Diagnostic Assessment of Adults' Reading Deficiencies in an Intelligent Tutoring System

Genghu Shi[1, 2], Anne M. Lippert[1, 2], Andrew J. Hampton[1, 2], Su Chen[1, 2], Ying Fang[1, 2], and Arthur C. Graesser[1, 2]
[1] University of Memphis, Memphis TN 38111, USA
[2] Institute for Intelligent Systems
gshi@memphis.edu

Abstract. In this paper, we investigate whether a version of AutoTutor that teaches comprehension strategies can be used to diagnose reading deficiencies in adults with low literacy. We hypothesized that the speed and accuracy with which participants answered questions during the AutoTutor conversation could be diagnostic of their mastery of reading comprehension components: *words*, the explicit *textbase*, the *situation model*, and *rhetorical structure*. We used linear mixed effect models to compare the accuracy and response times of 52 low literacy adults who worked on 29 AutoTutor lessons during a four-month intervention period. Our results show that adults' response accuracy for questions addressing more basic reading components (e.g., meaning of words) was higher than for those pertaining to deeper discourse levels. In contrast, question response time did not vary significantly among the theoretical levels. A correlation analysis between theoretical levels and performance (accuracy and time) supported this trend. These results affirm that adults with low literacy tend to have more proficiency for basic reading levels than for deeper discourse levels. In addition, the results of exact binomial test showed that hints or prompts were effective in scaffolding learning reading. Furthermore, we describe how response accuracy on the four comprehension components can provide a more nuanced diagnosis of reading problems than a single overall performance score. More fine-grained diagnoses can assist both educators wanting more detailed insight into learner difficulties, and ITS developers looking to improve the personalization and adaptivity of learning environments.

Keywords: CSAL AutoTutor, Reading strategies, Comprehension framework.

1 Introduction

One in six adults in the United States has low levels of literacy skills [1]. Low literacy has a negative impact on the social health and economic stability of entire countries as well as the personal well-being of its citizens [1, 2]. Adult literacy educational programs are often funded by government or non-profit organizations, but unfortunately these programs generally do not reach the level that can accommodate all adults in need. Moreover, it is difficult to teach comprehension strategies at deeper levels because few teachers and tutors in literacy centers are trained to cover these levels of reading difficulty. Intelligent tutoring systems can help close this gap and provide the necessary, deeper training. An intelligent tutoring system that can differentially diagnose reading deficits constitutes an important first step in adaptively remediating individuals' deficits. In this study, we explore the assessment capabilities of a version of a web-based intelligent tutoring system, AutoTutor [4, 7], specifically created for adults with low literacy. In particular, we use AutoTutor to classify the reading comprehension deficiencies of adults within the Graesser and McNamara [3] multilevel theoretical framework of reading comprehension.

AutoTutor for CSAL

The version of AutoTutor we developed was part of an intervention led by the Center for the Study of Adult Literacy (CSAL) [4, 7], and helps improve reading comprehension in low literacy adults. The system has two computer agents (one tutor and one peer student) that hold conversations with the human learners and with each other, called trialogues [4, 5]. Trialogues illustrate comprehension strategies to adult learners, help them apply these strategies, and give them feedback when assessing their performance, all in natural language. CSAL AutoTutor has 35 lessons that focus on distinct theoretical levels of reading comprehension [6, 7]. For each lesson, the system starts out assigning words or texts at a medium level of difficulty and AutoTutor asks 8-12 questions about the words or text, all embedded in an overarching conversation. Struggling readers tend to have even more pronounced difficulties in writing, so most of their responses are entered by clicking response options on the interface. Learner response accuracy on the medium level questions determines whether AutoTutor assigns new words or texts at a hard or easy (above or below some performance threshold) level [8]. When answers do not include all component parts of a good answer, the learner receives hints or prompts, providing another chance to pick an answer from the remaining two choices with somewhat more guidance.

CSAL AutoTutor was designed to "care" about the particular motivations, metacognitions and emotions of struggling adult readers. The caring aspect of CSAL AutoTutor is critical because most adults participating in literacy programs do so voluntarily, and if the instruction is not adult-oriented, engaging, and pertinent to adult daily life, they will stop attending. Thus, in addition to allowing easy access, individualized self-paced instruction, and intuitive design for low literacy adult learners, AutoTutor was designed to optimize engagement. First, lessons were carefully scripted to contain texts that have practical value to the adult (such as rental agreements, job applications, recipes, health information) or are expected to interest adults. Second, texts are adaptively selected by AutoTutor to be at a reading level that the student can handle (not too hard or too easy), so that the student does not become frustrated or bored. Third, trialogues were written to boost the self-esteem of the adult learner who may feel embarrassment or shame over his or her skill level. Both agents express positive encouraging messages when the adult is not performing well, and sometimes stage game-like competitions between the adult and a peer agent (with the adult always winning, thereby enhancing self-esteem). These caring functionalities of AutoTutor help create situations that users find engaging and welcoming and simultaneously allow the system to assess learner ability.

1.1 The Multilevel Framework of Comprehension

The Graesser and McNamara [3] framework identifies six theoretical levels: *words*, *syntax*, the *explicit textbase*, the *referential situation model,* the *discourse genre and rhetorical structure,* and *the pragmatic communication level* (between speaker and listener, or writer and reader). Because AutoTutor for CSAL includes only one lesson for syntax and none for pragmatic communication, we did not include these levels in our study. Of the levels we included, *word* represents the lower-level basic reading components that include morphology, word decoding, and vocabulary. The *textbase* consists of meaning of the explicit ideas in sentences and texts. The *referential situation model* (sometimes called the mental model) represents the subject matter that the texts are describing. *Genre and rhetorical structure* focuses on the type of discourse and its composition, such as narrative, persuasive, and informational genres, and also the subcategories of these genres. The last three theoretical levels (all except *word*) represent deeper discourse levels.

We hypothesize that the accuracy and time on questions in AutoTutor will be diagnostic of adult learners' mastery of comprehension components. By comparing the accuracy and time on questions of four theoretical levels [3], we can better pinpoint where adult learners' strengths and weaknesses in reading comprehension lie. Such results can provide a more nuanced diagnosis of reading problems than a single overall performance score and ultimately help improve the adaptivity of an ITS like AutoTutor. We also hypothesize that adult learners who do not answer correctly on the first attempt, and receive guidance through hints or prompts for the second attempt will perform better than chance on these questions. These results will provide insight into AutoTutor's effectiveness in helping adult learners with reading comprehension.

2 Method

2.1 Participants

The participants were 52 adults recruited from CSAL literacy classes in Metro-Atlanta (n = 20) and Metro-Toronto (n = 32). They worked on 29 lessons during a four-month intervention. Each lesson took 20 to 50 minutes to complete. Their ages ranged from 16–69 years (Mean = 40, SD = 14.97). Most of the participants were female (73.1%). All participants read at 3.0–7.9 grade levels, and 30% reported that they were either diagnosed as learning disabled or attended special education classes in their childhood.

2.2 Measures and Data Collection

Only the adults' initial responses (1 as correct, 0 as incorrect) of medium level questions in each of the 29 lessons contributed to the diagnostic analysis. This ensured a balanced design, as all participants were assigned the medium level texts, but not all participants subsequently received the easy or difficult texts. In addition, the medium level questions produce higher level discrimination. We used only the initial (as opposed to second) attempts to questions because we felt these would best reveal adults' actual mastery of the theoretical levels of comprehension. For these medium-level observations, we collected the accuracy (1 or 0) and the time to produce an answer (in seconds). Time was measured from the onset of the question to the onset of the participant's answer.

To assess the effectiveness of the hints or prompts, we collected accuracy (1 or 0) of the second attempt to all questions which were answered incorrectly on the first attempt by learners. Second attempts involved all difficulty levels (medium, easy, and hard).

We calculated accuracy and time measures for 29 lessons. Most of the lessons focus on more than one theoretical level (at most three) but have varying degrees of relevance within a lesson. For example, the lesson "Compare and Contrast" addresses mainly the *rhetorical structure* level, but also includes material involving the *textbase* and *situation model* levels. Thus, we included a relevance score for each of the four theoretical levels for each lesson. The most relevant theoretical level on a lesson received a score of 1.00, with scores of 0.67 and 0.33 assigned to the second and third order, respectively. The fourth theoretical level received a 0.00 and was thus nullified for that lesson.

2.3 Data Analysis

From each set of participant log files, we extracted time and accuracy data for the 29 lessons. We found that the distribution of response time per question was positively skewed. To alleviate the bias brought by potential outliers, we truncated the data by replacing observations falling outside three standard deviation above the mean with the corresponding value at three z-score units beyond the mean.

We first performed a descriptive analysis of the data by exploring the means and standard deviations of accuracy and time on questions of the four theoretical levels. Next we used mixed effect modeling [9], where item (question) was the unit of analysis, to test for differences in time and accuracy among the four theoretical levels. To account for the variability in participants, lessons, and questions, these components were included in the linear mixed effect models as random intercepts. We also added by-participant random slopes on different theoretical levels and random intercepts of the interaction between lesson and item for the nesting relationships. Follow-up correlational analyses were performed on the continuous measures of theoretical levels, as well as on the accuracy and time for the 29 lessons. In addition, we conducted an exact binomial test on the accuracy of second attempts to see if the proportion of correct responses is greater than chance (50%).

3 Results

Figures 1 and 2 show the means of accuracy and time on questions separately as a function of four theoretical levels. Here we see accuracy is highest and answer times are shortest for the *word* level (reference level in the analysis) compared to the three discourse levels (*textbase, situation model,* and *rhetorical structure*).

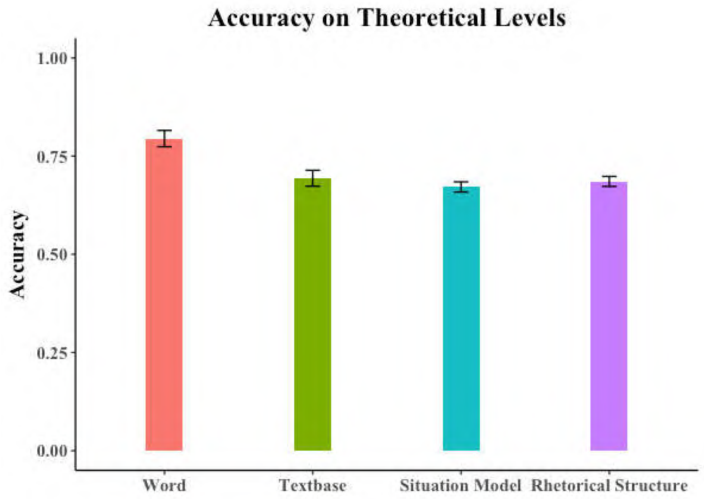

Figure 1. Adults' means accuracies (scale 0–1) on four theoretical levels, with error bars.

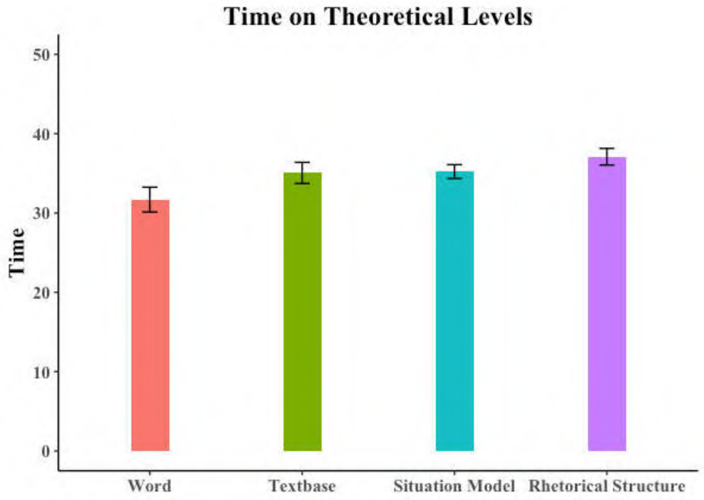

Figure 2. Adults' mean times (in seconds) to answer questions on four theoretical levels, with error bars.

Results from our logistic mixed effect model of response accuracy showed a significant difference ($\chi^2(3) = 8.34$, $p = 0.040$) in accuracy among the four theoretical levels.

Table 1 presents the output of the model. We can see that the estimated odds ratio (Estimated Odds) of *word* level is significantly higher than each of the three discourse levels (*textbase, situation model, rhetorical structure*). A post-hoc analysis with pairwise comparison showed that there was no significant difference among the three discourse levels. In contrast, results of our linear mixed effect model of suggested that time not significantly vary among theoretical levels. $F(3,25.8) = 0.058$, $p = 0.981$.

Table 1. Output of Mixed effect models on Performance and Time

		Word	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	--	0.058	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	--	0.804	0.716	0.694
	Predicted Time	34.3	36.5	37.1	37.7

Our correlational analysis showed a significant positive correlation between mean accuracies on 29 lessons and word level ($r = .386$, $p < .05$), but this correlation did not extend to any of the discourse levels. The times showed no significant correlations among theoretical levels. The pattern of correlations reinforced the results of mixed effect models of accuracy and time. In addition, the *word* level had a significant negative correlation with each of the three discourse levels (*textbase, situation model, rhetorical structure*, with *r* values of -0.365, -0.485, and -0.567, respectively).

The results of exact binomial test with 712 correct responses out of 1044 questions showed that the proportion of correct responses was significantly greater than chance (one tail *p*-value = 0.00).

4 Discussion and Conclusion

We performed mixed effect models and correlation analysis to see if there were differences among adult learners' accuracy and response times to questions in each of the four theoretical levels. As expected, the results indicated that adult learners' performance on *word* level was higher than the three discourse levels, and correlational analysis reinforced this trend. One reason for adult learners' higher performance for *word* level items is that *word* items tend to focus on individual words or single sentences. This type of stimulus is less taxing on working memory compared to items that address deeper discourse levels, which are more time-consuming, strategic, and taxing on cognitive resources.

In a previous study [6], learning gains within the four theoretical levels were tracked by considering performance on all items (medium, easy, and hard). Results revealed learning occurred for lessons involving *rhetorical structure*, but not on other theoretical levels. This implies that learning gains may be affected by the particular time frame (i.e., within lessons versus across lessons) used for assessment, the difficulty of the words and texts, and the specific theoretical levels being used. Future work is needed to further clarify these issues.

With respect to response time, we found no difference between theoretical levels, despite a trend in the data that suggested learners were slower to respond as theoretical level increased. Part of the explanation for this apparent discrepancy may be due to the modest sample size ($N = 52$), which did not provide adequate power to detect all differences. Another reason may be disengagement—the data may have been muddied by adult learners who became bored or distracted. Identifying chunks of disengagement and either removing or controlling for these periods in our analysis may reveal relevant response time variability.

The results of exact binomial test indicated that hints and prompts significantly increased a learner's probability of correctly answering a question that he or she had previously answered incorrectly. This led us to the conclusion that the trialogues in AutoTutor did help learners.

In summary, we showed how AutoTutor can be used to assess reading ability in low literacy adults and how AutoTutor trialogues scaffold learning of reading comprehension skills. By assessing comprehension within a multi-level theoretical framework, we attempted to provide a more nuanced diagnosis of adults' reading abilities than a single overall performance score. Future research could focus on designing comprehension tests for each of the theoretical levels of the multilevel comprehension framework. The results of these tests could be used to establish target population norms for each of the six components of comprehension. Knowing the range of abilities of the target adult population could help designers develop more adaptive intelligent tutoring systems for adult literacy and provide customized learning content to low literacy adults.

Acknowledgements

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

References

1. OECD (2013) OECD Skills Studies Time for the U.S. to Reskill? What the Survey of Adult Skills Says: What the Survey of Adult Skills Says. OECD Publishing
2. Vernon, J. A., Trujillo, A., Rosenbaum, S. J., & DeBuono, B. (2007). Low health literacy: Implications for national health policy.
3. Graesser AC, McNamara DS (2011) Computational analyses of multilevel discourse comprehension. Topics in Cognitive Science, 3 (2), 371-398

4. Graesser AC, Li H, Forsyth C (2014) Learning by Communicating in Natural Language with Conversational Agents. Curr Dir Psychol Sci 23:374–380
5. McNamara DS, O'Reilly TP, Best RM, Ozuru Y (2006) Improving Adolescent Students' Reading Comprehension with Istart. Journal of Educational Computing Research 34:147–171
6. Shi, G., Pavlik Jr., P., & Graesser, A.C. (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 X. Hu, T. Barnes, A. Hershkovitz, L. Paquette (Eds), Proceedings of the 10th International Conference on Educational Data Mining (pp.376-377). Wuhan, China: EDM Society.
7. Graesser, A.C., Cai, Z., Baer, W., Olney, A.M., Hu, X., Reed, M., & Greenberg, D. (2017). Reading Comprehension Lessons in AutoTutor for the Center for the Study of Adult Literacy. In S. Crossley and D. S. McNamara (Eds.), Adaptive Educational Technologies for Literacy Instruction (pp. 288─294). New York: Routledge.
8. 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.
9. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67:1–48