5-1-2019

Improving Reading Comprehension Of Science Texts With Computer Generated Cloze Item Practice

Davis Whaley

IMPROVING READING COMPREHENSION OF SCIENCE TEXTS WITH COMPUTER GENERATED CLOZE ITEM PRACTICE

by

Davis Whaley

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Psychology

The University of Memphis

May 2019

Abstract

This research studied whether computer-generated cloze items using natural language processing methods could promote learning and comprehension of science texts compared to human and random cloze items. Participants recruited from Amazon Mechanical Turk (*N* = 562) took a pretest on one of three science topics and then read a text on it. Participants then practiced cloze items about the text generated either by a computer (machine), human, or randomly. Cloze items were presented using the MoFaCTS adaptive practice system. After 24 hours participants took a post-test on the text. ANOVA showed a significant effect of cloze type on gain score, and pairwise comparisons found the human conditions had higher gain scores than machine or random conditions. A separate ANOVA on the circulatory system text showed machine had higher gain scores than random. Implications of these findings are discussed.

Table of Contents

With calls by the U.S. Department of Education to raise the educational level of American students in STEM fields, there is a great need to incorporate theories of reading, comprehension, and learning into current educational initiatives. What theories generally agree on is that reading is a constructive and active process with interactions between the reader and the text. For example, the DIME model of reading emphasizes summarizing, vocabulary, inference generation, and self-questioning in understanding a text (Cromley & Azevedo, 2007). These overlap with recommendations from the National Reading Panel (2000), which found that strategies such as cooperative learning, question answering, question generation, and summarization have a scientific basis for improving reading comprehension. Another theoretical framework supporting reader engagement is the ICAP (Interactive, Constructive, Active, Passive) hypothesis (Chi & Wylie, 2014). ICAP predicts learning outcomes are based on learner behavior. Specifically, readers who interact with text will learn more than passive readers. These ideas are not comprehension specific, but they do all require the reader to actively engage with the text. Cloze item practice may utilize all of these strategies because readers must answer questions requiring inferences which can test vocabulary. In encouraging active reading, cloze tasks may promote beneficial reading habits.

Whether readers can get interested in a text also depends on its subject matter. Science texts have their own unique language and vocabulary and may present extra challenges to teachers (Fang, 2006). Ozuru, Dempsey, & McNamara (2009) found that college students' comprehension of biology texts was positively correlated with their prior knowledge of the subject. Prior knowledge strongly influences how much new information students learn from a text, so much so that inaccurate prior knowledge may interfere with learning (Lipson, 1982). Similarly, reader memory for science texts is affected by incorrect prior knowledge (Kendeou, 2007). Clearly, what readers bring with them influences how much they can learn. This might be especially true for the science domains, which university students hold many misconceptions about (Stein, 2008). These misconceptions may come from parents, news, magazines, websites, or even from school, as teachers in elementary school are more likely to have misconceptions about science themselves.

Vocabulary is influential in reading comprehension too depending on context. Unlike vocabulary used in the news, emails, or among peers, academic vocabulary must be used within its own context for students to fully comprehend the words. This is a bigger problem for science domains with more domain-specific vocabulary (Nagy, 2012), as there may be less opportunity to practice it. Thus, according to models of reading comprehension, a reader has to comprehend a science text that may not engage them while overcoming difficult vocabulary and lack of prior knowledge. Cloze item practice may be uniquely suited to solve this problem.

Cloze items were first used in an educational setting by Wilson Taylor in 1953 to test the readability of textbooks used in schools in the United States (Taylor, 1953). Most early research in the 1960s and 70s used cloze items to measure English ability among ESL students or as a test of reading proficiency among English speakers (Alderson, 1979; Bormuth, 1965). Cloze items are essentially fill-in-the-blank questions derived from a text. The answers are usually open-ended and without a word bank. The reader produces an answer to the question using context and inferences. Vocabulary is also tested because the reader must generate words themselves, in contrast to multiple-choice or word banks which are more passive (Lipson, 1982). Theoretically then, cloze tasks can address problems readers face, while fitting well within models emphasizing interaction and engagement.

There is disagreement over how cloze operates: at the word and sentence level (Carlisle & Rice, 2004; Porter, 1983) or is instead intersentential (Jonz, 1990). Cloze items restricted to the word or sentence level are less likely to test for comprehension across a passage, although this is more likely when the deleted words are chosen without reason. For example, Gellert & Elbro (2013) constructed cloze tests by choosing words requiring inferences to be made over many sentences to test comprehension beyond a "local" level. They found that scores on a cloze test requiring inferences were highly correlated with a "gold standard" test of reading comprehension. More recently, Brown (2013) has suggested that whether a cloze item is sentential or intersentential depends on the reading level of the student.

How the word is chosen is given much attention in cloze literature and concerns the present study. The two main kinds of deletion are nth deletion, also called fixed-ratio, and rational deletion. The nth deletion has every word deleted after a certain number chosen by the creator. Nth deletion is not per-sentence but over the whole passage and may start at or near the beginning. There are variations of nth deletion that only count specific word classes, e.g., keeping proper nouns and numbers (Kobayashi, 2002). While the words targeted by nth deletion are mostly random, rational deletion does not delete words after a certain count and instead follows the creator's own method. In practice, this means rational deletion can target more specific word classes like nouns. A meta-analysis of cloze studies found 15 different styles of deletion, with 12th-word deletion being the most common (55 cases), followed by 7th-word deletion (25 cases), followed by rational deletion (20 cases; Watanabe, 2008). The same meta-analysis found 7th-word and rational deletion to be the most reliable of cloze tests measuring ESL proficiency.

While there are many ways to delete a word and create a cloze item, there is disagreement on which pattern of deletion produces cloze items that best measure comprehension of a text, especially longer texts that aren't meant for ESL. However, there is more agreement that different deletion patterns produce qualitatively different kinds of items. Abraham & Chapelle (1992) looked at the types of cloze answers in nth-word deletion and rational deletion, finding that nth-word required more retrieval from long-term memory and rational deletion words were more related to contextual clues. They also found that content words were more difficult than function words. This suggests that nth-word deletion might measure memory more than comprehension. Bridge (1982) studied deleting different kinds of words in cohesive relationships. She found that students were aware of these relationships and used them to help answer the questions. She concluded that cloze sentences using nth deletion patterns would be less likely to require intersentential processing than sentences with rationally deleted words whose answers require information over multiple sentences. Bachman (1985) conducted an experiment where one group of university students was given nth-word science passages compared to rational deletion, identifying four types of deletions such as within-clause, across clause-within sentence, across sentence, and extra-textual. Results showed nth-words were more difficult, though this was attributable to the greater frequency of the more difficult extra-textual words produced by nth-word. He suggested that a creator's ability to choose and tailor the deleted words might lead to a better test. In support of this, Greene (2001) advocated that the blank should require an inference that links other sentences. Brown (2013) has even stated: "a cloze test that is not tailored is just an inefficient collection of unpiloted items" (p. 27).

Most cloze research has evaluated reader comprehension, particularly in English as a Second Language (ESL) classes. Comparatively fewer studies focus on cloze tasks as interventions to improve comprehension, sometimes termed "instructional cloze," although there is some research. When used as instruction, Jongsma (1980) in his review of the cloze literature at the time, found that rational deletion was more effective than nth-word as a teaching technique. In a study by Sampson (1982), third-graders practiced cloze lessons on a story for 15 weeks. The cloze group read the story, completed the deletions, and then discussed their answers with a teacher. The control group had regular reading instruction. At the end of the study, the experimental group had better scores on a comprehension post-test of the same story than the control, although the cloze condition was confounded by discussing the cloze item answers with a teacher. Other researchers have tried to use cloze instruction in remedial reading classes with mixed results (Pessah, 1975). Cloze instruction has also been used to improve recall of social studies material in a sixth-grade class (Grant, 1976). Jongsma (1980) concluded that cloze instruction was most helpful in improving comprehension and in social studies content material.

Teachers interested in using and making cloze for instruction only have general tips to follow that were mostly meant for testing and evaluation of ESL classes. Deciding on a deletion pattern and making cloze items as a lesson would also take time. This puts a burden on teachers to produce such content. If the process could be automated and done with a computer, this would make it easier to incorporate cloze procedures in the classroom.

Internet-based software that generates cloze items from text is increasingly common, but it’s unclear how many of the computerized cloze products either freely available or on the educational market have been tested in controlled conditions. TELE-Web was a program used in the 90s that incorporated cloze exercises to teach sight-reading and vocabulary. Englert, Zhao, Collings, & Romig (2005) used TELE-Web to improve the word recognition and sight-reading skills of 1st graders, but the study was small and lacked a control group. Seamon & Levitt (2003) reviewed three popular cloze software products and found that the free Hot Potatoes was easy to use and that students liked the web environment. Cloze Generator is a newer program that automatically creates cloze words from text and takes into account the vocabulary and Lexile difficulty levels. It also allows some teacher input into selecting the words (Kitao & Kamiya, 2009). One advantage computer-based cloze tasks have is that the student is often shown the correct answer after a mistake. This could correct mistaken prior knowledge quickly and easily. While they are advances over what came before, these programs mainly use nth deletion that is sometimes mixed with limited rational deletion.

While a program that can make its own cloze items would be useful, how people practice these items is also important. MoFaCTS is an adaptive practice scheduling system (Pavlik, Kelly, & Maass, 2017) based on "chunk" theory (Johnson, 1970). It was developed from the FaCT system, which allowed users to practice domain material in a sequence of drills while immediately correcting wrong answers (Pavlik et al., 2007). Run in a web browser, MoFaCTS lets the teacher create content. Different content can be set up and presented in multiple ways which allows researchers to test the effects of different manipulations. The MoFaCTS algorithm "uses a computational model of memory to infer the best item to practice next" (Pavlik et al., 2017, p. 3). In presenting practice items, it accounts for whether a participant got an item wrong in the past, as well as model parameters from pilot experimentation (Pavlik et al., 2017). The system is adaptable and can be configured to the needs of a project, making it suitable for large, complex experiments. It has been used to practice Mandarin Chinese, music intervals, statistics, and cloze items on the circulatory system. Post-tests assessing learning on the circulatory system using MoFaCTS have found transfer effects after practice (Olney, Pavlik, & Maass, 2017).

In Olney et al. (2017), the researchers used the MoFaCTS system to deliver cloze items to participants in a sequential order to optimize learning and to test cloze items made by different methods. Unlike most other cloze studies where the cloze task itself is a measure of comprehension, Olney et al. (2017) used MoFaCTS as an instructional cloze technique to improve the reading comprehension of a science text. In that experiment, participants (N = 302) took a pretest to assess their prior knowledge and comprehension of the circulatory system. After reading the text, they were assigned to one of five conditions: Machine cloze, Human cloze, Random cloze, Do-nothing, and Re-read. For the cloze conditions, the condition determined what type of cloze items they would answer about the text. These cloze items were made by either human researchers, a machine, or at random. Participants answered the cloze items for five minutes, which were presented to them by MoFaCTS. Participants in the Re-read or Do-nothing conditions either reread the text or finished after the first reading. After at least a 24-hour delay, the participants returned to take a post-test that measured their reading comprehension. Half of the questions in both the pretest and post-test were transfer questions which required making inferences and whose answers could not be found word-for-word in the text. The other half were fact questions which were simpler, declarative questions. A comparison of the five conditions using ANCOVA with pre-test as a covariate found that only the machine condition was significantly better than the do-nothing control. A comparison of the three cloze conditions using ANCOVA with pre-test, proportion of correct cloze trials, and the number of trials found the machine had significantly higher post-test scores than human and random conditions. The second research question, whether machine cloze items support transfer, was tested with an ANCOVA that looked only at transfer questions. The significant effects described above were found for transfer questions alone and for fact and transfer combined.

An explanation for why the machine did better than the human might be revealed by the sentences and cloze items shared and not shared between conditions. Thirteen out of the 21 sentences used to make cloze items were shared between machine and human in the circulatory system text. For the sentences not shared in that study, the machine and human conditions were compared in terms of coreference chains, which measure how much a sentence is connected to other sentences in the text. With this measure, the sum of the machine cloze coreference chain lengths was 221, and the sum of the human cloze coreference chain lengths was 67, suggesting that the success of the machine may be due to stronger connections between the machine cloze items and the rest of the text.

The present study attempted to replicate and extend the previous study by using instructional cloze items made by a teacher (human), computer (machine), or randomly to promote reading comprehension across two new scientific texts in addition to the circulatory system text. The research questions were: (1) would previous findings that machine-generated cloze practice promotes more comprehension compared to the random and human conditions on the circulatory system text be replicated, (2) would the machine-generated cloze practice promote more comprehension than random or human conditions for two new texts, (3) would machine cloze conditions promote learning transfer in the two new texts.

Methods

Design

The study was a 3x3 between-subjects factorial design. The two independent variables manipulated were cloze and text. Cloze had three levels: machine, human, random. Text had three levels: circulatory system, nitrogen cycle, photosynthesis. There were nine conditions: human-circulatory, machine-circulatory, random-circulatory, human-nitrogen, machine-nitrogen, random-nitrogen, human-photosynthesis, machine-photosynthesis, random-photosynthesis. Variables measured were pre- post-test score, practice timeout, practice score, reading time, dates of study, and demographic data.

Participants

A power analysis done with GPower3.1 using ANCOVA to compute the required sample size showed that to get 95% power with an effect size f of 0.25 required a total sample size of 251 participants. Post-hoc t-tests showed to get 80% power required 64 participants per condition. The effect size was taken from the previous Olney et al. (2017) study. Participants were recruited using Amazon Mechanical Turk (AMT). All participants were required to be from the United States or Canada and to have completed at least 100 AMT tasks with a 95% approval rating or higher. Participants were paid $3 for the first part of the experiment and $2 for returning to do the second part after a 24-hour retention interval. Any subject who had completed at least part one of the study was excluded from taking the study again on a later date.

In total, 809 participants were recruited and paid for part one using AMT between June and October of 2018. We excluded all participants who took longer than 72 hours to return between part one and part two (excessive retention interval), or who timed out when answering cloze items greater than 20% of the time, or whose data had errors such as missing pretests or double post-tests, or who had taken part in the original Olney et al. (2017) study, or anyone who failed at least one attention check question embedded in the pre- and post-tests. Attrition was balanced across conditions.

The retention interval cutoff had to be chosen as participants would eventually forget any learned material. We kept to a 72-hour interval because that was the design used in the previous study. For the practice timeout percentage, our study was set up so that a participant could timeout each question and still complete the cloze practice section and advance on. Filtering participants who timed out was necessary in order to exclude bots or people who showed satisficing behavior that met only the minimum threshold necessary to complete the HIT and get paid. A 20% or greater timeout exclusion cutoff was chosen to identify more of these satisficing participants and because at 20%, participants were more evenly distributed across conditions. This pattern of recruiting and excluding participants not meeting the above criteria was repeated until there were near 65 participants per condition. However, due to an error in counting participants who failed an attention check question we had to exclude 23 additional participants after the data collection had finished and ended with 562 participants for analysis. See Table A1 below.

Table 1.

Number of Participants

Text	Human	Machine	Random
Circulatory System	64	63	62
Nitrogen Cycle	61	63	61
Photosynthesis	61	64	63

Demographic data collected showed nearly half (48.22%) had college degrees or higher, 38.79% had some college, and 12.63% had a high school degree or less. Participants were 52.14% male and 47.33% female. On age, 40.39% of participants were 35-54 years old, 35.23% were 26-34, 16.01% were 18-25, and 8.36% were age 55 or older. The vast majority of participants (93.23%) had not worked in a profession that dealt with knowledge of their text.

Materials

The three texts used were on the circulatory system, the nitrogen cycle, and photosynthesis. The original text on the circulatory system was about how blood moves through the human body powered by the heart. The new text on photosynthesis described in great detail how a plant cell converts light into energy. The new text on the nitrogen cycle followed nitrogen moving from sky to ground to sea. The text on the circulatory system was taken from Olney et al. (2017). In that study, their text was based on an elementary school level text from Wolfe (1998) with some revisions that were detailed. The texts on the nitrogen cycle and photosynthesis were based on detailed passages and descriptions from several educational websites. Extraneous information was excluded to highlight the main points of each topic. Some sentences were reworded to keep the texts about the same word length and same Flesh-Kincaid level as the circulatory system text. The topics of the nitrogen cycle and photosynthesis were chosen because, like the circulatory system, each described a complex, multi-step scientific process. In order to understand the text, the reader was required to use inferences and memory, two abilities that cloze tasks tap into well. More details of these texts are in Table 2.

Table 2.

Measures across Conditions

Measure	Nitrogen Cycle	Photosynthesis	Circulatory System
Number of Words	997	1004	1001
Flesh-Kincaid	8.3	8.4	6.5
Lexile	800-900	900-1000	900-1000
Number of Cloze Items	66	74	53

Thirty-four questions for each text were used as a test bank for pre- and post-test questions. The questions on the circulatory system were reused from the prior study. Questions on the new texts about photosynthesis and the nitrogen cycle were made by a researcher in the present study. The researcher was blind to the cloze items while making the test questions to minimize bias. Fact and transfer questions were created by focusing on a concept within a text. For example, “Nitrogen uptake is when the roots bring fixed nitrogen, like ammonium, into the plant body. Plants can absorb fixed nitrogen from the soil or bacteria.”, was used to make the fact question “What is nitrogen uptake?” and the transfer question “If plants could fix and use nitrogen directly from the air, what might they not need?”. 16 of these concepts were identified within the text, and from each one, a fact and transfer question were made, totaling 32 questions. The remaining two questions were attention check questions to detect “bots” or random clicking and were not used to calculate gain scores. These questions were reviewed by two experts external to the study who checked to make sure the questions matched with a concept in the text.

Four multiple-choice items were created as possible answers for each question and once created, piloting of the new questions used item-analysis techniques. Participants read the questions without the texts before answering. Distractor choices were made from free response versions of the questions as had been done for the circulatory system and recommended by Gierl, Bulut, Guo, & Zhang (2017). Multiple rounds of piloting were done using Qualtrics with participants recruited through AMT. This continued until each correct answer was selected by less than 50% of the participants and with the goal to have each item selected an equal number of times using distractors. The std. Cronbach's alpha calculated from piloting final versions of the questions from the two new texts of photosynthesis and nitrogen cycle were 0.7 and 0.62, respectively. When interpreting the standardized Cronbach's alpha for our texts, it should be understood that the texts contain wide and varied information within themselves. For example, the nitrogen cycle has questions about the atmosphere, soil, plants, oceans, sewage, and bacteria, and though these are all connected through nitrogen, the text could draw upon quite different domains of knowledge. Questions made from them may not be expected to have high reliability.

Cloze items were created either by a human, machine, or randomly. The human cloze items for the circulatory system text were reused from the prior study. A researcher of that study created them by selecting sentences with main ideas and from those selecting words most central to each sentence's meaning. The number of sentences (21) and words (53) that the researcher chose was held constant to make the random and machine cloze items for that text. Thus, all cloze conditions within the circulatory system had the same number of sentences (21) and words (53) but differed in content. However, there was some flexibility in the process. For example, if the researcher selected two sentences with one and two cloze items from each sentence, then the machine or random algorithms could select two sentences and reverse the order, with two and one cloze items for each sentence, as long as the total number of sentences and items selected was the same. Cloze items for the photosynthesis and nitrogen cycle texts were made by a high school teacher who taught biology and was unaware of the design and goal of the experiment. The teacher was emailed the texts and given examples of cloze items. They were told to make 45-55 items for each text but with the freedom to deviate from this range to make items that would best help students remember and understand the material. With this in mind, the teacher made cloze items for the nitrogen cycle and photosynthesis texts that came from 26 sentences and 66 words, and 24 sentences and 74 words, respectively. Like for the circulatory system, these numbers were held constant for each text while making their respective machine and random cloze items. Within each text, all cloze conditions had the same number of practice items.

The researchers created random cloze items by randomly selecting at least one word from each sentence that was longer than two characters, and not including "the" or "and." Six sets of random cloze items were created for each random condition to minimize the impact of any unusually good or bad random items. As before, the number of sentences and words matched the human condition. On average, this meant that random cloze words generated by this process were more likely to test inference and promote learning than most n^{th} deletion methods.

Natural language processing techniques from Olney et al. (2017) were used to make the machine cloze items. The text was parsed using semantic, syntactic, and discourse parsers. This labeled the text using word form, named entities, syntactic dependencies, verbal and nominal predicates, argument roles, coreference chains, elementary discourse units, and discourse dependencies. Sentences were chosen that had at least three coreference chains that were at least two chains long. The goal was to make sure the sentences were connected to the rest of the text.

This process is equivalent to identifying anaphora, where pronouns are resolved to their referents, and correspondingly involved argument overlap. Sentences were then filtered if their elementary discourse units did not have the meaning of their discourse relationships, i.e., did not contain a discourse nucleus (Carlson, Marcu, & Okurowski 2003). Possible cloze words were then chosen based on whether it was an argument in a coreference chain, a semantic argument, or a syntactic subject or object with a noun. The final cloze words were chosen from this group, but not if they were part of the 1000 most frequent English words. This excluded common words in the circulatory system text like "heart," "blood," and "body" (Olney et al., 2017).

Additionally, several changes were made to the stimuli and procedure from Olney et al. (2017). The random condition in Olney et al. (2017) was not in sequential order, which may have weakened it. For example, instead of presenting the first cloze item about the size of the heart, which is in the first sentence of the text, MoFaCTS could instead first present a cloze item about heart valves, which is at the end of the text. In the present study, stimuli in the random condition were presented in sequential order by default to make it more directly comparable to the human and machine conditions. Another change made was that participants had to practice the cloze items for a minimum of 15 minutes compared to 5 minutes in the previous study. This was changed to give the participants a greater "dose" of instructional cloze items to promote more learning and comprehension. Finally, the MoFaCTS algorithm was updated to provide a more accurate model of learning that made for more optimally scheduled practice.

Procedure

IRB approval was obtained before the start of the study and participants read and agreed to an informed consent page before the study began. Participants recruited through AMT took the experiment on a computer using the web interface of MoFaCTS. Upon agreeing to take the

Human Intelligence Task (HIT), participants were randomized to one of nine conditions and given the link to the MoFaCTS website. Once there, they completed the informed consent. On the next page, they read a summary of the six parts of the experiment (pretest, read text, practice cloze, return in one day, post-test, and demographics). The next page gave basic information on the pretest. Participants were instructed not to take notes or search for answers online and to make the best guess if they didn't know the answer. They then took the pretest which was nine questions: eight randomly selected from the question bank plus one being the attention check. The eight pretest questions were randomly chosen from four of 16 concept clusters, with one fact and one transfer question for each concept. The order of the multiple-choice answers for each question was randomized. Participants had 35 seconds to choose an answer before the system timed them out and displayed the correct answer. Thirty-five seconds was chosen based on the researchers' estimation for how long participants needed to read the longest questions and answers. If MoFaCTS timed out or an incorrect answer was given, the system would display the correct answer for 12 seconds before showing the next question.

After the pretest, participants were instructed to spend five minutes minimum and ten minutes maximum reading the text that followed, and not to take notes or search online. After the reading, participants practiced cloze items based on the condition they were randomized to. The participant typed their word and pressed enter to submit an answer. There are different ways to score a cloze question, including accepting only the exact-word as correct, or some combination of syntactic and semantic similarity to the correct answer (Park, 2011). Our study ruled an answer correct if 85% or more of the letters of the entered word matched the letters of the correct word. The choice not to accept semantically similar words was not as relevant for our study because the score on the cloze items was not used as a measure or dependent variable.

Participants who answered incorrectly in the practice were immediately shown the right answer. Participants had 15 seconds to enter an answer before they timed out. If they timed out or answered incorrectly, the right answer was displayed for eight seconds. If they got the cloze item right, they moved on to a new question. If they got the item wrong, they were more likely to get the item again later. MoFaCTS tried to get every participant to a high practice score and to cycle through every item. After choosing to practice the cloze items for between 15-25 minutes, participants were paid for part 1 and emailed instructions to return for part 2.

Participants had to wait at least 24 hours from completing part one to return for part two. Those who returned within 72 hours took the post-test and had their data analyzed. The post-test format was identical to the pretest except it consisted of 17 out of the remaining 25 questions in the test bank, with one being a required attention check question. After the post-test, participants answered four demographic questions on age, gender, education, and if they had held a job related to the text. Participants were then paid for part 2.

Results

Results could be analyzed using a two-way ANCOVA on post-test with pretest score and correct cloze responses (practice score) as covariates or a two-way ANOVA after converting pretest and post-test to a proportion and calculating proportion post-test – proportion pretest to get proportion gain score, hereafter referred to simply as "gain score." There's a long history of debate over whether to use ANCOVA or gain score (Smolkowski, 2018). After accounting for Cohen's d, sample size, estimated intrinsic association between initial status and growth, and estimated reliability, it may be possible to determine which approach has the least bias (Kelly & Ye, 2017). Calculating the bias for our study ("Comparing Change-Score And ANCOVA," n.d.), we found gain scores and ANCOVA to be equivalently biased but in opposite directions, with gain score negatively biased and ANCOVA positively biased. Thus, other criteria were taken into consideration. Since two of the planned covariates, number of trials and trial score, were correlated with condition, r(560) = .250, *p* < .001, r(560) = .299, *p* < .001, respectively, they couldn't be used for ANCOVA. With ANCOVA more limited in usefulness, gain score was chosen due to its relative simplicity.

Due to experimental error, the texts the participants read were similar to but different from the cloze item text in the nitrogen and photosynthesis conditions. In these conditions, the cloze item text and the text participants read differed with some rewordings of sentences, information shifted to a different part of the text, or new sentences that described the same process in a different way and sometimes with different content words. The number of cloze items made from text which was not identical to the text read was 4/66 for machine-nitrogen, 7/66 for human-nitrogen, 7/74 for machine-photosynthesis, and 11/74 for human-photosynthesis. It's estimated 2-4 test questions in the nitrogen cycle covered similar or same content as the cloze items affected, whereas 6-8 test questions in photosynthesis had similar or same content as the cloze items affected. The cloze items were all made from the same texts, but differences in how the machine or human chose their items may have interacted in unknown ways with differences in the edited version of the texts the participants read. Contrastingly, in the circulatory system condition, there were no differences between the text participants read and the text of the cloze items.

A two-way ANOVA was performed to look at the effects of text and cloze type on gain score. The assumptions of ANOVA were tested by analyzing residuals. Residuals were normally distributed as assessed by Shapiro-Wilk, p < .05. There was homogeneity of variances as assessed by Levene's test for equality of variances, p = .732. Inspection of a boxplot revealed no extreme outliers for gain score. No outliers were removed or transformed for the ANOVA. There was a statistically significant difference in pretest scores between cloze conditions in the photosynthesis text as measured by ANOVA, F(2, 185) = 4.410, p = .013, partial η^2 = .046.

There was no statistically significant interaction between text and cloze on gain score, F(4, 553) = 1.892, p = .110, partial η^2 = .014. The estimated marginal means of text and cloze type on gain score are shown in Figure 1. There was a statistically significant main effect of cloze on gain score, F(2, 553) = 7.349, p = .001, partial η^2 = .026. Pairwise comparisons were run with p-values Bonferroni-adjusted and 95% confidence intervals. The unweighted marginal means of gain score for human, machine, and random were .158 (*SE* = .017), .099 (*SE* = .017) and .067 (*SE* = .017), respectively. The human factor was associated with a mean gain score .06, 95% CI [.002, .117] higher than the machine factor, a statistically significant difference, p = .04. The human factor was associated with a mean gain score .091, 95% CI [.033, .149] higher than the random factor, *p* = .001. There was no statistically significant difference between machine and random factors, *p* = .573. Results are shown in Figure 1.

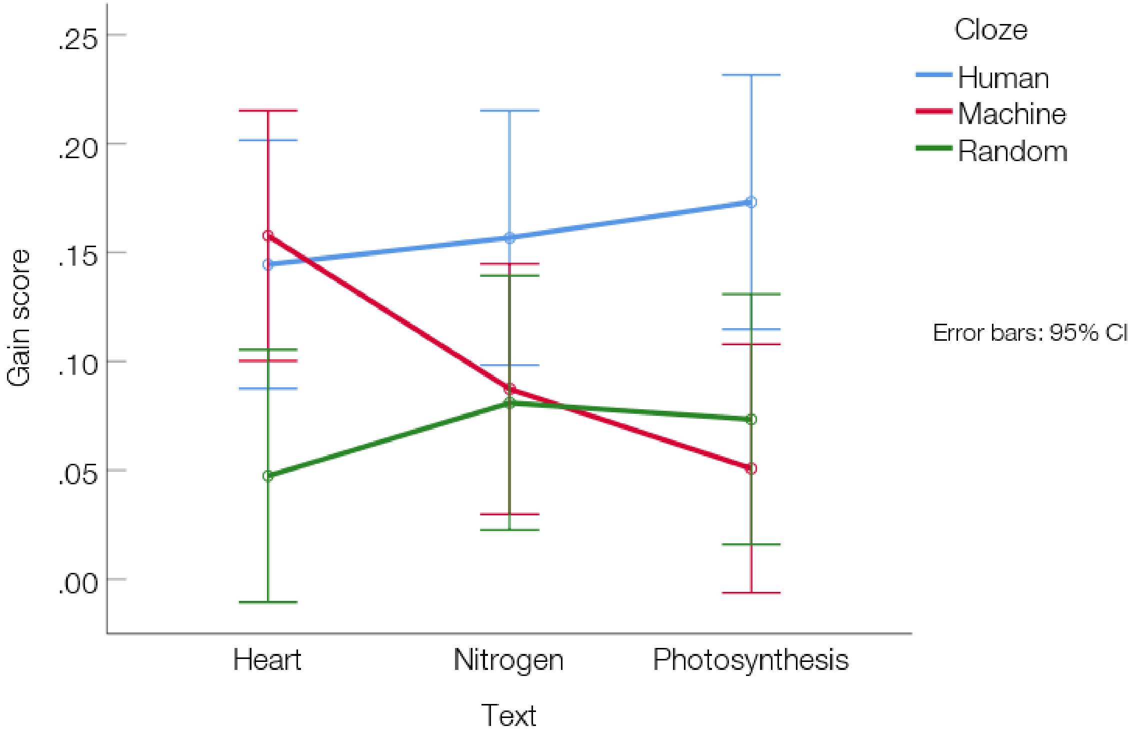


Figure 1. Estimated marginal means of gain score across text and cloze type.

Additionally, two-way ANOVAs were run on both fact and transfer questions to examine the effects of these question types on gain score. The residuals of both fact and transfer gain scores were normally distributed as assessed by Shapiro-Wilk, *p* < .05. There was homogeneity of variances as assessed by Levene's test for equality of variances for fact and transfer gain scores, *p* = .728, *p* = .785, respectively. No extreme outliers were identified in boxplots.

For fact questions, there was no significant interaction between text and cloze on gain score, *F*(4, 553) = 1.403, *p* = .232, η^2 = .010. There was a statistically significant main effect of cloze on gain score, *F*(2, 553) = 4.588, *p* = .011, partial η^2 = .016. Pairwise comparisons were run with p-values Bonferroni-adjusted and 95% confidence intervals. The unweighted marginal means of fact gain score for human, machine, and random were .182 (*SE* = .022), .110 (*SE* = .022) and .092 (*SE* = .022), respectively. The only statistically significant difference was between the human and random factors, with the human factor having a mean gain score .090, 95% CI [.015, .166] higher than the random factor, *p* = .013.

For transfer questions, there was no significant interaction between text and cloze on gain score, *F*(4, 553) = 2.242, *p* = .063, η^2 = .016. There was a statistically significant main effect of cloze on gain score, *F*(2, 553) = 4.439, *p* = .012, partial η^2 = .016. Pairwise comparisons were run with p-values Bonferroni-adjusted and 95% confidence intervals. The unweighted marginal means of transfer gain score for human, machine, and random were .134 (*SE* = .022), .087 (*SE* = .021) and .043 (*SE* = .022), respectively. The only statistically significant difference was between the human and random factors, with the human factor having a mean gain score .091, 95% CI [.018, .165] higher than the random factor, *p* = .009.

Since statistically significant associations were seen in fact and transfer questions on gain score, an ANOVA was performed post-hoc to look for possible interactions between text, cloze, question type (fact or transfer), on gain score. There was a statistically significant interaction between question type and text on gain score, *F*(2, 1106) = 4.648, *p* = .010, η^2 = .008. Analysis of simple main effects for text was done with a Bonferroni adjustment for statistical significance. There was a statistically significant difference in gain score between transfer and fact questions from the photosynthesis text, *F*(1,1118) = 9.535, *p* = .002, partial η^2 = .008. For fact and transfer questions from the photosynthesis text, mean gain score was .096, 95% CI [.035, .158] higher for fact than transfer. For the photosynthesis text, mean gain score for fact questions was .146 (SD = .307) and .05 (SD = .307) for transfer questions.

In order to determine if the results from the original study had been reproduced, a one-way ANOVA was run only on the circulatory system, the text which had been used in Olney et al. (2017). There was a statistically significant main effect of cloze on gain score, $F(2, 186) = 4.136$, $p = .017$, $\eta^2 = .043$. Pairwise comparisons were run with p-values Bonferroni-adjusted and 95% confidence intervals. The unweighted marginal means of gain scores for human, machine, and random were .145 (*SE* = .029), .158 (*SE* = .030) and .047 (*SE* = .030), respectively. The only statistically significant difference was between the machine and random factors, with the machine factor having a mean gain score .110, 95% CI [.009, .212] higher than the random factor, $p = .027$, with the human factor having a mean gain score .097, 95% CI [-.004, .198] higher than the random factor, $p = .063$. Two one-way ANOVAs were run for cloze on fact and transfer gain scores separately within the circulatory system. No statistically significant effect was found for fact or transfer, $p = .121$, $p = .068$, respectively.

Post-hoc tests were run in order to discover additional differences between texts so as to better understand the results. Reading time data could not be analyzed with an ANOVA because it was highly skewed. Instead, a Kruskal-Wallis test was performed to find possible differences in reading times between texts. Distributions of reading times were similar for all texts as shown by visual inspection of a boxplot. Median reading times were statistically significantly different between texts, $\chi^2(2) = 11.049$, $p = .004$. Therefore, pairwise comparisons were run using a Bonferroni correction for multiple comparisons (Dunn, 1964) with adjusted p-values shown. The pairwise comparisons showed statistically significant differences in reading time as measured in seconds between the circulatory (328) and photosynthesis (358) ($p = .003$) texts, but not between nitrogen (338) and any other combination.

A one-way ANOVA of text on pre-test score was also run posthoc to determine if any text may have been more difficult. Data from the texts were not normally distributed as assessed by Shapiro-Wilk, $p < .001$. However, visual inspection of the histograms and Q-Q plots revealed no extreme outliers and data looked normally distributed and unimodal ($N = 562$), so the decision was made to proceed. Inspection of a boxplot revealed no extreme outliers. There was homogeneity of variances as assessed by Levene's test for equality of variances, $p = .165$. There was a statistically significant main effect of text on pretest score, $F(2, 559) = 23.671$, $p < .001$, partial $\eta^2 = .078$. Pairwise comparisons performed with p-values Bonferroni-adjusted and 95% confidence intervals revealed that pre-test score for circulatory system ($M = .487$), 95% CI [.458, .515] was significantly higher than photosynthesis ($M=.354$), 95% CI [.326, .383], $p < .001$ and nitrogen cycle ($M = .378$), 95% CI [.350, .407], $p < .001$.

Discussion

Cloze items have a long history of use in schools, both to evaluate and assist learning. Much research has concentrated on using cloze tasks to evaluate reading comprehension in education with comparatively less focus on using it as an intervention to improve comprehension as we did. Olney et al. (2017) introduced a novel way to create cloze items with results suggesting that a machine could create cloze items for practice as well as a human. This research tried to replicate that study and see if their results could extend to other texts.

Within the circulatory system text, the machine condition had statistically significant higher gain scores than the random condition, which suggests those results from Olney et al. (2017) were partially replicated. Caution is warranted though as that study used ANCOVA as the method of analysis whereas we used an ANOVA with gain scores. For example, within the circulatory system text, Olney et al. (2017) found that machine had higher post-test scores relative to human, but controlled for variables we could not. Instead, our pairwise comparisons found machine superior to the random but not to human. Although the estimates of marginal means of gain scores for machine were higher than human in this condition, the scores were not statistically significant. However, if the results of the two studies are not directly comparable because of the different analyses, they still seem to be in agreement that the machine-generated cloze items were better than the random cloze items in the circulatory system text.

In contrast to the circulatory system, we found the human had higher gain scores than the machine and random conditions in the two new texts of nitrogen cycle and photosynthesis. Figure 1 illustrates how the human conditions performed consistently well across all texts compared to machine or random conditions, whereas the machine performed well in the circulatory system but faltered in the nitrogen cycle and photosynthesis texts. This finding was unexpected but not unsurprising, as it wasn't known how the machine would fare when making new cloze items out of these two new texts. This suggests that there is something different about the nitrogen cycle and photosynthesis texts compared to the circulatory system that gave the machine algorithm problems.

The two texts of nitrogen cycle and photosynthesis were made by a researcher in the present study, whereas the circulatory system was created by a different researcher from Olney et al. (2017). In creating the new texts for our study, an effort was made to make them similar to the circulatory system in terms of the number of words, Flesch-Kincaid, and Lexile. However, that does not mean there are not still significant differences uncaptured by these metrics. The topics of the texts may fall very broadly within science but are from different domains. For example, the nitrogen cycle may tap into multiple domains of knowledge involving the atmosphere, soil, plants, water, and farming, while the photosynthesis text takes place inside plant cells and is much narrower in scope but is highly detailed. It is possible that despite the similar word counts and reading levels of the text they contain different amounts of information that is not captured by a Lexile or Flesch-Kincaid score. For instance, it is possible the information in one of the texts is more difficult to comprehend, with harder concepts to grasp, and requires more sentences to be reread. Support for this would be that mean reading time was 30 seconds longer for photosynthesis than the circulatory system.

Our finding that pretest scores for nitrogen cycle and photosynthesis were significantly lower than the circulatory system could also fit this interpretation. It could be that the photosynthesis pretest questions were harder because they reflect a more difficult text. Or, the causality could be flipped, such that worse scores on the pretest for photosynthesis motivated participants to read the text for a longer time. Unusually difficult pretest questions could cause longer reading times instead of a more difficult text. Different researchers were responsible for creating pretest and post-test questions for the circulatory system compared to the nitrogen cycle and photosynthesis, and differences on pretests may reflect different approaches to creating the questions. One other possibility is a reader hypothesis, that lower pretest scores show that participants came into the study with less prior knowledge of the nitrogen cycle or photosynthesis compared to the circulatory system text, and prior knowledge does affect the reading of science texts (Kendeou et al., 2007; Ozuru et al., 2009). Other than a pretest, we did not measure the prior knowledge or reading ability of our participants.

In this context, it is interesting that the teacher we recruited made 66 and 74 cloze items for nitrogen and photosynthesis respectively, which is more than the 45-55 items for each text she was encouraged to create. This may suggest the teacher thought the new texts were sufficiently complicated they deserved more cloze items than the circulatory text (53), but it should also be remembered that circulatory cloze items were created by a different researcher and that these numbers may just reflect their own unique method. The nitrogen cycle had 7/26 shared sentences between its human and machine conditions, with 10/66 cloze items being the same. Similarly, photosynthesis had 8/24 sentences shared between human and machine conditions, with 18/74 cloze items being the same. The circulatory system had 13/21 shared sentences between human and machine conditions used to make cloze items.

Overall, the results on reading time and pretest scores from posthoc tests suggest that the texts differed in readability or difficulty. Why a less readable or more information dense text would pose problems for the generation of cloze items by the machine and not the human is unclear. Any gain scores differences between cloze conditions should be attributable to the sentences and words not shared between them in cloze practice, making an examination of these unshared cloze items warranted. Since our study copied the circulatory system materials used in Olney et al. (2017), we know that 13 out of the 21 sentences used to make cloze items were shared between machine and human in the circulatory system text. If the previous study is any indication, then differences in coreference chains in the new texts and how connected the cloze items are to the discourse would be interesting to explore.

Finally, we tested whether fact or transfer questions could promote learning across conditions. This is relevant because transfer questions should require inferences to be made to answer correctly, and cloze tasks using rational deletion may also require inferences. The Olney et al. (2017) study used ANCOVA to look at transfer post-test proportion correct with cloze condition, pre-test score, number of trials, proportion correct across trials, and the interaction of number of trials and post-test proportion correct as predictors and found a main effect of condition, pre-test proportion correct, and number of trials. In that study, posthoc comparisons using Tukey's HSD found that machine cloze had significantly higher transfer post-test proportion correct than the human and random conditions. Our study used an ANOVA to find a main effect of cloze on gain score and this same pattern of effects was found for transfer and fact gain scores. Human conditions had statistically significant higher transfer or fact gain scores than random conditions, although the differences between human and machine conditions for fact or transfer were not statistically significant. These results do not replicate Olney et al. (2017) where the machine conditions scored higher on transfer than the human or random conditions, but they do follow the same pattern where the cloze condition with the highest gain scores overall also has the highest transfer gain scores. The interaction between text, and question type (fact or transfer) on gain score which found a simple main effect of fact and transfer in the photosynthesis text was not predicted. It is unclear exactly why participants would achieve lower transfer gain scores than fact within this text alone. Since the interaction between text, cloze, and question type on gain scores was not significant, it is less likely that the machine-generated cloze items are responsible for why transfer scores were lower. This result does seem to add to the evidence that the photosynthesis text is different in another important way.

We must add that an important limitation to the study is how participants read parts of the text that weren't used to create cloze items. It is unknown what effect, if any, this may explain the differences found between conditions. Converting proportional gain scores into raw post-test – pretest scores gives a difference of 1.11 test questions between human and machine conditions in the nitrogen cycle, and 1.96 test questions difference between human and machine for photosynthesis. In theory, because up to 6-8 questions were possibly affected by the text-cloze differences for photosynthesis, and 2-4 for nitrogen, the results might be easily explained by this limitation. But it is unclear how powerful this effect could be or which direction it could be biased towards. The results from the circulatory system conditions were not at all affected or limited by this.

We managed to partially replicate Olney et al. (2017) with similar results in the circulatory system conditions. However, the machine failed to equal human in the photosynthesis and nitrogen cycle conditions. Analysis of fact and transfer had the same pattern of results, with the human outperforming the machine and with lower transfer gain scores compared to fact for photosynthesis. There are many possible reasons why the machine-generated cloze items weren't as effective in the two newer texts. It is possible but speculative that the two new texts are different in a fundamental way from the circulatory system. What this research shows is that a computer algorithm can be as effective as a human teacher in making cloze items from one type of text, but not as effective in others.

Conclusion

This study attempted to replicate the results in Olney et al. (2017) and test whether cloze items created by natural language processing techniques could effectively improve reading comprehension when presented by MoFaCTS, an adaptive practice scheduling system. Results suggest that such techniques used in the machine condition of the circulatory text could significantly improve comprehension compared to random cloze generation. Unfortunately, these findings did not extend to the new texts of photosynthesis and the nitrogen cycle, where the human condition had higher gain scores than the machine and random conditions. How the differences between texts may have made the machine less effective at generating cloze items is unknown. This research is important because quality cloze items being automatically generated from any text could be used to improve educational outcomes in reading comprehension. Understanding the type of texts the machine can best generate cloze items from and how to improve it is a subject for future research.

References

Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *Modern Language Journal*, *76*(4), 468-479.

Alderson, J. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*(2), 219-227.

Alonzo, A. C. (2002). Evaluation of a model for supporting the development of elementary school teachers' science content knowledge. Proceedings of the *Annual International Conference of the Association for the Education of Teachers in Science.* Charlotte, NC. January 10-13, 2002.

Bachman, L. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*(3), 535-556.

Bormuth, J. R. (1965), Optimum sample size and cloze test length in readability. *Journal of Educational Measurement*, *2*(1), 111-116.

Bridge, C., & Winograd, P. (1982). Readers' awareness of cohesive relationships during cloze comprehension. *Journal of Literacy Research*, *14*(3), 299-312.

Brown, J. D. (2013). My twenty-five years of cloze testing research: So what? *International Journal of Language Studies*, *7*(1), 1–32.

Carlisle, J. F., & Rice, M. S. (2004). Assessment of reading comprehension. In Stone, C. A. (2006*). Handbook of language and literacy : Development and disorders.* New York : Guilford Press.

Carlson L., Marcu D., & Okurowski M.E. (2003). *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory.* In van Kuppevelt J., Smith R.W. (Eds.), *Current and New Directions in Discourse and Dialogue. Text, Speech and Language Technology* (Vol. 22, pp.85-112). Springer, Dordrecht.

Comparing Change-Score And ANCOVA Methods of Analyzing Data. (n.d.). Retrieved from https://lilygray.github.io/ancova-change-score/

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 99(2),* 311-325.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243.

Durkin, D. (1993). *Teaching them to read*. Boston, MA: Allyn and Bacon.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6,* 241-252.

Englert, C.S., Zhao, Y., Collings, N. & Romig, N. (2005). Learning to read words: The effects of internet-based software on the improvement of reading performance. *Remedial and Special Education, 26*(6), 357-371.

Fang, Z. (2006). The language demands of science reading in middle school. *International Journal of Science Education*, *28*(5), 491-520.

Gellert, A. S., & Elbro, C. (2013). Cloze tests may be quick, but Are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment, 31*(1), 16–28.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082-1116. doi:10.3102/0034654317726529

Grant, P. (1976). *The cloze procedure for improving sixth grade students' reading comprehension and understanding of social studies materials.* Unpublished master's thesis, Rutgers University, New Brunswick, New Jersey.

Greene Jr., B. B. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading*, *24*(1), 82-98.

Harris, T.L., & Hodges, R.E. (1995). *The Literacy Dictionary: The vocabulary of reading and writing*. Newark, DE: International Reading Association.

Johnson, N. F. (1970). The role of chunking and organization in the process of recall. *Psychology of Learning and Motivation,* 4, 171-247. 1970.

Jongsma, E. A. (1980). *Cloze instruction research : A second look*. Newark, Del. : International Reading Association

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly, 24*(1), 61-83.

Kelly, S., & Ye, F. (2017). Accounting for the relationship between initial status and growth in regression models. *Journal of Experimental Education*, *85*(3), 353–375.

Kendeou P, van den Broek P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory and Cognition*, *35*(7), 1567–1577.

Kitao, K., & Kamiya, K. (2009). Using cloze generator to make cloze exercises. *International Journal of Pedagogies & Learning*, *5*(2), 67-79.

Kobayashi, M. (2002). Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *The Modern Language Journal, 86*(4), 571-586.

Lipson, M. Y. (1982). Learning new information from text: The role of prior knowledge and reading ability. *Journal of Reading Behavior*, *14*(3), 243-261.

Nagy, W., Townsend, D., Lesaux, N., & Schmitt, N. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*(1), 91-108.

National Institute of Child Health and Human Development. 2000. *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,* Washington, DC: U.S. Government Printing Office. (NIH Publication No. 00–4769)

Olney, A. M., Pavlik, P. I., & Maass, J. K. (2017). Improving reading comprehension with automatically generated cloze item practice. Lecture Notes in Computer Science. *Artificial Intelligence in Education*, 262-273. doi:10.1007/978-3-319-61425-0_22

Ozuru, Y., Dempsey, K., & McNamara, D.S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19*(3), 228-242.

Park, C. (2011). Making cloze tests more valid. *Journal of the Korea Academia-Industrial Cooperation Society*, *12*(2), 640-645.

Pavlik, P.I., Kelly, C., & Maass, J.K. (2017). The mobile fact and concept training system (MoFaCTS). *Intelligent Tutoring Systems Lecture Notes in Computer Science*, 247–253.

Pavlik Jr., P. I., Presson, N., Dozzi, G., Wu, S.-m., MacWhinney, B., & Koedinger, K. R. (2007). The FaCT (Fact and Concept Training) System: A new tool linking cognitive science with educators. In D. McNamara & G. Trafton (Eds.). *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society,* (pp. 397–402)

Pessah, N. (1975). The effect of various teaching techniques, involving the cloze procedure, upon the reading achievement of community college students. Annual Meeting of the International Reading Association. New York City, NY. May 13-16, 1975.

Porter, D. (1983). The effect of quantity of context on the ability to make linguistic predictions: A flaw in a measure of general proficiency. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 63-74). London: Academic Press.

Sampson, M. R., Valmont, W. J., & Van Allen, R. (1982). The effects of instructional cloze on the comprehension, vocabulary, and divergent production of third-grade students. *Reading Research Quarterly*, *17*(3), 389-399.

Seamon, M. P., & Levitt, E. J. (2003). Quiz makers: A product comparison of Hot Potatoes, WebPractest, & Interactive Exercise Makers. *Library Media Connection*, *21*(5), 59-61.

Smolkowski, K. (2018, October 1). Gain score analysis. Retrieved from https://homes.ori.org//keiths/Tips/Stats_GainScores.html

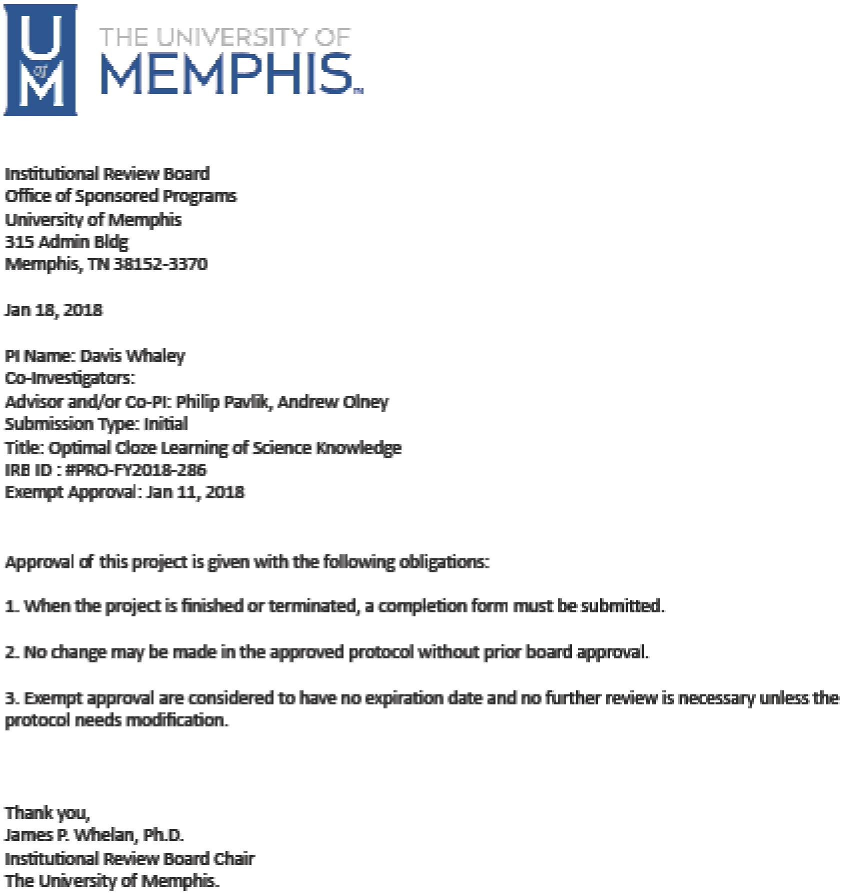
Stein, M., Larrabee, T. G., & Barman, C. R. (2008). A study of common beliefs and misconceptions in physical science. *Journal of Elementary Science Education*, *20*(2), 1–11.

Taylor, W. L. (1953). "Cloze Procedure": A new tool for measuring readability. *Journalism Quarterly*. 30, 415-433.

Watanabe, Y., & Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies, 26(2)*, 103–133.

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes: A Multidisciplinary Journal*, *25*(2–3), 309–3

Appendix



THE UNIVERSITY OF MEMPHIS

Institutional Review Board
Office of Sponsored Programs
University of Memphis
315 Admin Bldg
Memphis, TN 38152-3370

Jan 18, 2018

PI Name: Davis Whaley
Co-Investigators:
Advisor and/or Co-PI: Philip Pavlik, Andrew Olney
Submission Type: Initial
Title: Optimal Cloze Learning of Science Knowledge
IRB ID : #PRO-FY2018-286
Exempt Approval: Jan 11, 2018

Approval of this project is given with the following obligations:

1. When the project is finished or terminated, a completion form must be submitted.

2. No change may be made in the approved protocol without prior board approval.

3. Exempt approval are considered to have no expiration date and no further review is necessary unless the protocol needs modification.

Thank you,
James P. Whelan, Ph.D.
Institutional Review Board Chair
The University of Memphis.