Improving Reading Comprehension with Automatically Generated Cloze Item Practice

Andrew M. Olney(✉), Philip I. Pavlik Jr., and Jaclyn K. Maass

University of Memphis, Memphis, TN 38152, USA
{aolney,ppavlik,jkmaass}@memphis.edu

Abstract. This study investigated the effect of cloze item practice on reading comprehension, where cloze items were either created by humans, by machine using natural language processing techniques, or randomly. Participants from Amazon Mechanical Turk ($N = 302$) took a pre-test, read a text, and took part in one of five conditions, Do-Nothing, Re-Read, Human Cloze, Machine Cloze, or Random Cloze, followed by a 24-hour retention interval and post-test. Participants used the MoFaCTS system [27], which in cloze conditions presented items adaptively based on individual success with each item. Analysis revealed that only Machine Cloze was significantly higher than the Do-Nothing condition on post-test, $d = .58$, $CI_{95}[.21, .94]$. Additionally, Machine Cloze was significantly higher than Human and Random Cloze conditions on post-test, $d = .49$, $CI_{95}[.12, .86]$ and $d = .71$, $CI_{95}[.34, 1.09]$ respectively. These results suggest that Machine Cloze items generated using natural language processing techniques are effective for enhancing reading comprehension when delivered by an adaptive practice scheduling system.

Keywords: Reading comprehension · Natural language processing · Testing effect

1 Introduction

Reading has long been one of the preeminent means of learning new information. Reading to learn necessarily involves comprehension, the process by which information in the text is reconciled with prior knowledge. Theorists differ on the precise mechanisms underlying the role of prior knowledge in reading comprehension, though there is considerable overlap across theories [19]. The differences that exist between theories may be partly attributable to differing ideas about how knowledge is represented and applied. Experimental results, however, have broadly found that prior knowledge exhibits a strong positive effect on reading comprehension [1,3,15]. Prior knowledge also moderates the effect of reading ability on comprehension. When prior knowledge is high, the effect of reading ability on comprehension vanishes [28]. Prior knowledge also influences whether reading ability interacts with text difficulty to influence comprehension [26].

E. Andr´e et al. (Eds.): AIED 2017, LNAI 10331, pp. 262–273,
2017. DOI: 10.1007/978-3-319-61425-0 22

Paper presented at 18th annual conference on Artificial Intelligence in Education,
June 28- July 02, 2017 Wuhan, China

Altogether the evidence suggests that prior knowledge has a central role, if not the central role, in reading comprehension.

If reading to learn requires prior knowledge, but the goal of reading to learn is to acquire new knowledge, then it seems there is a kind of circular causality between knowledge and reading. In educational practice, this relationship becomes apparent when the curricular focus shifts from the mechanics of reading, i.e. decoding fluency, to content area reading with the emphasis on learning from text. This shift is often marked by a sudden drop in reading scores, particularly in students from low income families [5]. Long referred to as the "fourth-grade slump," evidence now suggests that the disparity between learning to read and reading to learn starts much earlier but becomes apparent as tasks and assessments shift from narrative to informational, content-area reading [9,21]. Unfortunately, the fourth-grade slump neither begins in fourth grade, nor does it end there. Rather, the evidence suggests that early differences in reading skill widen over time. Those with high reading comprehension skill read more and become more skilled by practice, a positive-feedback loop [20]. Those with low reading comprehension skill read less, and their slowness in decoding delays identification of words by sight, which delays vocabulary growth, which in turn diminishes comprehension [30].

The importance of reading to learn has led to calls for interventions that embed comprehension activities in the learning of content areas [23]. The advantage of targeting comprehension in content areas is that, in addition to teleological prior knowledge [28], content areas typically have their own specialized vocabulary and style distinct from narrative and informal conversation, making normal mechanisms for acquiring vocabulary and grammar, like implicit learning, less efficient because of children's reduced exposure to content-area text [7,22]. Vocabulary and comprehension are deeply intertwined because text must be decoded, disambiguated, and linked with prior knowledge for comprehension to occur [12]. Multiple studies investigating the impact of unknown words on comprehension suggest that the number of unknown words should be no lower than 1 in 20 if serious comprehension deficits are to be avoided [13], which is roughly less than one unknown word per sentence.

Reading comprehension activities in educational contexts typically center around the instruction and practice of reading strategies. The definition of strategy is wide ranging and can include activities that occur before, during, or after reading of the text. Moreover, the strategies can be covert, artifact-producing, or interactive. For example, of the seven comprehension strategies recommended by the National Reading Panel (NRP) [23], comprehension monitoring and question generation are covert and occur during reading, graphic organizers and summarization are artifact producing and occur after reading, and cooperative learning, question answering, and reciprocal teaching are interactive and occur during reading. Arguably, activities that occur after reading, or tasks that are interactive, fall more into the realm of instructional activities than comprehension strategies. Nevertheless, such activities can be highly effective for increasing comprehension of text. One possible explanation for the effectiveness of these

activities is the ICAP Hypothesis [6], which predicts that learning outcomes will follow the order *interactive* > *constructive* > *active* > *passive* because of the cognitive processes required by interactive, constructive, active, and passive activities. Of the NRP comprehension activities, all but monitoring are either constructive or interactive in nature, meaning that they require generating outputs or co-generating outputs, respectively.

Although interactive educational technologies have been developed, most notably in dialogue-based intelligent tutoring systems (ITS) [24], these systems currently have two weaknesses with respect to reading comprehension. First, these systems are primarily content-oriented rather than reading-oriented, meaning that students using the ITS may not do any particular reading during the learning process (though see [14] for a counterexample). Secondly, ITS content must be authored manually, and it is commonly believed that it takes several hundred hours of authoring effort to create one hour of instruction for an ITS using traditional methods [2], though research is beginning to make progress in automated authoring [25]. Because of authoring needs and challenges, it is not currently possible to automatically create a high-quality, interactive ITS for a given piece of text on demand. Accordingly, there are two options for educational technology. First, one could focus on interactive strategy training divorced from content with the aim of strategy transfer to other texts [18]. This is a worthwhile strategy but it does not directly support comprehension of an arbitrary piece of text. Secondly, one could step back from interactive activities and instead focus on constructive activities, which is the focus of the present work.

This paper investigates an automated method for generating cloze items and the effect of practice with these items on reading comprehension. In a cloze task, a participant is asked to restore words that have been deleted from a text. Cloze tasks are well established for both vocabulary and comprehension instruction in addition to vocabulary and comprehension assessment [7,17,23]. Additionally, according to the ICAP theory, practice with cloze items is constructive because students must generate fill-in-the-blank answers, and constructive activities facilitate transfer of learning to novel contexts. In this work our primary research questions are therefore (1) whether practice with machine generated cloze items promotes reading comprehension, (2) whether reading comprehension with machine generated cloze items is equivalent to reading comprehension with human generated or random cloze items, and (3) whether reading comprehension supported by machine cloze practice supports transfer.

2 Method

2.1 Design

This study used a between-subjects design with the following conditions: Do-Nothing, Re-Read, Human Cloze, Machine Cloze, and Random Cloze conditions. All participants took pre-tests and read a text before being assigned to one of the conditions. Therefore, the Do-Nothing condition participants did nothing beyond the pre-test and reading. The Do-Nothing condition can be considered a business

as usual control condition, the Re-Read a stronger control condition where reading time is consistent with practice time in the cloze conditions, and Random Cloze another control condition where cloze practice occurs but items may not be optimal. All participants also took a post-test after a 24-hour delay. Test items with simple declarative answers, or *fact* questions, were concept-matched to test items with contextualized application questions, or *transfer* questions, such that a concept either appeared on the pre-test or on the post-test but not both. The purpose of concept matching was to eliminate the possibility that the pre-test cued participants on what to study for the post-test.

2.2 Participants

Participants were recruited through the Amazon Mechanical Turk (AMT) marketplace between September and November of 2016. In this study, participants were required to be English speakers from the U.S. or Canada and required to have completed at least 50 previous AMT tasks with at least a 95% approval rating. Experience/approval criteria were applied to prevent automated programs from attempting the experiment (i.e. "bots") and to ensure quality from human participants. Participants were paid $3 for the first phase of the experiment and $2 for the second phase following the 24-hour retention interval.

Age of participants in years was 18–25 (11%), 26–34 (45%), 35–54 (36%), 55–64 (6%), and over 65 (2%), and participants were slightly more female (52%) than male (47%). Educational attainment of participants included less than high school (<1%), high school (12%), some college (35%), bachelor's degree (43%), and graduate degree (9%). Over 95% of participants reported never having worked in a profession dealing with the circulatory system.

2.3 Materials

A text on the heart and circulatory system was derived from experimental materials used by [33], which used four versions of the text ranging from elementary school to medical school difficulty. The text used in the present study was derived from elementary school level text, with modifications primarily removing the extraneous information present in the original. Examples of removed sentences include motivational/interest statements like "You probably think you know what the heart looks like. But you may be wrong.", statements involving reader-oriented imagery like "You can feel the thumps if you press there with your hand. You can hear them with your ear.", and statements that are thematically relevant but not directly relevant to the functioning of the heart and circulatory system like "When a fire burns, carbon dioxide is formed." Both fact and transfer test items were created from the derived text by matching on a particular concept. For example, the *heart is a pump* concept has the associated fact question "Which component(s) of the circulatory system acts as a pump?" and the associated transfer question "Why doesn't oxygen rich blood flow directly from the lungs to the rest of the body?" A total of 16 concept clusters were created, each having one associated fact and transfer question for a total of 32 questions. All questions were in multiple-choice format.

Cloze items for the three cloze conditions were created either by human, randomly, or by machine using an algorithm described below. Human cloze items were created by the same researcher who derived the text and created the pre- and post-test items. The researcher selected, at their discretion, the sentences capturing the main ideas of the text and the words central to each selected sentence's meaning. The number of sentences (21) and words (53) selected by the human were then held constant in the random and machine generated cloze item conditions. Accordingly, all cloze conditions contained the same number of items, the items in each condition were generated from 21 sentences and 53 words within those sentences, but each condition differed in terms of which 21 sentences and 53 words were selected.

Random cloze items were created by randomly selecting 21 sentences from all sentences in the text and randomly selecting between one and four words in each sentence such that the words were longer than two characters, words did not include "the" or "and," and 53 words were selected in total. The random cloze generation procedure was repeated six times to create six sets of random cloze items, to minimize the chance the effects from this condition were due to an unusual random sample.

Machine generated cloze items were selected by using natural language processing techniques at the word, sentence, and discourse level. Specifically, the entire text was parsed using syntactic, semantic, and discourse parsers [10,16,29]. These parsers annotated the text with a variety of information, including part of speech, word form/lemma, named entities, syntactic dependencies, verbal and nominal predicates, argument roles, coreference chains, elementary discourse units, and discourse dependencies. Because no labeled data was available, we used applied intuition and linguistic knowledge to develop a relatively simple heuristic for the selection of sentences and words. Sentences were selected primarily based on the number of coreference chains they contained (at least three) and the length of those chains (at least two). These criteria ensured that only sentences that were well connected to the discourse were preserved. Alternatively these criteria can be considered as argument overlap where anaphora, e.g. pronouns, have been resolved to their referents (cf. [4,31]). Once selected, sentences were filtered if they consisted of only satellite discourse units, i.e. discourse units that did not carry the core meaning of the discourse relationships in which they participated. Candidate cloze words for these sentences were selected based on whether the word was an argument in a coreference chain, a semantic argument, or a syntactic subject or object with a noun or modified noun part of speech. Final cloze words were chosen from candidates if they did not belong to the 1000 most frequent words of English. For example, in the heart and circulatory system text, excluded candidate words included "heart," "middle," "blood," and "body."

2.4 Procedure

The experiment was delivered through the web interface of the MoFaCTS system [27] to AMT participants. Participants completed informed consent and

then took the pre-test. For each participant, 12 concept clusters were randomly selected from a test bank of 16 concept clusters. Four concepts were randomly assigned for pre-test, and eight concepts were randomly assigned for post-test. Since each concept had an associated fact and transfer question, the selection process yielded eight pre-test items and 16 post-test items. Order of items on each test was randomized. After the pre-test, participants read a text on the heart and circulatory system for at least 5 min and up to 10 min if they so chose. After reading the text, each participate completed one of five conditions: Do-Nothing, Re-Read, Human Cloze, Machine Cloze, or Random Cloze. Except for Do-Nothing, each of these conditions lasted from 5 min up to 25 min. Continuing longer than 5 min was purely by participant choice. The text presented in the Re-Read condition was the same as the original text. Participants in the three cloze conditions received items specific to their condition. However all items were adaptively sequenced using the MoFaCTS system based on the success history of each item and model parameters inferred from pilot experimentation. During the cloze conditions, cloze items were presented on the screen and participants were asked to fill in the missing word(s) with a 15 s timeout that was reset whenever the participant typed. After an incorrect response, the correct response was displayed for 8 s. Upon completing their condition, participants were paid for the first phase of the experiment. After a 24-hour retention interval, participants were contacted via email from MoFaCTS to complete the second phase. The second phase consisted of a post-test, consisting of items not selected on the pre-test, presented in random order. Following the post-test, participants completed a demographic survey and were paid for the second phase of the experiment.

3 Results and Discussion

Although 365 participants attempted the experiment, 13 were excluded for various reasons including using a friend's account, server crashes, and collection errors, and 50 were excluded because they did not return for the post-test, i.e. were lost to attrition ($N = 302$). Each condition had approximately the same attrition ($M = 11.6$, $SD = 1.64$), within the acceptable range for attrition and differential attrition for educational research [32]. No outliers were removed or transformed. None of the demographic variables collected (age, gender, educational attainment, professional knowledge of circulatory system) were significantly related to assigned condition under a chi-square test of independence. Table 1 shows the condition sample sizes and means, standard deviations, and 95% confidence intervals for pre- and post-test proportion correct.

Learning outcomes could not be analyzed as normalized gain scores, i.e. $(post - pre)/(1 - pre)$, because this value was undefined for some participants. The choice of analysis between ANOVA on gain scores and ANCOVA on post-test using pre-test as a covariate was informed by recent guidance suggesting that when, as in the present study, differences in pre-test between conditions are substantial, $d = .2$, and correlation between simple learning gains and pre-test

Table 1. Proportion correct

Group	n	Pre-test		Post-test	
		M (SD)	95% CI	M (SD)	95% CI
Do-Nothing	62	.46 (.23)	[.41, .52]	.54 (.20)	[.49, .59]
Re-Read	61	.46 (.19)	[.41, .51]	.57 (.23)	[.51, .63]
Random Cloze	58	.46 (.18)	[.41, .51]	.56 (.18)	[.51, .61]
Human Cloze	60	.51 (.18)	[.46, .55]	.61 (.21)	[.56, .67]
Machine Cloze	61	.50 (.20)	[.45, .55]	.67 (.22)	[.61, .73]

Note: CI = confidence interval.

is large, $r(300) = -.5$, ANOVA on gain scores is more likely to be biased than ANCOVA (see Table 5 of [11]). Therefore ANCOVA was adopted for all analyses. We conducted statistical tests at $\alpha = .05$ to address our research questions.

To answer our first research question, whether practice with machine generated cloze items promotes reading comprehension, we ran an ANCOVA with condition and pre-test proportion correct as predictors and post-test proportion correct as the dependent variable. The model controlled for differences in pre-test across participants so that differences in post-test can be attributed to condition. The ANCOVA revealed a significant main effect of condition, $F(4, 296) = 3.04$, $p = .02$, $\eta_p^2 = .04$, as well as a main effect of pre-test proportion correct, $F(1, 296) = 53.95$, $p < .001$, $\eta_p^2 = .15$. Post hoc comparisons between predicted marginal means using Tukey's HSD revealed that the Machine Cloze had significantly higher post-test proportion correct ($M = .66$, $SE = .03$) than the Do-Nothing condition ($M = .55$, $SE = .03$), $t(296) = 3.21$, $p = .01$, $d = .58$, $CI_{95}[.21, .94]$. No other pairwise comparisons were significant.

An additional exploratory analysis was performed to investigate whether other variables or interactions omitted from the ANCOVA might qualify or limit these results. An exploratory ANCOVA model with condition, text reading time (log transformed), pre-test proportion correct, and all interactions as predictors and post-test proportion correct as the dependent variable was created and refined using backward elimination variable selection based on the Akaike information criterion (AIC). The only significant predictors in the exploratory model were condition and pre-test proportion correct, which were the same predictors in the a priori model. Diagnostic plots revealed no concerning departures from normality, heterogeneity, or violations of independence, suggesting the model was well-fitted.

To answer our second research question, whether reading comprehension with machine generated cloze items is equivalent to reading comprehension with human generated or random cloze items, we ran an ANCOVA with the three cloze conditions, pre-test proportion correct, and variables controlling for the learning experience within the cloze conditions as predictors and post-test proportion correct as the dependent variable. The measured variables controlling for the learning experience within the cloze conditions included proportion correct

across trials, number of trials, and time. Because time and number of trials were highly correlated, $r(176) = .94$, and number of trials (log transformed) was more normally distributed than time, trials was included in the model and time was not included. Furthermore, because the learning experience necessarily involves correctness over time, an interaction between number of trials and proportion correct across trials was included. Thus the model controlled for differences in pre-test scores, number of trials, proportion correct across trials, and the interaction of number of trials and proportion correct across trials so that differences in post-test can be attributed to condition.

The ANCOVA revealed a significant main effect of condition, $F(2, 171) = 7.89$, $p < .001$, $\eta_p^2 = .08$, a main effect of pre-test proportion correct, $F(1, 171) = 5.78$, $p = .02$, $\eta_p^2 = .03$, and a main effect of number of trials, $F(1, 171) = 9.80$, $p = .002$, $\eta_p^2 = .05$. A main effect of proportion correct across trials was not significant $F(1, 171) = 1.57$, $p = .21$, but the interaction of proportion correct across trials and the number of trials was significant, $F(1, 171) = 10.27$, $p = .002$, $\eta_p^2 = .06$. Examination of the interaction slope revealed that participants with low proportion correct across a high number of trials fared poorly on post-test proportion correct. Note that while only the main effect of condition was relevant to our hypothesis, the effects of condition, number of trials, and the interaction of the number of trials and proportion correct across trials are statistically significant with Bonferroni adjusted alpha levels of .01 per test ($\alpha = .05/5$). Post hoc comparisons between predicted marginal means using Tukey's HSD revealed that the Machine Cloze had significantly higher post-test proportion correct ($M = .66$, $SE = .02$) than the Human Cloze condition ($M = .58$, $SE = .02$), $t(171) = 2.69$, $p = .02$, $d = .49$, $CI_{95}[.12, .86]$ and significantly higher post-test proportion correct than the Random Cloze condition ($M = .54$, $SE = .02$), $t(171) = 3.88$, $p < .001$, $d = .71$, $CI_{95}[.34, 1.09]$.

An additional exploratory analysis was performed to investigate whether other variables or interactions omitted from the ANCOVA might qualify or limit these results. An exploratory ANCOVA model with condition, text reading time (log transformed), pre-test proportion correct, number of trials, proportion correct across trials, and all two-way interactions as predictors and post-test proportion correct as the dependent variable was created and refined using backward elimination variable selection based on the Akaike information criterion (AIC). The significant predictors in the exploratory model were identical to the a priori model except for the addition of a pre-test proportion correct by number of trials interaction, $F(1, 170) = 5.50$, $p = .02$, $\eta_p^2 = .03$. Examination of the interaction slope revealed that participants with low pre-test proportion correct who experienced a high number of trials fared better on post-test proportion correct while participants with high pre-test proportion correct who experience a high number of trials fared more poorly. Though this interaction is sensible, it should be treated with caution because it was obtained through variable selection [8]. The most useful finding of the exploratory ANCOVA is that it did not alter the significant effect of condition or contrasts found in the a priori ANCOVA. Diagnostic plots revealed no concerning departures from normality, heterogeneity, or violations of independence, suggesting the model was well-fitted.

To answer our final research question, whether reading comprehension with machine generated cloze items supports transfer, we re-ran ANCOVAs with test scores based on the transfer questions alone. An ANCOVA for transfer post-test proportion correct using condition and transfer pre-test proportion correct as predictors yielded virtually the same effects and contrasts as the ANCOVA for all test items. There was a significant main effect of condition, $F(4, 296) = 2.59$, $p = .04$, $\eta_p^2 = .03$, as well as a main effect of pre-test proportion correct, $F(1, 296) = 23.34$, $p < .001$, $\eta_p^2 = .07$. Post hoc comparisons between predicted marginal means using Tukey's HSD revealed that Machine Cloze had significantly higher transfer post-test proportion correct ($M = .61$, $SE = .03$) than the Do-Nothing condition ($M = .50$, $SE = .03$), $t(296) = 2.82$, $p = .04$, $d = .51$, $CI_{95}[.15, .87]$. No other pairwise comparisons were statistically significant. An ANCOVA for transfer post-test proportion correct using the three cloze conditions, pre-test proportion correct, number of trials, proportion correct across trials, and the interaction of number of trials and proportion correct as predictors also yielded virtually the same effects and contrasts as the ANCOVA for all test items. There was a significant main effect of condition, $F(2, 171) = 6.52$, $p = .002$, $\eta_p^2 = .07$, a main effect of pre-test proportion correct, $F(1, 171) = 3.98$, $p = .05$, $\eta_p^2 = .02$, and a main effect of number of trials, $F(1, 171) = 9.13$, $p = .003$, $\eta_p^2 = .05$. A main effect of proportion correct across trials was not significant $F(1, 171) = 0.56$, $p = .46$, but the interaction of proportion correct across trials and the number of trials was significant, $F(1, 171) = 7.45$, $p = .007$, $\eta_p^2 = .04$. Examination of the interaction slope revealed that participants with low proportion correct across a high number of trials fared poorly on post-test proportion correct. Post hoc comparisons between predicted marginal means using Tukey's HSD revealed that the Machine Cloze had significantly higher transfer post-test proportion correct ($M = .61$, $SE = .02$) than the Human Cloze condition ($M = .52$, $SE = .03$), $t(171) = 2.71$, $p = .02$, $d = .5$, $CI_{95}[.13, .86]$ and significantly higher transfer post-test proportion correct than the Random Cloze condition ($M = .49$, $SE = .03$), $t(171) = 3.42$, $p = .002$, $d = .63$, $CI_{95}[.26, 1.0]$.

Our main findings were that the Machine Cloze condition led to superior post-test outcomes relative to other conditions, including Human Cloze when learning experience variables are controlled for, and that these findings hold both overall and for a subset of pre- and post-test questions specifically targeting transfer. The causal mechanism behind the advantage for the Machine Cloze condition is currently unclear. An examination of the Human Cloze and Machine Cloze items revealed 13 sentences in common out of 21. Presumably differences in learning between the Human and Machine Cloze conditions are attributable to the items not shared and their interactions with the items in common. Recall that the primary features for selecting the Machine Cloze sentences were based on coreference chains. Sentences with more chains and with longer chains are more connected to the discourse by virtue of echoing or extending ideas present in other sentences. For the eight items not shared, the sum of Machine Cloze coreference lengths was 221 and the sum of Human Cloze coreference weights was

67, meaning that the Machine Cloze items were approximately three times more connected to the discourse than the Human Cloze items. Whether differences in coreference chains can explain differences in post-test performance is a matter for future research.

4 Conclusion

Results from the study suggest that cloze items generated by machine using natural language processing techniques are effective for enhancing reading comprehension when delivered by an adaptive practice scheduling system. Because such cloze items can be generated automatically, ostensibly for any text, our findings potentially have broad implications for improving reading comprehension in educational settings. An important limitation on these implications, however, is that these results were obtained for a single text only and in comparison to human-generated items by a single individual. It may be that the natural language processing techniques used were particularly suitable to this text and would not be as effective for other texts or that these techniques would not fare as well against items generated by a domain expert. Two important targets for future research are to replicate this finding with other texts in other domains and to better understand the properties of the machine generated cloze items that made them more effective than human generated cloze items.

Acknowledgements. This work was supported by the National Science Foundation Data Infrastructure Building Blocks program (NSF; ACI-1443068), by the Institute of Education Sciences (IES; R305C120001) and by the Office of Naval Research (ONR; N00014-00-1-0600, N00014-12-C-0643; N00014-16-C-3027). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, or ONR.

References

1. Ahmed, Y., Francis, D.J., York, M., Fletcher, J.M., Barnes, M., Kulesz, P.: Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. Contemp. Educ. Psychol. **4445**, 68–82 (2016)
2. Aleven, V., Mclaren, B.M., Sewall, J., Koedinger, K.R.: A new paradigm for intelligent tutoring systems: example-tracing tutors. Int. J. Artif. Intell. Educ. **19**(2), 105–154 (2009)
3. Bransford, J.D., Johnson, M.K.: Contextual prerequisites for understanding: some investigations of comprehension and recall. J. Verbal Learn. Verbal Behav. **11**(6), 717–726 (1972)
4. Britton, B.K., Gülgöz, S.: Using Kintsch's computational model to improve instructional text: effects of repairing inference calls on recall and cognitive structures. J. Educ. Psychol. **83**(3), 329–345 (1991)
5. Chall, J.S., Jacobs, V.A.: Writing and reading in the elementary grades: developmental trends among low SES children. Lang. Arts **60**(5), 617–626 (1983). http://www.jstor.org/stable/41961511

6. Chi, M.T.H., Wylie, R.: The ICAP framework: linking cognitive engagement to active learning outcomes. Educ. Psychol. **49**(4), 219–243 (2014). http://dx.doi.org/10.1080/00461520.2014.965823
7. Fang, Z.: The language demands of science reading in middle school. Int. J. Sci. Educ. **28**(5), 491–520 (2006). http://dx.doi.org/10.1080/09500690500339092
8. Harrell, F.: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Graduate Texts in Mathematics. Springer, New York (2001)
9. Hirsch, E.D.: Reading comprehension requires knowledge-of words and the world. Am. Educator **27**(1), 10–13, 16–22, 28–29, 48 (2003)
10. Johansson, R., Nugues, P.: Dependency-based syntactic-semantic analysis with PropBank and NomBank. In: Proceedings of the Twelfth Conference on Computational Natural Language Learning, CoNLL 2008, pp. 183–187. Association for Computational Linguistics, Morristown (2008)
11. Kelly, S., Ye, F.: Accounting for the relationship between initial status and growth in regression models. J. Exp. Educ. **85**(3), 353–375 (2017). http://dx.doi.org/10.1080/00220973.2016.1160357
12. Kintsch, W.: Comprehension: A Paradigm for Cognition. Cambridge University Press, Cambridge (1998)
13. Laufer, B.: Lexical thresholds for reading comprehension: what they are and how they can be used for teaching purposes. TESOL Q. **47**(4), 867–872 (2013). http://www.jstor.org/stable/43267941
14. Leelawong, K., Biswas, G.: Designing learning by teaching agents: the Betty's Brain system. Int. J. Artif. Intell. Educ. **18**(3), 181–208 (2008)
15. Lipson, M.Y.: Learning new information from text: the role of prior knowledge and reading ability. J. Read. Behav. **14**(3), 243–261 (1982)
16. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60. Association for Computational Linguistics, Baltimore, June 2014. http://www.aclweb.org/anthology/pp.14-5010
17. McKeown, M.G., Beck, I.L., Omanson, R.C., Pople, M.T.: Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. Read. Res. Q. **20**(5), 522–535 (1985). http://www.jstor.org/stable/747940
18. McNamara, D., Levinstein, I., Boonthum, C.: iSTART: interactive strategy training for active reading and thinking. Behav. Res. Methods **36**, 222–233 (2004). doi:10.3758/BF03195567
19. McNamara, D.S., Magliano, J.: Toward a comprehensive model of comprehension. In: The Psychology of Learning and Motivation, Psychology of Learning and Motivation, vol. 51, pp. 297–384. Academic Press (2009). https://www.sciencedirect.com/science/article/pii/S0079742109510092
20. Mol, S.E., Bus, A.G.: To read or not to read: a meta-analysis of print exposure from infancy to early adulthood. Psychol. Bull. **137**(2), 267–296 (2011)
21. Moss, B.: Making a case and a place for effective content area literacy instruction in the elementary grades. Read. Teach. **59**(1), 46–55 (2005). http://dx.doi.org/10.1598/RT.59.1.5
22. Nagy, W., Townsend, D.: Words as tools: learning academic vocabulary as language acquisition. Read. Res. Q. **47**(1), 91–108 (2012). http://dx.doi.org/10.1002/RRQ.011

23. National Institute of Child Health and Human Development: Report of the National Reading Panel. Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. NIH Publication No. 00-4769, U.S. Government Printing Office, Washington, DC (2000)
24. Nye, B.D., Graesser, A.C., Hu, X.: AutoTutor and family: a review of 17 years of natural language tutoring. Int. J. Artif. Intell. Educ. **24**(4), 427–469 (2014). http://dx.doi.org/10.1007/s40593-014-0029-5
25. Olney, A.M., Brawner, K., Pavlik, P., Koedinger, K.: Emerging trends in automated authoring. In: Sottilare, R., Graesser, A., Hu, X., Brawner, K. (eds.) Design Recommendations for Intelligent Tutoring Systems, Adaptive Tutoring, vol. 3, pp. 227–242. U.S. Army Research Laboratory, Orlando (2015)
26. Ozuru, Y., Dempsey, K., McNamara, D.S.: Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. Learn. Instr. **19**(3), 228–242 (2009)
27. Pavlik, P.I., Kelly, C., Maass, J.K.: The mobile fact and concept training system (MoFaCTS). In: Micarelli, A., Stamper, J., Panourgia, K. (eds.) ITS 2016. LNCS, vol. 9684, pp. 247–253. Springer, Cham (2016). doi:10.1007/978-3-319-39583-8_25
28. Recht, D.R., Leslie, L.: Effect of prior knowledge on good and poor readers' memory of text. J. Educ. Psychol. **80**(1), 16–20 (1988)
29. Surdeanu, M., Hicks, T., Valenzuela-Escarcega, M.A.: Two practical rhetorical structure theory parsers. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 1–5. Association for Computational Linguistics, Denver. http://www.aclweb.org/anthology/N15-3001
30. Torgesen, J.K.: Avoiding the devastating downward spiral: the evidence that early intervention prevents reading failure. Am. Educator **28**(3), 6–19 (2004)
31. Vidal-Abarca, E., Martínez, G., Gilabert, R.: Two procedures to improve instructional text: effects on memory and learning. J. Educ. Psychol. **92**(1), 107–116 (2000)
32. What Works Clearinghouse: Assessing attrition bias. Technical report, Institute of Eduation Sciences (2012). https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_attrition_v2.1.pdf
33. Wolfe, M.B., Schreiner, M., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., Landauer, T.K.: Learning from text: matching readers and texts by latent semantic analysis. Discourse Process. **25**(2–3), 309–336 (1998). http://dx.doi.org/10.1080/01638539809545030