. This comprehension subtest is

a complex and conceptually driven processing task that measures the ability to produce the mental representations of the text during reading. In the test, participants silently read passages and fill in the missing word. The reliability is 0.83 for ages 5–19 and 0.88 for adults (McGrew and Woodcock 2006).

2.4 Data analyses

To compare the differences in accuracy and time among the four theoretical levels, we first computed descriptive statistics on the accuracy and time data for the four theoretical components. Any trends we observed from the descriptive analysis were further investigated using mixed effect models (Bates et al. 2014). We used a logistic mixed effect model to predict question accuracy (1: correct, 0: incorrect) and a linear mixed effect model to predict time spent on a question (in seconds). Item (question) was the unit of analysis for both models. Our rationale for using a linear mixed effect model instead of an ANOVA for analysis was to avoid the language-as-fixed-effect fallacy, or more properly, the stimuli-as-fixed-effect fallacy (Baayen et al. 2008; Clark 1973). The AutoTutor curriculum contains different lessons that address different comprehension levels, as well as different questions in each lesson. It is not appropriate to assume the independence among the observed outcomes (accuracy and time) since participants completed multiple questions within multiple lessons for a specific comprehension level. In addition, the variability in accuracy (and/or time) was mixed with variability among subject (participants), lessons, as well as items (questions) for each comprehension level (fixed effect). Thus, to test whether the accuracy (and/or time) differs among the four theoretical levels of comprehension (fixed effect), subject (participants), item (question), and lesson were added into the linear mixed effect models as random intercepts. We also included random-subject (participant) slopes on theoretical levels, and random effects for specific questions nested within lessons in the model since participants' performance might vary on different theoretical levels, and questions were designed to be nested in lessons. To confirm the results of mixed effect models, a follow-up correlational analysis was performed between the four continuous measures of the theoretical levels and adult learners' mean accuracies and average time per question (in seconds) on the 29 lessons (in Table 1). Through this approach, we will determine whether adult learners' accuracy and time measures on the 29 lessons are correlated with the reading comprehension components that the lessons focused on.

Analyses were also conducted on the relationship between performance on four theoretical levels and reading grade level. We used multiple linear regression to predict reading grade level from accuracy on each of the four theoretical levels. Prior to this, we computed the correlation matrix on these variables and tested potential multicollinearity among the performance of the four theoretical levels. We used R programing language (R Core Team 2013) to carry out all aspects of our data analysis.

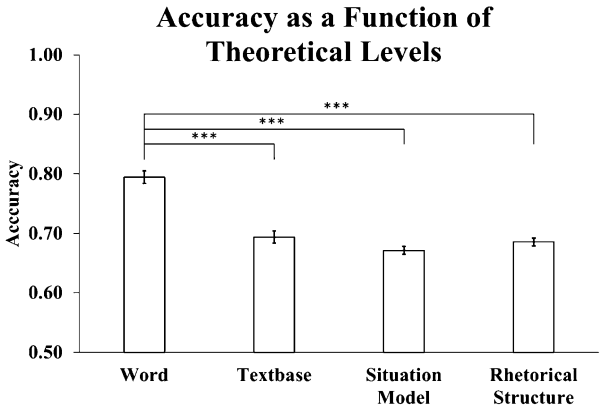
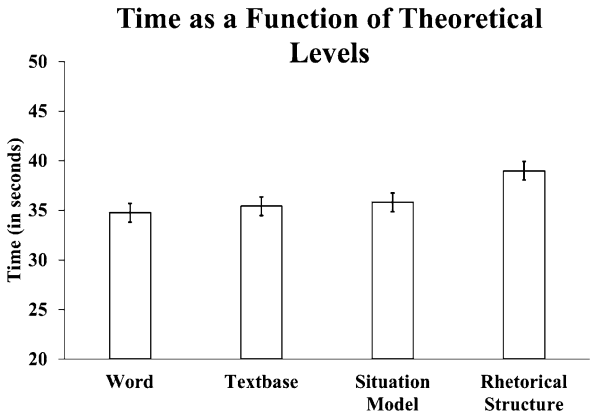


Fig. 2 Means and error bars of accuracy on four theoretical levels. The symbol "*" indicates the significance level between two categories in the figure. ***indicates $p < 0.001$

Fig. 3 Means and error bars of time (in second) on four theoretical levels



3 Results

3.1 Differences in participants' accuracy and time by comprehension levels

Our hypothesis that accuracy and time would vary between theoretical components was partially supported. In particular, we expected that participants would be more accurate and spend less time on Word level questions compared to the other three discourse levels. The means and standard errors of accuracy and time for questions within each of the four theoretical level categories are shown in Figs. 2 and 3. Figure 2 shows that adults had the highest accuracy for *Word* level questions compared to the three deeper discourse levels. Figure 3 appears to show longer times for Rhetorical Structure questions, but statistically analyses needed to be conducted to confirm that.

The results of the logistic and linear mixed effects models provide partial support of our hypothesis that time and accuracy would vary between theoretical components. In particular, our results indicated that accuracy, but not time, differs among theoretical levels. Table 2 shows results of the logistic (upper half of Table 2) and linear (lower half of Table 2) models that compared the differences in accuracy and time, respectively, among the four theoretical levels. For the model using theoretical level to predict accuracy, we see the estimated odds ratio (estimated odds) of *Word* level (*Intercept*, the reference level in the model) is significantly higher than each of the three discourse levels (*Textbase, Situation Model, Rhetorical Structure*). After using a log function to convert the odds ratios to predicted accuracies of adult learners, we can say the predicted accuracy of adult learners on *Word* level was higher than each of the three discourse levels. We used a Type II Wald χ^2 test to check for differences in accuracy among the four theoretical levels. The results indicated a significant difference ($\chi^2(3) = 8.34$, $p = 0.040$) among the four theoretical levels, so we conducted a post hoc analysis using pairwise comparisons. This analysis revealed differences occurred for only pairs involving the *Word* level (all *p values* were less than 0.001), suggesting that compared to other theoretical levels, the learner's accuracy on the Word level is unique.

Our linear mixed effect model, however, did not support the idea that time to answer questions varied among comprehension components. Results of a type III ANOVA with a Satterthwaite approximation indicated that time did not differ significantly among the four theoretical levels ($F(3, 25.8) = 0.058$, $p = 0.981$). Thus, even though the descriptive analysis suggested that adults use the least time to answer Word questions, and the most time on *Rhetorical Structure* questions, the linear mixed effects model did not quite support this trend.

Table 3 shows the results of the follow-up correlation analysis between mean accuracy, mean time, and comprehension components for each of the 29 AutoTutor lessons. Here we see that mean accuracy was positively and significantly correlated with Word level ($r = 0.386$, $p < 0.05$), but not associated with any of the discourse

Table 2 Results of mixed effects models (accuracy and time)

	Word (Intercept)	Textbase	Situation model	Rhetorical structure
No. of items	1455	1981	5049	5071
Accuracy				
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
Time				
Model parameter	34.3	2.23	2.84	3.15
P value	0.000	0.804	0.716	0.694
Predicted time	34.3	36.5	37.1	37.7

Model parameter represents the estimates of Intercept, Textbase, Situation Model, and Rhetorical Structure by the models in which Word was the base (Intercept), and the other three levels compared to the base. The same condition was for time

Table 3 Correlations among the four theoretical relevance measures and the accuracy and time per question on lessons

	W	TB	SM	RS	Time
Word					
Textbase	− 0.365*				
Situation Model	− 0.485*	− 0.084			
Rhetorical Structure	− 0.467*	0.098	− 0.318		
Time	− 0.168	0.207	− 0.088	0.236	
Accuracy	0.386*	0.009	− 0.253	− 0.142	− 0.188

Time and accuracy represent adult learners' mean accuracies and time on 29 lessons

W, TB, SM, and RS refer to Word, Textbase, Situation model, and Rhetorical structure

levels (*Textbase, Situation Model, Rhetorical Structure*). In contrast, average time was not significantly correlated with any of the four theoretical levels. These patterns of correlational analysis reinforced the results of the mixed effect models on accuracy and time measure. Additionally, the Word level was negatively and significantly correlated with each of the three discourse levels (*Textbase, Situation Model, Rhetorical Structure*).

3.2 Using accuracy on theoretical levels to predict grade level

A linear multiple regression analysis was conducted to predict adults' reading grade level with their accuracy on questions corresponding to the four theoretical levels. Prior to conducting this analysis, we computed correlations among adult learners' accuracy on the four theoretical levels and their reading grade level. The lower triangular matrix in Table 4 shows that adult learners' accuracies on the four theoretical levels were significantly correlated with their reading grade level, and also with each

Table 4 Reading grade levels related to the four theoretical levels (n = 50)

	Zero-order *r*					*B*	sr^2	*b*
Variable	RS	SM	TB	*W*	RGL			
Word					0.32*	0.15*	0.02	2.04
Textbase				0.35*	0.36**	− 0.09	0.00	− 1.11
Situation Model			0.66***	0.30*	0.42**	0.32***	0.06	5.15
Rhetorical Structure		0.38**	0.59**	0.35*	0.44**	0.32***	0.07	6.94
						Intercept = − 5.19*		
M	0.684	0.680	0.691	0.785	3.89	$F(4, 45) = 4.64$**		
SD	0.078	0.102	0.133	0.119	1.67	$R^2 = 0.292$**		

W, *TB*, *SM*, and *RS* refer to adult learners' accuracy on Word, *Textbase*, *situation model*, and *Rhetorical structure*, respectively. *RGL* refers to Reading Grade level.. indicates $p < 0.1$, * indicates $p < 0.05$; ** indicates $p < 0.01$; ***indicates $p < 0.001$

Indicates $p < 0.1$, * indicates $p < 0.05$; ** indicates $p < 0.01$; ***indicates $p < 0.001$

other. These moderately high correlations might signal multicollinearity when using adult learners' accuracies on the four theoretical levels to predict their grade levels with a multiple regression model. Therefore, we performed the *Variance Inflation Factor* (VIF) test on the model. A VIF value under 4 (or a tolerance value above 0.25) has been used as a rule of thumb to reject the hypothesis of multicollinearity (O'Brien 2007). The values met the criteria of VIF and tolerance.

Our hypothesis that participants' accuracy within each of the four theoretical levels would predict their reading grade levels was supported. A significant regression equation was found with an R^2 of 0.292. The output of the multiple linear regression model gave the two-sided p values which tested whether the betas were equal to zero or not. In the study, we were only curious whether the contribution of the adults' accuracies on the four theoretical levels were greater than zero. Therefore, we performed a one-sided test on the parameters of the model, and these results are shown in Table 4. Looking at Table 4, we see *Word*, *Situation Model*, and *Rhetorical Structure* were significant predictors of adult learners' reading grade level. Notably, the coefficient of *Textbase* was negative and not significant in either the one-sided test or two-sided test. When we removed *Textbase* from the multiple linear regression model, the coefficients (b) of Word, Situation Model, and Rhetorical Structure decreased to 1.95, 4.43, and 6.22, respectively, while the R^2 also decreased to 0.289. Therefore, this non-significant negative result reflected a suppression effect (Ludlow and Klein 2014; Thompson and Levine 1997).

4 Discussion

This paper presents a systematic evaluation of AutoTutor as a potential assessment tool for comprehension in adults with low literacy. We were particularly interested in the use of AutoTutor as a more nuanced assessment tool since it provides measures for four different theoretical levels of comprehension rather than just a single, general measure. To this end, our study aimed to (1) determine whether participants' accuracy and time spent on questions in AutoTutor could indicate reading comprehension ability within the four comprehension levels, and (2) to see if participants' accuracy within each of the four theoretical levels would predict their reading grade levels.

Our results concerning the first goal were mixed since they suggest accuracy, but not time, can be used to track adults' proficiency for different components of comprehension. For accuracy, we found that participants were more accurate answering *Word* level questions than for questions on discourse levels (*Textbase, Situation Model,* and *Rhetorical Structure*). One explanation for this result is that questions for the *Word* level focus on individual words or single sentences which require low loads on working memory. Solving the context-based questions of deep discourse levels is time-consuming, may require complex strategies, and is often taxing on cognitive resources (Carretti et al. 2009; Cutting and Scarborough 2006).

With respect to time, descriptive statistics suggested that adults were somewhat slower to answer questions for the discourse levels requiring deeper comprehension. In other words, time to answer questions showed the following pattern:

Word < *Textbase* < *Situation Model* < *Rhetorical Structure*. However, mixed effect modeling did not validate this trend and instead suggested time spent on questions is not affected by the comprehension component being tested. Time studies have indicated that time spent might be affected by confusion, motivation, personality, cognitive abilities, fluency, actual performance, and other variables (Goldhammer et al. 2015), but apparently not theoretical level in this study. The random effects of participants were taken into consideration in mixed effect model which ruled out most of the individual difference from the results. Perhaps clustering analyses on the participants would show systematic individual differences in the patterns of times among different reader clusters. The modest sample size most likely accounts for the small impact of theoretical level on time to answer questions.

Correlations between the four theoretical levels with time and accuracy confirmed these results. In particular, we found no significant correlation between time and theoretical levels but did find that adult learners were more accurate on lessons that focused mainly on the Word level of comprehension. Of additional interest is that accuracy for *Word* level was negatively correlated with accuracy on the three other theoretical levels. This gives credence to the theory that proficiency on basic reading processes and deeper discourse levels may be separable (Graesser et al. 2011).

Our second goal was to investigate whether participants' accuracy within each of the four theoretical levels could predict their reading grade levels. We found some evidence to confirm this expectation. A multiple regression analysis indicated that participants with higher accuracy for *Word*, *Situation Model*, and *Rhetorical Structure* level questions had higher reading grade levels. In addition, the analysis suggested a suppression effect whereby Textbase overlapped other theoretical levels. Accuracy on *Textbase* level questions did not predict the reading grade level after statistically removing the contributions of the other theoretical levels. One possible interpretation of the *Textbase* data is that the Woodcock–Johnson reading test does not tap the *Textbase* theoretical level as a component of reading comprehension. Another explanation is that there was a suppression effect because the *Textbase* component was highly correlated with *Situation Model* ($r = 0.66$) and *Rhetorical Structure* ($r = 0.59$).

The present work is important for a number of reasons. We explored the potential of an ITS to assess levels of reading comprehension in adults who have low literacy skills, which, to our knowledge, is the first study of its kind. In doing so, we offer researchers a new assessment approach that is efficient and capable of providing a more complete story about reading comprehension deficiencies than what is relayed using a single overall performance score.

By assessing proficiency on four major theoretical components of comprehension, we have taken a first step toward the idea of establishing performance norms that address individual aspects of comprehension. Theories of reading comprehension assume there are multiple levels of comprehension (e.g., Graesser and McNamara 2011), but existing assessments do not address deficiencies that occur within each of these theoretical components. There are other potential benefits of segregating the status of the particular theoretical components. It is conceivable that a low comprehension score could be increased by teaching strategies that cater to the specific components that particular readers struggle with. An instructional system that

tailors training to particular reader deficits would be expected to increase learning efficiency and learner engagement. Students will not be re-learning previously mastered material and teachers will not spend time on strategies that fail to deal with the precise deficits. Once norms are established, scientists can use them to develop more personalized curriculum for different populations of readers that can be used both in traditional classroom settings and in ITSs. Thus, future work should continue to collect data that could be used to establish ranges of proficiency across all six theoretical levels of comprehension.

We also explored the relationship between theoretical levels of comprehension and reading grade level, which offers a venue for putting theory into practice. Since we showed how accuracy on the theoretical levels of comprehension can be directly related to a reading grade level, learning gains assessed through an ITS can be used by teachers in traditional, real-world educational settings. This can help integrate ITSs into schools that may be concerned about having to spend time and money on training to interpret unfamiliar outputs.

Despite the many strengths of the study, we realize there may be potential concerns regarding the methodology. For example, the sample size used for the study was modest ($N=52$). This may have contributed to our inability to find differences in time to answer questions concerning different theoretical components. Ideally, future work should use a larger sample to increase power and the ability to detect an effect of theoretical components on response time. We also note that the data we used corresponded to a set of texts that did not have large variations in difficulty. We are unsure if we would find similar results if we considered easy-, medium-, and hard-level texts. For instance, there may be an interaction effect such that time and accuracy depend on both the text difficulty level and theoretical level being tested, but further work is needed to see if this is the case.

5 Implication

In summary, we have shown that an ITS can offer an efficient and thorough route to assessing reading comprehension in adults. This type of assessment goes beyond giving a total correct score in its capacity to differentiate students' abilities for specific levels of comprehension. Both teachers and students will presumably benefit from a more discriminating assessment tool. For example, a student who needs help with the Textbase level of comprehension but not with the Word level will primarily or exclusively receive this type of tailored instruction if the teacher has access to this information. Intelligent tutoring systems can provide a way to track both time and accuracy on different types of items, and a way to systematically conduct a nuanced assessment of comprehension abilities. This is particularly important for adults with low literacy skills who struggle with traditional types of interventions, and who may benefit from the more personalized, component-specific interventions that an ITS can provide.

Acknowledgements The research reported here is supported by the Institute of Education Sciences, US Department of Education, through Grant R305C120001, and the National Science Foundation Data

Infrastructure Building Blocks program under Grant no. (ACI-1443068). The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education and the National Science Foundation.

References

Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 59(4):390–412

Baker L (1989) Metacognition, comprehension monitoring, and the adult reader. Educ Psychol Rev 1(1):3–38

Bates D, Mächler M, Bolker B, Walker S (2014) Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823

Bohn-Gettler CM (2014) Does monitoring event changes improve comprehension? Discourse Process 51:398–425

Carretti B, Borella E, Cornoldi C, De Beni R (2009) Role of working memory in explaining the performance of individuals with specific reading comprehension difficulties: a meta-analysis. Learn Individ Diff 19(2):246–251

Clark HH (1973) The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. J Verbal Learn Verbal Behav 12(4):335–359

Cutting LE, Scarborough HS (2006) Prediction of reading comprehension: relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. Sci Stud Read 10(3):277–299

Fletcher JM (2006) Measuring reading comprehension. Sci Stud Read 10(3):323–330

Goldhammer F, Naumann J, Greiff S (2015) More is not always better: the relation between item response and item response time in Raven's matrices. J Intell 3(1):21–40

Graesser AC (2015) Deeper learning with advances in discourse science and technology. Policy Insights Behav Brain Sci 2(1):42–50

Graesser AC (2016) Conversations with AutoTutor help students learn. Int J Artif Intell Educ 26(1):124–132

Graesser AC, McNamara DS (2011) Computational analyses of multilevel discourse comprehension. Top Cognit Sci 3(2):371–398

Graesser AC, McNamara DS, Kulikowich J (2011) Coh-Metrix: providing multilevel analyses of text characteristics. Educ Res 40(5):223–234

Graesser AC, Li H, Forsyth C (2014a) Learning by communicating in natural language with conversational agents. Curr Dir Psychol Sci 23(5):374–380

Graesser AC, McNamara DS, Cai Z, Conley M, Li H, Pennebaker J (2014b) Coh-Metrix measures text characteristics at multiple levels of language and discourse. Elem Sch J 115(2):210–229

Graesser AC, Cai Z, Baer WO, Olney AM, Hu X, Reed M, Greenberg D (2016a) Reading comprehension lessons in AutoTutor for the Center for the Study of Adult Literacy. In: Crossley SA, McNamara DS (eds) Adaptive educational technologies for literacy instruction. Taylor & Francis Routledge, New York, pp 288–293

Graesser AC, Hu X, Nye B, Sottilare R (2016b) Intelligent tutoring systems, serious games, and the generalized intelligent framework for tutoring (GIFT). In: O'Neil HF, Baker EL, Perez RS (eds) Using games and simulation for teaching and assessment. Routledge, Abingdon, pp 58–79

Graesser AC, Rus V, Hu X (2017) Instruction based on tutoring. In: Mayer RE, Alexander PA (eds) Handbook of research on learning and instruction. Routledge Press, New York, pp 460–482

Hacker DJ, Dunlosky J, Graesser AC (eds) (1998) Metacognition in educational theory and practice. Erlbaum, Mahwah

Johnson AM, Guerrero TA, Tighe EL, McNamara DS (2017) Confronting adult low literacy with intelligent tutoring for reading comprehension. In: Andre E, Baker R, Hu X, Rodrigo M, du Boulay B (eds) International conference on artificial intelligence in education. Springer Cham, pp 125–136

Kulik JA, Fletcher JD (2016) Effectiveness of intelligent tutoring systems: a meta-analytic review. Rev Educ Res 86(1):42–78

Ludlow L, Klein K (2014) Suppressor variables: the difference between 'is' versus 'acting as'. J Stat Educ 22(2):1–28

Malicky G, Norman C (1994) Participation patterns in adult literacy programs. Adult Basic Educ 3(3):144–156

McGrew KS, Woodcock RW (2006) Woodcock–Johnson III technical manual: WJ III. Riverside Publishing, Rolling Meadows

Meyer BJF, 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. Read Res Q 45(1):62–92

National Research Council [NRC] (2011) Improving adult literacy instruction: options for practice and research. The National Academies Press, Washington

Nye BD, Graesser AC, Hu X (2014) AutoTutor and family: a review of 17 years of natural language tutoring. Int J Artif Intell Educ 24(4):427–469

O'Brien RM (2007) A caution regarding rules of thumb for variance inflation factors. Qual Quant 41(5):673–690

Olney AM, Bakhtiari D, Greenberg D, Graesser A (2017) Assessing computer literacy of adults with low literacy skills. In: Hu X, Barnes T, Hershkovitz A, Paquette L (eds) Proceedings of the 10th international conference on educational data mining. International Educational Data Mining Society, Wuhan, pp 128–134

Perfetti CA (2007) Reading ability: lexical quality to comprehension. Sci Stud Read 11(4):357–383

Rayner K, Foorman BR, Perfetti CA, Pesetsky D, Seidenberg MS (2001) How psychological science informs the teaching of reading. Psychol Sci Public Interest 2(2):31–74

Sabatini JP, O'Reilly T, Albro E (eds) (2012) Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences. R&L Education, Lanham

Shi G, Pavlik Jr P, Graesser A (2017) Using an additive factor model and performance factor analysis to assess learning gains in a tutoring system to help adults with reading difficulties. In: Hu X, Barnes T, Hershkovitz A, Paquette L (eds) Proceedings of the 10th international conference on educational data mining. EDM Society, Wuhan, pp 376–377

R Core Team (2013) R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Thompson FT, Levine DU (1997) Examples of easily explainable suppressor variables in multiple regression research. Mult Linear Regres Viewp 24(1):11–13

Tighe EL, Barnes AE, Connor CM, Steadman SC (2013) Defining success in adult basic education settings: multiple stakeholders, multiple perspectives. Read Res Q 48(4):415–435

Van den Broek P, Bohn-Gettler C, Kendeou P, Carlson S, White MJ (2011) When a reader meets a text: the role of standards of coherence in reading comprehension. In: McCrudden MT, Magliano J, Schraw G (eds) Relevance instructions and goal-focusing in text learning. Information Age Publishing, Greenwich, pp 123–140

VanLehn K (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ Psychol 46(4):197–221

Woodcock RW, McGrew KS, Mather N (2001) Woodcock–Johnson tests of achievement. Riverside Publishing, Itasca

Woolf BP (2009) Building intelligent tutoring systems. Morgan Kaufman, Burlington

Zapata-Rivera D, Jackson T, Katz IR (2015) Authoring conversation-based assessment scenarios. In: Sottilare R, Graesser AC, Hu X, Brawner K (eds) Design recommendations for intelligent tutoring systems: authoring tools, vol 3. Army Research Laboratory, Orlando, pp 191–200

Zwaan RA, Radvansky GA (1998) Situation models in language comprehension and memory. Psychol Bull 123(2):162–185

Zwaan RA, Magliano JP, Graesser AC (1995) Dimensions of situation model construction in narrative comprehension. J Exp Psychol Learn Mem Cogn 21(2):386–397