An Evaluation of the Testing Effect with Third Grade Students

Chandra L. Brojde (chandrab@colorado.edu)
Department of Psychology, CB 345
University of Colorado, Boulder, Colorado 80309

Barbara W. Wise (Barbara.wise@colorado.edu)
CLEAR Computational Language and EducAtion Research, CB 594
University of Colorado, Boulder, Colorado 80309

Abstract

Previous research shows that being tested after studying a set of material will improve recall of that material on a later test. While this effect has been shown to be robust in adults using both simple material, such as word lists, and complex texts, little research has been done looking at this effect in children. In the current study, third-grade students were asked to read stories while being tested with multiple-choice questions requiring inferences and understanding of the main ideas, with questions requiring knowledge of facts, or with no questions. Afterwards, they were asked to give a summary of the text and answer short open ended questions. Results showed that children did better on the summary and open-ended questions after having answered questions than not. However, there was no difference between inference/main idea questions and fact questions. These results suggest that the testing effect is a robust effect that transcends type of question and age.

Keywords: Testing effect; Recall; Memory; Development; Comprehension; Reading; Education

Introduction

Traditional educational settings have always relied heavily on testing as a means of evaluating student learning. However, in recent years, researchers have begun to recognize that testing is not only a good way to evaluate what a person has learned but also to support further learning (for a review McDaniel, Anderson, Derbish, & Morrisette, 2007; Roediger & Karpicke, 2006a). Specifically, research shows that testing an individual while they are learning a set of material or immediately after their initial study session produces a strong learning effect such that retention of that material is greater when assessed at a later time than if no initial testing had occurred. This effect is often called the "testing effect".

The testing effect has been studied extensively with undergraduate students using a wide variety of material, including word lists (Tulving, 1967) and texts from educationally relevant material (Roediger & Karpicke, 2006b). However, this effect has yet to be extensively studied in lower levels of education, especially in real classroom settings with educationally relevant material. The study reported here is a first attempt at addressing two questions related to this issue. First, we ask whether or not elementary age students also show this testing effect using educationally relevant material. Specifically, we evaluated whether initial testing of material using multiple-choice questions leads to better retention of that material on later assessments using texts similar to those used in real classroom settings. A second question asked in this study was whether or not an overall testing effect (having questions during the initial reading of the story or not) would be mediated by the type of questions children encountered while reading the story. This is an important issue as the testing effect may be due to participants creating more elaborate memories (e.g. integrating material with background information, building main ideas of the story) during the initial testing. This more elaborate network of connected ideas would then, presumably, be helpful for later retrieval at the final test. In particular, we asked whether multiple choice questions that required children to have a more elaborate deeper understanding of a text would yield stronger retention than questions that simply required them to access more superficial facts.

The Testing Effect

The first goal of this study was, then, to evaluate whether or not young children (third grade students) would show an overall testing effect. Very little research exists at this educational level as the testing effect was originally described as a way to aid learning in word list studies and paired-associate studies with undergraduate students. In these domains, the testing effect has been shown to be a very robust effect. It is still unclear, however, whether the testing effect will be as robust for young children as it is for adults given that reading comprehension skills continue to develop throughout childhood. Testing may not be an efficient means to support reading comprehension at this grade level. In addition, other factors such as smaller working memory capacities, which has been shown to be related to reading comprehension (e.g. Gottardo, Stanovich, & Siegel, 1996; Haarmann, Davelaar, & Usher, 2003; Siegal & Ryan, 1989), may limit the size of the testing effect.

In studies designed to test this effect, participants are often asked to study a set of material (e.g. a list of words or a passage of text). They are then tested on the material or asked to re-study the material. Finally, all participants are given a final test, or post-test, on the material. Results consistently show that if the final post-test is given immediately, participants do better after having re-studied the material than having been given a pre-test. However, recall of the material long-term (e.g. 1 week later) is best after having been pre-tested on the material rather than having re-studied

the material (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b; Wheeler & Roediger, 1992). Moreover, several pre-tests are better than one pre-test, suggesting that initial testing causes students to make more elaborative representations of the material and therefore reduces the amount of forgetting (Wheeler & Roediger, 1992).

In addition to the implications of the testing effect, several recent studies have begun to provide support for the use of testing in educational contexts. For example, Roediger and Karpicke (2006b) asked undergraduate students to study passages of text describing general scientific phenomenon after which they were given a free-recall test or they were asked to re-study the passage. On a final recall test, students who had re-studied the material did better after a 5 minute delay, but did worse than students given the initial recall test after a 2 day or 1 week delay. These results, among others (Bangert-Drowns, Kulik, & Kulik, 1991; Butler & Roediger, 2007; McDaniel et al., 2007; McDaniel, Roediger, & McDermott, 2007) , suggest that the testing effect found in the laboratory can be generalized to an applied educational setting.

Reading Comprehension

Research on test-taking in younger children has mainly 1) focused on basic research tasks, such as learning lists of words (Gates, 1917), 2) has been done using complex material only relevant at the high-school level (Bangert-Drowns et al., 1991), or 3) has focused on testing children's ability to comprehend material or using testing to provide further support for comprehension of that material.

Research focusing on testing material as a way to further aid reading comprehension suggests that the format, content, and design of test questions may be important for comprehension of the material (Kintsch, 2005; Wise, Van Vuuren, Struemph, & Richey, 2006). This leads to the second question addressed by this research: whether or not the content of the initial questions will mediate the overall testing effect.

Recent research suggests that questions that promote greater learning require children not only to know details from the material, or facts, but also require them to make inferences about the material and understand the main idea of the text (Mannes & Kintsch, 1987) (for a more detailed description of inference and main idea type questions see Kintsch, 2005). These inference and main idea questions are thought to promote comprehension because they require the reader to connect details from the story in new ways, understand the main point or gist of the story, and think beyond the material that is learned to include previous background knowledge (Cain & Oakhill, 1999; Oakhill, 1984). One can think of these questions as promoting a *deeper* more elaborate understanding of the text. Fact questions on the other hand only require the reader to remember explicit details or facts from the story (Beck & McKeown, 1981). This can be thought of as a *surface* understanding of the text. Because inference and main idea questions are often thought to promote greater learning, we wondered whether questions requiring a deeper understanding would translate into a greater testing effect because more elaborate memories would be created when initially learning the material. A secondary question, then, was whether students would recall more on a later final test if they were given inference and main idea questions as opposed to fact questions while reading.

Study Overview

In order to investigate 1) whether the testing effect does indeed occur in younger children and 2) whether this effect, if present, depends on the type of question asked while learning the material (inference/main idea vs. fact), we asked each student to read three different texts: once with inference and main idea questions, once with fact questions, and, as a control, once with no questions. Immediately after reading the stories (and answering the multiple choice questions where appropriate), students were asked to give an oral summary of the story and answer 10 short-answer questions about the text that they had just finished reading. Final testing was conducted immediately rather than after a retention interval as children at this age often show difficulty with reading comprehension while reading a text, let alone after a delay, suggesting that the testing effect may be largest immediately. We are, however, currently planning follow up studies to investigate the effect of including a retention interval.

We chose third-grade students as our participants in order to assess the testing effect at lower grade-levels than has been done previously. In addition, the third-grade is typically cited as a critical time for these reading comprehension skills to develop (Snow, 2002). We also chose to test word reading and reading comprehension to ensure that students were at an appropriate reading level for the third grade.

Method

Participants

Eighteen children (eight males and ten females) with an average age of nine years and one month participated. All children were contacted in their third grade classrooms during the second half of the school year.

Materials

Three stories written at the third grade level from the Houston Museum of Science Horizon Plus Science Stories Series (Summers & Conant, 1992) were used in the study. Each story aimed to teach a lesson in one or more area of science (e.g. weight/measurement or plant reproduction) using a short narrative sequence. Each story contained from 900 to 1,000 words.

In addition to the stories, 10 multiple-choice questions were created following guidelines developed for a related study (Kintsch, 2005; Wise, Van Vuuren, Cole, & Kintsch, 2003; Wise et al., 2006). For each story, five questions were fact questions that asked about details provided in the

story and five questions were inference or main idea questions that required a deeper understanding of the text. Inference and main idea questions required students to integrate information from several points in the story, understand implicit information hinted at in the text, or speculate about events that had not yet occurred in the story. Each question consisted of the stem and four alternative choices – one correct answer and three distractors.

The books and multiple-choice questions were presented via an interactive computer program (for a detailed description see Cole, Wise, & van Vuuren, 2006; Wise et al., 2007) designed to improve children's reading. In addition to recording the child as they read the story orally, this program allowed the child to select words by clicking on them with a mouse so that the word was pronounced by the interactive tutor. This program also administered the multiple-choice questions, allowing the children to answer the question more than once until they obtained the correct answer. The system also provided children with feedback on their success at answering the questions. This same program was also used to record children's final oral summaries and answers to questions about the story.

Children were also asked to complete two tests to assess their reading skill level before beginning the study. Word reading was tested using the word reading sub-section of the Wide Range Achievement Test – 3 (WRAT3) (Wilkinson, 1989). This task simply required the child to read as many words correctly as possible with no time constraint. Second, reading comprehension was tested using the passage comprehension sub-test of the Woodcock Johnson – III (WJ-III) (Woodcock, 1987). This task simply consisted of reading short sentences or passages and filling in an appropriate word to complete them.

Design and Procedure

Each child read three stories, one week apart for three weeks. The three stories differed only in the questioning that occurred while the children were reading: 1) in the fact condition, children read a story and answered five *fact* multiple choice questions, 2) in the inference/main idea condition, children read a story and answered the five *inference/main idea* multiple choice questions, and 3) in the no question condition, children read through the story but received no multiple-choice questions. Thus, type of question was a within-subjects factor. In each of the first two conditions, children answered three of the questions while reading and two questions directly after having finished the story. In the fact condition, all questions asked about details in the story. In the inference/main idea condition, the first three questions required inferences and the last two questions at the end of the story required the child to understand the main ideas from the entire story. The story was not available while children answered these multiple-choice questions.

All children read the story orally and were recorded while reading. Each story was prepared for each of the three conditions and each version of each story was assigned to one of three different conditions in a counterbalanced fashion. Immediately following reading, children were asked to summarize the story and answer short open-ended questions.

Introduction Phase. At the beginning of the first session, children completed the word reading and reading comprehension test. In addition, the experimenter walked the child through an example story to familiarize them with the layout of the books and the interactive environment. Children were allowed to click on words in the story that they did not know and the interactive tutor pronounced the words for them. As monitored by the experimenter, children who struggled were reminded that they could have the words pronounced.

Reading Phase. During each of the three sessions, children were given a headset with microphone to wear as they read the story. This allowed the program to record children as they read. Each story was five pages long including graphics depicting plot points in the story.

Testing Phase. At the end of each story, and after all multiple choice questions were answered, the children were then asked to give an oral summary of the story they had just finished reading. Although most children readily supplied a detailed summary, children were prompted to give more information about a topic if they were reluctant to answer or gave a short one to two sentence summary.

Next, when children indicated that they were done with their summary, they immediately answered the 10 short open-ended questions that required an answer only one or two words in length. These questions asked about the most prominent themes in the story and represented the content of the multiple-choice questions from the two experimental procedures equally.

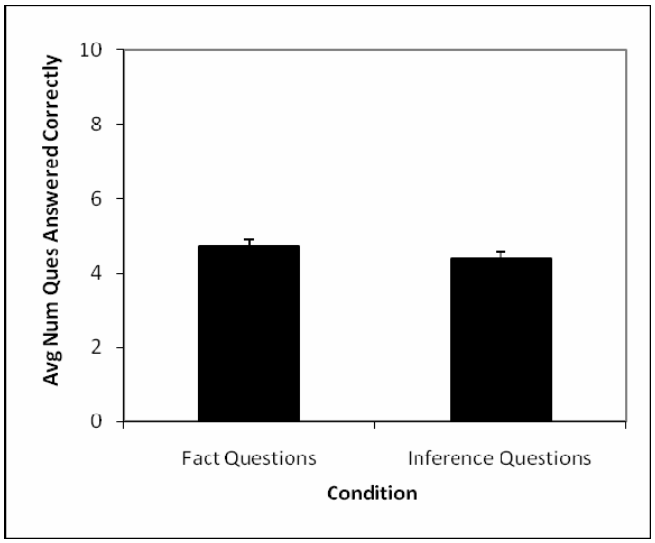

Figure 1. Mean number of multiple-choice questions answered correctly on the first try according to condition. Error bars reflect the standard error of the mean.

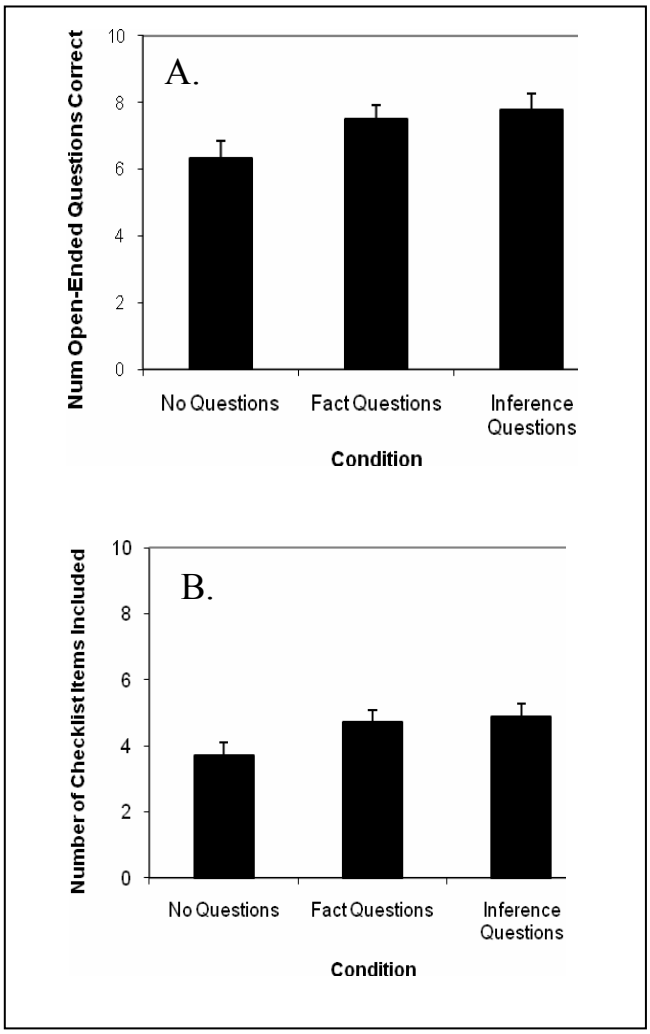

Figure 2. The number of correct items at the final testing for both the A) summary and B) open-ended questions measure. Error bars reflect the standard error of the mean.

Grading

Summaries obtained from the participants were transcribed. These transcriptions were analyzed using a 10-point checklist that addressed similar material to the end-of-story questions. This was assumed to be an acceptable measure of content as there was a high correlation between the number of end questions answered correctly and scores on two reading skill measures (word reading, $r(18) = .52$, $p = .03$, and reading comprehension, $r(18) = .72$, $p < .01$). Each summary was given a score between zero and ten according to the number of checklist items addressed in the summary.

Likewise, children were given a score from zero to ten based on the number of end-of-story questions answered correctly.

Results

Reading Ability

The average standard score of word reading ability (as measured by the WRAT-R) was 103.11 ($SD = 2.81$) with an average grade level equivalent of approximately third grade ($M = 3.67$, $SD = .37$). In addition, the average grade level of reading comprehension (as measured by the WJ-III) was equivalent to the fourth grade ($M = 4.53$, $SD = .42$). Thus, on average children were at or above their grade level in reading skill

Performance on Multiple-Choice Questions

In all of the following analyses, both word reading ability and reading comprehension ability were controlled by using them as covariates. However, similar results obtained without these covariates.

In order to investigate children's ability to answer the multiple-choice questions correctly, a paired-samples t-test was conducted to look for any differences between number of multiple-choice questions answered correctly for the fact and inference/main-idea question conditions. Children were considered as having successfully answered the question only when they chose the correct answer on the first try. For each condition, a score of 5 indicated a perfect score. A non-significant trend showed that children did better on the fact questions ($M = 4.72$, $SD = .75$) than the inference/main-idea questions ($M = 4.39$, $SD = .70$), $t(17) = 2.062$, $p = .06$ (see Figure 1). This suggests that the fact questions may have been easier for the children to answer correctly right away than the inference questions. However, in both conditions children did quite well, answering, on average, more than 4 questions correctly (out of 5) on their first try.

Testing Effect

The testing effect was evaluated using both the summaries that children provided as well as the open-ended questions. First, summaries were graded from zero to ten for each child for each condition. A one-way repeated-measure ANOVA with planned contrasts showed that there was a significant difference between conditions with questions and the condition without questions, $F_{(1,15)} = 4.54$, $p = 0.05$, such that scores were higher for conditions with questions ($M = 4.81$, $SD = 1.58$) than the condition without questions ($M = 3.72$, $SD = 1.67$). However, there was no significant difference between the fact condition and inference/main idea question condition, $F_{(1,15)} = 0.11$, $p = 0.75$ (see Figure 2a).

Similarly, for the open-ended questions, children were given a score from zero to ten for each condition. A second one-way repeated-measures ANOVA with planned contrasts showed that there was a significant difference between conditions containing multiple-choice questions and the condi-

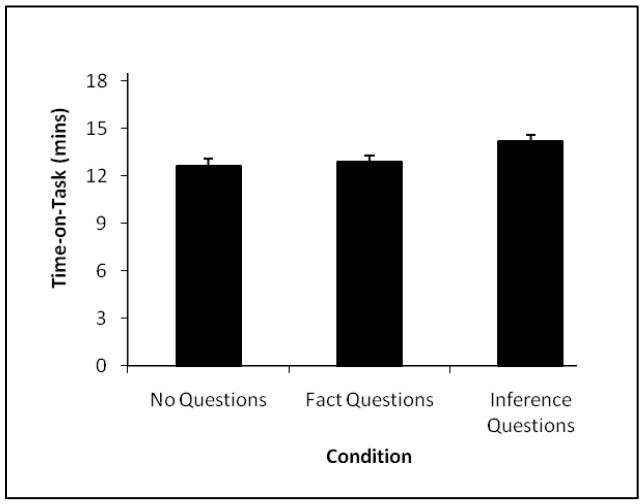

Figure 3. Average amount of time (in minutes) spent reading and answering multiple-choice conditions per condition. Error bars reflect the standard error of the mean.

tion that did not, $F_{(1,15)} = 5.802$, $p = 0.03$, such that when children read stories containing multiple-choice questions ($M = 7.6$), they answered significantly more of the end questions correctly than did children who did not answer multiple-choice questions ($M = 6.3$) (see Figure 2b). In addition, there was again no significant difference between the number of end questions answered correctly after answering inference/main idea questions or answering fact questions, $F_{(1,15)} = 0.195$, $p = 0.67$.

Time-on-task

A final analysis addressed the amount of time that children spent reading and answering questions for each condition. Results obtained here may have been due to children spending more time with the material overall in the question conditions. However, a one-way repeated measure ANOVA showed no significant difference in the overall amount of time spent with the material in the three conditions, $F_{(2,34)} = 1.67$, $p = 0.20$ (see Figure 3).

Discussion

One goal of this study was to evaluate the testing effect in a group of third-grade students. A second goal of this project was to evaluate the testing effect using different types of questions. To address these two issues, students were asked to read stories with questions (inference/main idea or fact questions) or without questions. After finishing the story, students then immediately produced oral summaries of the story and answered short-open ended questions. Overall, the results clearly demonstrate that 1) the testing effect does generalize to lower levels of education, third-grade in this case, 2) the testing effect is just as strong using inference/main idea building questions as it is using fact or detail questions and 3) the testing effect is as strong when using either summary or short-answer question format at the final test.

The fact that these children showed a robust testing effect in both summary content and open-ended questions at the final test is particularly interesting given that the children 1) only answered five quick multiple-choice questions while reading the text and were tested immediately after reading and 2) showed no difference in overall amount of time spent with the material across conditions (most likely because the multiple-choice questions were very short). Testing is no doubt important for promoting comprehension – a statement that is clearly consistent with recent assertions that questions are useful for both enhancing learning as well as assessing it (Beck & McKeown, 2001; Graesser & Person, 1994; Kintsch, 2005).

It is, however, surprising that children did not show a larger testing effect for inference/main idea questions than for fact questions (Beck & McKeown, 1981; Kintsch, 2005; Oakhill, 1984). There are several reasons why we might have seen this pattern. First, it may be that the cause of the testing effect is not related to the participants' creating more elaborate representations of the text, but rather due to something else, such as re-rehearsal of the material.

A second potential reason for no difference between fact and inference/main idea questions is that these questions may lead to more elaborate representations only after a delay. Inference and main idea questions may promote comprehension of text through elaboration processes that require time to have a significant effect. In addition, the process of summarizing the story right after reading it may have led to the same kind of integration of material that we hoped would be caused by the multiple-choice questions. Perhaps the immediate post-testing overwhelmed any difference of the two types of questions during reading.

Thus, it is likely that a delay between the initial test and the final test will show a greater testing effect given several studies showing that the testing effect, in adults, at least, is larger after a delay (e.g. McDaniel et al., 2007; McDaniel et al., 2007; Roediger & Karpicke, 2006a). In adults, the testing effect is weaker, if present at all, immediately after learning a set of material but is strongest after a delay (Roediger & Karpicke, 2006b). It is still an open question, however, whether a delay will lead to a greater testing effect following questions that promote elaboration (inference/main idea questions).

Thus, it will be important in future research to extend the current study to include a delay between the initial reading/testing session and the final testing. It is particularly remarkable that children were able to recall more information about a story after having answered five multiple choice questions, but it would be even more remarkable if this effect lasted or was potentially greater after a delay, especially if this were true more for inference/main idea questions than fact questions. This is an important question that we plan to address in future studies.

These results thus add two important pieces of information to the current literature on reading in children. First, testing children while reading, even if only with basic multiple-choice questions, is beneficial to later recall of the ma-

terial. Second, these questions can be written to require simple fact knowledge of the text or to require children to make inferences and understand the main idea – any type of question is better than no questioning in terms of later recall.

Acknowledgments

We would like to thank Richard Olson and Eliana Colunga for comments on a prior draft and Eileen Kintsch, Sarel Van Vuuren, Ron Cole, Taylor Struemph and Nattawut Ngampatipatpong for help design the materials used in this project.

References

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85*, 89-99.

Beck, I. L., & McKeown, M. G. (1981). Developing questions that promote comprehension: The story map. *Language Arts, 58*, 913-917.

Beck, I. L., & McKeown, M. G. (2001). Test talk: Capturing the benefits of read-aloud experiences for young children. *The reading teacher, 55*, 10-20.

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*, 514-527.

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing, 11*, 489-503.

Cole, R., Wise, B., & van Vuuren, S. (2006). How Marni teaches children to read. *Educational Technology, 47*, 14-18.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*.

Gottardo, A., Stanovich, K. E., & Siegel, L. S. (1996). The Relationships between Phonological Sensitivity, Syntactic Processing, and Verbal Working Memory in the Reading Performance of Third-Grade Children. *Journal of Experimental Child Psychology*, 63(3), 563-582.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American educational research journal, 31*, 104-137.

Haarmann, H. J., Davelaar, E. J., & Usher, M. (2003). Individual differences in semantic short-term memory capacity and reading comprehension. *Journal of Memory and Language*, 48(2), 320-345.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning & Verbal Behavior. Vol., 10*, 562-567.

Kintsch, E. (2005). Comprehension theory as a guide for the design of thoughtful questions. *Topics in Language Disorders, 25*, 51-64.

Mannes, S. M., & Kintsch, W. (1987). Knowledge organization and text organization. *Cognition and Instruction, 4*, 91-115.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*, 494-513.

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*, 200-206.

Oakhill, J. (1984). Inferential and memory skills in children's comprehension of stories. *British Journal of Educational Psychology, 54*, 31-39.

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181-210.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255.

Siegal, L. S., & Ryan, E. B. (1989). The Development of Working Memory in Normally Achieving and Subtypes of Learning Disabled Children. *Child Development*, 60(4), 973-980.

Snow, C. E. (2002). *Reading for understanding: Toward an R & D program in reading comprehension*. Arlington, VA: RAND.

Summers, C., & Conant, T. (1992). *Houston museum of science: Horizons plus science stories level 3*. Morristown, NJ: Silver Burdett & Ginn.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning & Verbal Behavior, 6*, 175-184.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3*, 240-245.

Wilkinson, G. S. (1989). Wide range achievement test--revised. *Newmark, Charles S, 2*.

Wise, B., Cole, R., van Vuuren, S., Schwartz, S., Snyder, L., Ngampatipatpong, N., et al. (2007). Learning to read with a virtual tutor. In C. K. Kinzer & L. Verhoeven (Eds.), *Interactive literacy education: Facilitating literacy environments through technology*. NY: Erlbaum, Taylor & Francis Group.

Wise, B., Van Vuuren, S., Cole, R., & Kintsch, E. (2003). *Building thoughtful multiple choice questions*. In-house documents: Univ of Colorado, CSLR.

Wise, B., Van Vuuren, S., Struemph, T., & Richey, L. (2006). *How we build thoughtful questions for interactive books*. In-house documents: Univ of Colorado, CLEAR.

Woodcock, R. W. (1987). *Woodcock reading mastery tests-revised for form g*. Circle Pines, NM: American Guidance Service.