EXAMINING EYE MOVEMENTS OF ELEMENTARY STUDENTS DURING READING COMPREHENSION ASSESSMENT

by

ANDREA MARIE ZAWOYSKI

(Under the Direction of Scott P. Ardoin)

ABSTRACT

High-stakes reading comprehension assessment is commonplace in schools, given the increased emphasis on accountability for educational outcomes (Barksdale-Ladd & Thomas, 2000). Yet, not all assessment practices have empirical support. This two-study dissertation examined areas related to reading comprehension assessment: test-taking strategies and question format. Both studies employed eye tracking technology in order to examine participants' underlying reading comprehension behavior.

Due to conflicting findings regarding the effectiveness of test-taking strategies during reading comprehension assessment (e.g., Wiesendanger, Birlem, & Wollenberg, 1982), the purpose of Chapter 2 was to investigate elementary students' reading behavior during a reading comprehension assessment under typical conditions and when required to read the passage first (PF) or read the questions first (QF). Participants were 84 third- and fourth-grade students who first completed a control condition and then completed PF and QF conditions in randomized order. Eye movement data revealed that the QF strategy was generally less efficient than the PF strategy.

The purpose of Chapter 3 was to investigate the impact of anticipated question format (i.e., Multiple-Choice; MC or Short Answer; SA) on elementary students' eye movements during reading. This study addressed concerns that MC questions may alter what reading assessments measure (Martinez, 1999; Rupp, Ferne, & Choi, 2006). Additionally, this study examined the influence of SA questions on reading behavior, given that new reading assessments will include more SA questions (Polikoff, 2014). Participants included 87 third- and fourth-grade students who were randomly assigned to the MC or SA condition. Condition assignment dictated the type of questions participants expected to answer after reading a passage. Results indicated that participants in the SA condition engaged in different eye movements than participants in the MC condition.

Overall, this dissertation extends research in the fields of school psychology and cognitive psychology. Implications of the studies reported upon in Chapters 2 and 3 may inform future studies regarding the validity of reading comprehension assessment practices. Implications of findings from the two studies for classroom practices were also discussed.

INDEX WORDS: Reading comprehension, Elementary students, Eye movements, Test-taking strategies, Multiple-choice questions, Short-answer questions

EXAMINING EYE MOVEMENTS OF ELEMENTARY STUDENTS DURING READING COMPREHENSION ASSESSMENT

by

ANDREA MARIE ZAWOYSKI

BS, University of Florida, 2011

MA, University of Georgia, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2017

EXAMINING EYE MOVEMENTS OF ELEMENTARY STUDENTS DURING READING COMPREHENSION ASSESSMENT

by

ANDREA MARIE ZAWOYSKI

Major Professor:	Scott P. Ardoin
Committee:	Katherine S. Binder Stacey Neuharth-Pritchett Amy L. Reschly

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2017

ACKNOWLEDGEMENTS

My most important acknowledgement goes to my advisor, Dr. Scott Ardoin. In the past six years, Scott has provided unwavering support that has guided me through countless moments of doubt. I will forever be grateful for his availability, kindness, and expertise. I know that at the very least, I will "always have something to be working on" until this dissertation is accepted for publication, but I hope that my scholarly endeavors with Scott will continue far beyond that point. I could not have asked for a better mentor.

I would also like to acknowledge my dissertation committee members, Dr. Katherine Binder, Dr. Stacey Neuharth-Pritchett, and Dr. Amy Reschly, for their guidance and support throughout this project. I have greatly benefited from Dr. Binder's rich knowledge of eye movement research and cognitive psychology throughout all the projects we have worked on together. Additionally, I value Dr. Neuharth-Pritchett and Dr. Reschly's insightful feedback during my prospectus defense, which shaped my dissertation into its current state.

Next, I would like to acknowledge the support of the ArdoinLab members I worked closely with during my time at UGA: Stacy-Ann January, Jeff Hine, Laura Rogers, Tori Foster, Christina Simmons, Cameron Oddone, Patrick Morin, Lily Wagner, Katie Bangs, Kristen McConkey, and Josh Mellott. In accordance with our Bylaws, but certainly not only because of the Bylaws, ArdoinLab members have always helped and supported me. I will eternally admire the team's hard work and dedication to research. I would also like to give special recognition to Cameron, Kristen, and Josh, who assisted with a significant portion of my dissertation data collection.

In addition, I am sincerely grateful for the help of my fantastic research assistants: Shannon Schebell, Anna Butler, Erinn Whiteside, Madeline Brown, Sharon Udemba, Justin Li, and Adam Lowe. It was a pleasure to train Shannon, Anna, and Erinn, who learned the intricacies of eye tracking data cleaning and even learned to enjoy the process. I will always remember early morning data collection fondly, thanks to the energy and enthusiasm from Madeline, Sharon, Justin, and Adam, who were quick to learn new and challenging tasks.

Throughout my graduate career, I was fortunate to learn from the faculty members in the School Psychology program at UGA, as well as many of the faculty members in the Special Education program. I hope that our partnership continues to prosper! Furthermore, I am appreciative of the guidance from my numerous practicum supervisors and mentors from Gwinnett County Public Schools, Children's Healthcare of Atlanta, and of course, the Marcus Autism Center. Additionally, I must acknowledge the support of my current internship supervisors at the Munroe-Meyer Institute. Together, these placements have enriched my clinical and research skills to an immeasurable extent.

Looking back, I would not be where I am today without the support of my undergraduate mentors, Dr. Timothy Vollmer and Dr. Amanda Bosch, who somehow saw potential in an 18-year-old girl who was dazzled by the science of behavior. I will never forget when Dr. Vollmer pulled up Scott's picture in his office and told me to "go find him." I will always be amazed by seemingly minor, yet life-changing moments.

Also important to the success of this project was the assistance of the highly skilled tech-support employees at SR Research, who worked around the clock to answer my many questions and assist in the creation of my eye tracking programs. I suspect that at this point they have blocked my email address, or at least draw straws to see who must respond to me. Additionally,

I am thankful for the financial support provided by the Society for the Study of School Psychology and the University of Georgia Innovative and Interdisciplinary Research grant, which funded necessary materials and incentives for my dissertation research. Certainly, my dissertation could not have been completed without the time and effort of the my third- and fourth-grade research participants and the patience of their teachers and school principals. I truly believe that their participation will benefit other students in the future.

Finally, this dissertation could not have been completed without the constant, life-long support from my family, especially my parents and my brother. Additionally, I am grateful for the hard work and encouragement from my extended family, both living and deceased. I am also thrilled to have my reliable pup, Nemo at my side. (He is reliable in the sense that I could always count on him to put his face on my keyboard while I was trying to type.) Last but certainly not least, my fiancé, Nate, earns my most heartfelt gratitude for his continuous words of encouragement and his selfless support of my goals.

To everyone mentioned above - It's been a long journey and finally, we're here.

TABLE OF CONTENTS

LIST OF TABLES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Over 30 years ago, a report entitled *A Nation at Risk: The Imperative for Educational Reform* revealed the poor state of academic performance in graduates from American schools and the startling statistic that 23 million Americans were functionally illiterate (National Commission on Excellence in Education, 1983). Afterwards, numerous efforts were made to help all students gain proficiency in reading, such as improving the breadth and quality of reading research, increasing schools' accountability for student performance, and creating resources that identify evidence-based interventions (e.g., What Works Clearinghouse). Despite these commendable efforts, over half of fourth-grade students are reading at or below the basic reading level, and fourth-grade students' reading scores have not improved significantly since 2007 (National Center for Education Statistics, 2014).

Current reading assessments use outcome measures (i.e., accuracy on comprehension questions) as a proxy for comprehension. Yet, an accuracy score does not capture all the underlying reading behaviors that occur before a student selects his or her answer, such as how often the student reviewed relevant material or what kind of strategies the student engaged in when reading. Additionally, the extent to which students engage in comprehension on a passage may differ based on the task they expect to complete afterwards. Researchers studying reading comprehension must examine not only the products of reading comprehension, but also the processes of reading comprehension (Rapp, van den Broek, McMaster, Kendeou, & Espin, 2007). Eye tracking can provide a real-time examination of underlying reading behavior because it allows for direct recording of students' attention allocation during reading and provides insight

into comprehension processes (Just & Carpenter, 1980; Radach & Kennedy, 2013; Rayner, Chace, Slattery, & Ashby, 2006). By viewing students' eye movements during reading, researchers can examine the process of comprehension and gain a better understanding of why students attain the accuracy scores they do. Eye tracking broadens opportunities for research by providing a means to observe reading behavior that was previously considered unobservable.

Eye Movement Research

Modern eye movement research requires a participant to read stimuli on a computer screen as a camera records his or her eye movements. The camera obtains a record of where, how long, and how often participants make *fixations* on words (i.e., pauses on the text that allow the reader to gain information). The record also provides information about the length and locations of participants' *saccades* (i.e., eye movements) and *regressions* (i.e., backward eye movements that are thought to represent a reader's attempt to review previously read information or correct for overshooting eye movements). Later, these individual fixations and saccades are analyzed and information about individual dependent eye movement measures can be examined. Different eye movement measures are thought to reflect different aspects of reading, including lower-level processing (i.e., word recognition and decoding), and higher-level processing (comprehension) (Binder, 2003; Binder & Morris, 1995; Foster, Ardoin, & Binder, in press; Radach & Kennedy, 2013). Additionally, because eye movements are measured in real-time, their temporal occurrence can provide valuable information about reading behavior. For example, the duration of the first fixation a reader makes on a word is likely to reflect that reader's initial attempt at reading the word, whereas a fixation the reader makes back on that word after moving to a different word is likely to reflect the reader's attempt to gain more information from that word. Extant research suggests that the duration and frequency of

fixations can reflect the difficulty level of words or a text as a whole, because readers make longer and more frequent fixations when text is more challenging (Rayner et al., 2006).

Until recently, eye movement research in reading was conducted almost exclusively with skilled adult participants in a laboratory at a university, and stimuli often included single sentences. With advancements in technology improving portability and comfort of the eye tracking system, eye movement research in schools with child participants is now possible (Rayner, Ardoin, & Binder, 2013). Within the last decade, eye movement research has demonstrated its value in contributing to the understanding of underlying reading processes in children. Studies investigated reading interventions (e.g., Ardoin, Binder, Foster, & Zawoyski, 2016; Ardoin, Binder, Zawoyski, & Foster, 2017; Ardoin, Binder, Zawoyski, Foster, & Blevins, 2013), examined instructional procedures such as Repeated Readings (e.g., Foster, Ardoin, & Binder, 2013; Zawoyski, Ardoin, & Binder, 2015), compared behavior of readers at different skill levels (e.g., Valle, Binder, Walsh, Nemier, & Bangs, 2013; Zawoyski et al., 2015), analyzed phonemic awareness (e.g., Ashby, Dix, Bontrager, Dey, & Archer, 2013), contrasted eye movements during oral and silent reading (e.g., Vorstius, Radach, & Lonigan, 2014), and investigated psychometric qualities between eye movements and reading outcome measures (Foster et al., in press). Eye movement research on comprehension with school-aged participants is also increasing, as evidenced by studies examining eye movements during comprehension monitoring (e.g., Vorstius, Radach, Mayer, & Lonigan, 2013) as well as full-length passage comprehension (e.g., van der Schoot, Vasbinder, Horsley, & van Lieshout, 2008), and reading behavior during a question-answering task in seventh-grade students (i.e., Solheim & Uppstad, 2011) and second-grade students (i.e., Ardoin, Zawoyski, Wagner, Bangs, & Binder, 2015). A common conclusion amongst all these studies is that eye movement research in schools should

be continued because it allows researchers to gain rich information about reading behavior occurring at the millisecond level. Another consensus is that the eye tracker is not a diagnostic tool and is not recommended as a replacement for established procedures used to identify students in need of intensive intervention or special education. Rather, eye movement research provides a means for understanding how reading comprehension occurs and permits analysis of individual differences in comprehension processes, with the goal of improving reading assessment and intervention (Rayner et al., 2013; Rayner et al., 2006). In light of the aforementioned considerations, the following two-study dissertation aimed to improve understanding of reading comprehension in elementary students by using eye tracking technology to observe test-taking strategies and the impact of anticipated question format on reading behavior.

Chapter 2: Test-Taking Strategies

Chapter 2 described the rationale, methodology, and analyses for Study 1. The chapter first reviewed extant research on test-taking strategies observed in college-aged students (i.e., Cerdán, Vidal-Abarca, Martínez, Gilabert, & Gil, 2009; Farr, Pritchard, & Smitten, 1990; Lewandowski, Gathje, Lovett, & Gordon, 2013) and how forced strategy usage influences reading behavior and accuracy on comprehension questions (i.e., Daneman & Hannon, 2001). Specifically, a test-taking strategy encouraging students to read the questions prior to reading the text (i.e., the questions-first strategy (QF)) was compared to a traditional passage-first strategy (PF), in which students read the passage prior to reading the text. Findings from the few studies examining PF and QF strategy usage with elementary students were presented (e.g., Bishop, 2001; Bishop & Frisbie, 1999; Wiesendanger, Birlem, & Wollenberg, 1982; Wiesendanger & Wollenberg, 1978). Also discussed were eye movement studies examining reading

comprehension in children (e.g., van der Schoot et al., 2008), question-answering strategies with older students (e.g., Solheim & Uppstad, 2011), and test-taking strategies in second-grade students (e.g., Ardoin et al., 2015). Then, Chapter 2 provided methodology detailing the procedures for assessment and comparison of third- and fourth-grade students' eye movements and accuracy under typical reading conditions and when engaging in the PF and QF strategy. Analyses included mixed ANOVAs and *t*-tests evaluating participants' eye movements during reading and on areas of the text containing and not containing answers to literal questions. Additionally, mixed ANOVAs were conducted to examine differences in participants' accuracy on comprehension questions and the overall efficiency of the PF and QF strategies. The discussion summarized findings from analyses on eye movements, accuracy, and efficiency, detailed the implications of results, and presented avenues for future eye movement research as well as recommendations for teachers who encourage students to use test-taking strategies.

Chapter 3: Question Format

Chapter 3 provided a rationale for Study 2, which aimed to examine the impact of question format on elementary students' reading behavior. Participants expected to answer multiple-choice (MC) or short-answer (SA) questions after reading a corresponding passage. Their eye movements on the passage were monitored during reading. MC and SA questions were compared based on concerns that MC questions may change the construct of reading comprehension (Martinez, 1999; Rupp, Ferne, & Choi, 2006). Additionally, newer reading assessments, such as the Common Core State Standards Assessments, may include more SA items (Polikoff, 2014). Yet, there is a significant lack of research regarding how question format impacts reading comprehension performance. MC and SA formats were reviewed and compared to examine the benefits and drawbacks of each type. Studies examining similarities and

differences in results obtained from questions of differing formats were also presented. Chapter 3 also included a discussion of passage dependency (i.e., the extent to which readers can answer questions without reading the corresponding passage). Then, Chapter 3 reviewed the existing research on eye movements during reading comprehension (e.g., Ardoin et al., 2015; Solheim & Uppstad, 2011; Vorstius et al., 2013) and question format, which amounts to one study conducted by Feng et al. (2012), during which undergraduate students' expectations for question type were manipulated. Chapter 3 extended methodology similar to that utilized by Feng et al. to examine eye movements of elementary-aged students. Procedures for examining elementary students' eye movements during reading comprehension were presented, as well as methodology for manipulating their expectation of question format based on group assignment. Results from *t*-tests examining between-groups differences in participants' eye movements during reading as well as their comprehension question accuracy were presented. Additionally, Chapter 3 included findings from correlational analyses and multiple linear regressions examining the relations between eye movement measures, reading measures, and outcome measures. The discussion summarized results and provided recommendations for test developers as well as future eye movement research on reading comprehension in elementary students.

Purpose

The following two studies aimed to advance the fields of cognitive psychology and school psychology by demonstrating the benefits of conducting eye movement research to better understand the reading comprehension and test-taking behavior of elementary students. The studies provided a foundation for research in unexplored areas and extended upon previous findings with skilled readers. Additionally, both studies accounted for the impact of working

memory on reading comprehension, based on evidence suggesting that successful comprehension requires intact working memory abilities (Cain, 2006).

The purpose of the study presented in Chapter 2 was to examine the natural reading behavior of third- and fourth-grade students during a reading comprehension assessment in a typically presented format (i.e., passage presented next to corresponding MC questions) using eye tracking technology. This permitted comparison of natural test-taking strategies in third- and fourth-grade students, who are beginning to learn test-taking strategies, with those exhibited by second-grade students (Ardoin et al., 2015) and adults (Farr et al., 1990; Lewandowski et al., 2013). Additionally, Chapter 2 involved manipulation of question and passage presentation to investigate the utility of the PF and QF strategies. Findings (a) revealed differences in elementary students' reading strategy usage and eye movements during reading, (b) detailed how accuracy and efficiency was impacted by the PF and QF strategies, and (c) identified which strategy was most likely to evoke eye movements indicative of comprehension.

The study presented in Chapter 3 aimed to investigate the impact of question format on eye movements during reading for third- and fourth-grade students. Participants expected to answer either MC questions or SA questions based on their assignment experimental condition. Eye movements on the passage were analyzed in order to assess for differences in reading behavior reflective of comprehension across the text and on meaningful areas of the text. Differences in accuracy on comprehension questions was also examined. Results provided information regarding how the expectation of question format impacts elementary students' eye movements during reading.

References

Ardoin, S. P., Binder, K. S., Foster, T. E., & Zawoyski, A. M. (2016). A randomized control design study examining the effects of repeated readings on reading achievement and reading behavior. *Journal of School Psychology, 59,* 13-38. doi:10.1016/j.jsp.2016.09.002

Ardoin, S. P., Binder, K. S., Zawoyski, A. M., & Foster, T. E. (2017). *Examining the maintenance and generalization effects of repeated practice: A comparison of three interventions.* Manuscript submitted for publication.

Ardoin, S. P., Binder, K. S., Zawoyski, A. M., Foster, T. E., & Blevins, L. A. (2013). Using eye-tracking procedures to evaluate generalization effects: Practicing target words during repeated reading within versus across texts. *School Psychology Review, 42*, 477-495.

Ardoin, S. P., Zawoyski, A. M., Wagner, L., Bangs, K., & Binder, K. S. (2015). *Measuring test-taking behavior: Different behaviors but similar outcomes.* Manuscript submitted for publication.

Ashby, J., Dix, H., Bontrager, M., Dey, R., & Archer, A. (2013). Phonemic awareness contributes to text reading fluency: Evidence from eye movements. *School Psychology Review, 42,* 157-170.

Binder, K. S. (2003). Sentential and discourse topic effects on lexical ambiguity processing: An eye movement examination. *Memory and Cognition, 31,* 690-702. doi:10.3758/BF03196108

Binder, K.S., & Morris, R. K. (1995). Eye movements and lexical ambiguity resolution: Effects of prior encounter and discourse topic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1186-1196. doi:10.1037/0278-7393.21.5.1186

Bishop, N. S. (2001, April). *The validity of reading comprehension test scores: Evidence of generalizability across difference test administration conditions.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, Washington.

Bishop, N. S. & Frisbie, D. A. (1999, April). *The effects of different test-taking conditions on reading comprehension test performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Cain, K. (2006). Children's reading comprehension: The role of working memory in normal and impaired development. In S. J. Pickering (Ed.), *Working memory and education* (pp. 61-91). Cambridge, MA: Academic Press.

Cerdán, R., Vidal-Abarca, E., Martínez, T., Gilabert, R., & Gil, L. (2009). Impact of question answering tasks on search processes and reading comprehension. *Learning and Instruction, 19,* 13-27. doi:10.1016/j.learninstruc.2007.12.003

Daneman, M. & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology, 130,* 208-223. doi:10.1037/0096-3445.130.2.208

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27,* 209-226. doi:10.1111/j.1745-3984.1990.tb00744.x

Feng, G., Gorin, J., Sabatini, J., O'Reilly, T., Wall, C., & Bruce, K. (2012, July). Reading for understanding: How comprehension facilitates answering questions, and what questions enhance understanding. In G. Feng (Chair), *Higher order literacy skills.* Symposium

conducted at the meeting of the Society for the Scientific Study of Reading, Montreal, Canada.

Foster, T. E., Ardoin, S. P., & Binder, K. S. (2013). Underlying changes in repeated reading: An eye movement study. *School Psychology Review, 42,* 140-156.

Foster, T. E., Ardoin, S. P., & Binder, K. S. (in press). Reliability and validity of eye movement measures of children's reading. *Reading Research Quarterly.*

Just, M. A. & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 4,* 329-354. doi:10.1037/0033-295X.87.4.329

Lewandowski, L., Gathje, R. A., Lovett, B. J., & Gordon, M. (2013). Test-taking skills in college students with and without ADHD. *Journal of Psychoeducational Assessment, 31,* 41-52. doi:10.1177/0734282912446304

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist, 34,* 207–218. doi:10.1207/s15326985ep3404_2

National Center for Education Statistics (2014). The Nation's Report Card: A first look: 2013 mathematics and reading (NCES Publication No. 2014-451). Washington, DC: U.S. Government Printing Office.

National Commission on Excellence in Education (1983). A nation at risk: The imperative for educational reform. *The Elementary School Journal, 84,* 112-130.

Polikoff, M. S. (2014). Common core state standards assessments: Challenges and opportunities. Retrieved from: https://cdn.americanprogress.org/wpcontent/uploads/2014/04/CCCAssessments-report.pdf

Radach, R., & Kennedy, A. (2013). Eye movements in reading: Some theoretical context. *The Quarterly Journal of Experimental Psychology*, *66*, 429-452. doi:10.1080/17470218.2012.750676

Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P. & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11,* 289–312. doi:10.1080/10888430701530417

Rayner, K., Ardoin, S. P., & Binder, K. S. (2013). Children's eye movements in reading: A commentary. *School Psychology Review*, *42,* 223-233.

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading, 10,* 241-255. doi:10.1207/s1532799xssr1003_3

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shape the construct: A cognitive processing perspective. *Language Testing, 23*, 441-474. doi:10.1191/0265532206lt337oa

Solheim, O. J. & Uppstad, P. H. (2011). Eye-tracking as a tool in process-oriented reading test validation. *International Electronic Journal of Elementary Education, 4,* 153-168.

Valle, A., Binder, K. S., Walsh, C. B., Nemier, C., & Bangs, K. E. (2013). Eye movements, prosody, and word frequency among average-and high-skilled second-grade readers. *School Psychology Review*, *42,* 171-190.

van der Schoot, M., Vasbinder, A. L., Horsley, T. M., & van Lieshout, E. C. D. M. (2008). The role of two reading strategies in text comprehension: An eye fixation study in primary

school children. *Journal of Research in Reading, 31,* 203–223. doi:10.1111/j.14679817.2007.00354.x

Vorstius, C., Radach, R., & Lonigan, C. J. (2014). Eye movements in developing readers: A comparison of silent and oral sentence reading. *Visual Cognition, 22,* 458-485. doi:10.1080/13506285.2014.881445

Vorstius, C., Radach, R., Mayer, M., & Lonigan, C. (2013). Monitoring local comprehension monitoring in sentence reading. *School Psychology Review, 42,* 191–206.

Wiesendanger, K. D., Birlem, E. D., & Wollenberg, J. (1982). A summary of studies related to the effect of question placement on reading comprehension. *Reading Horizons, 23,* 15-21.

Wiesendanger, K. & Wollenberg, J. (1978). Prequestioning inhibits third graders' reading comprehension. *The Reading Teacher, 31,* 892-895.

Zawoyski, A. M., Ardoin, S. P., & Binder, K. S. (2015). Using eye tracking to observe differential effects of repeated readings for second-grade students as a function of achievement level. *Reading Research Quarterly, 50,* 171-184. doi:10.1002/rrq.91

CHAPTER 2

THE IMPACT OF TEST-TAKING STRATEGIES ON EYE MOVEMENTS OF ELEMENTARY STUDENTS DURING A READING COMPREHENSION ASSESSMENT[1]

[1] Zawoyski, A. M. and S. P. Ardoin. To be submitted to *School Psychology Review*.

Abstract

Teachers often encourage students to use test-taking strategies during reading comprehension assessments, but these strategies are often not evidence-based. One common strategy dictates that students should read the questions prior to reading the passage. Results from studies examining the effectiveness of a questions-first (QF) strategy in comparison to a passage-first (PF) strategy have been mixed (Wiesendanger, Birlem, & Wollenberg, 1982). The current study employed eye tracking technology to record 84 third- and fourth-grade participants' eye movements as they completed comprehension assessments. Participants engaged in different test-taking strategies under typical conditions. Additionally, PF and QF strategies significantly differed in their effects on participants' reading behavior as well as in their accuracy and testing efficiency. Implications for test-taking strategy effectiveness and future eye movement studies are discussed.

INDEX WORDS: Reading, Reading comprehension, Elementary students, Eye movements, Reading strategies, Test-taking strategies

Introduction

High-stakes testing greatly impacts the lives of students, teachers, and the entire school community. Teachers, in particular, are under significant pressure to help their students achieve exceptional test scores (Barksdale-Ladd & Thomas, 2000). Given the emphasis on testing outcomes, teachers may try to help students improve their performance by encouraging them to use test-taking strategies.

With respect to reading comprehension assessments, one test-taking strategy students are often taught is to read comprehension questions prior to reading the passage (i.e., use a QF strategy). Empirical support for the QF strategy is limited (Bishop & Frisbie, 1999; Wiesendanger & Wollenberg, 1978), yet this strategy remains popular amongst teachers and test-preparation materials across grade levels. For example, Rupp, Ferne, and Choi (2006) found that some test-preparation materials for the Graduate Record Examination and the Test of English as a Foreign Language recommended the questions-first (QF) strategy. Additionally, a simple search of the internet reveals numerous websites suggesting that parents and teachers instruct students to utilize a QF strategy when taking reading comprehension assessments (See Appendix A for a list of several test-taking recommendations). Most likely, the QF strategy arose from ideas that teachers should prime students' schemas before they read (Bishop & Frisbie, 1999), thus helping them retain more information about the text because they can connect this information with previously acquired knowledge. Opposing views suggest that by reading questions first, students may search the passage for answers, rather than read the passage for its content. Evidence for this latter assumption comes from research with adults (e.g., Cerdán, Vidal-Abarca, Martínez, Gilabert, & Gil, 2009). Additionally, reading the questions first typically takes more time (Bishop, 2001; Bishop & Frisbie, 1999; Cerdán et al., 2009; Daneman & Hannon, 2001), which must be allocated wisely on timed standardized reading assessments

(Kwabi, Xu, Binder, Nemier, & Ardoin, 2015). Furthermore, the QF strategy might negatively impact performance because this strategy contradicts test directions indicating that students should read the passage before viewing the questions; students who are trying to comply with test directions may become uncertain about how to complete the assessment (Perlman, 2003).

In order to understand how test-taking strategies can impact reading comprehension, it is important to understand the process of comprehension. Comprehension is often considered the goal of reading because successful comprehension implies that the reader has understood the meaning of the text. Reading comprehension requires integration of skills such as decoding, fluency, and vocabulary (Rayner, Chace, Slattery, & Ashby, 2006). In addition to foundational reading skills, another critical component of reading is working memory. Working memory refers to a cognitive system involved in information processing, manipulation, and storage (Baddeley, 2003; Baddeley & Hitch, 1974). According to Cain (2006), without strong working memory abilities, readers struggle to comprehend text, even when prerequisite reading skills are intact (e.g., word reading). Additionally, working memory independently predicts reading comprehension in children (Andreassen & Bråten, 2010; Cain, Oakhill, & Bryant, 2004). Reading comprehension skills and working memory abilities are closely related because working memory capacity allows a reader to store information and integrate it with other areas of a text.

Applied Reading Research on Test-Taking Strategies

Studies examining natural test-taking strategies in skilled readers. Most reading comprehension assessments require students to answer multiple-choice (MC) questions associated with a passage that they have access to when answering the questions. The dependent measure for such assessments is students' response accuracy. Yet, accuracy scores do not reflect

how students arrive at their responses. Students engage in varying processes during test-taking that are not illuminated by their final choice (Gorin, 2006).

Farr, Pritchard, and Smitten (1990) examined reading behavior of undergraduate students during a reading test as they participated in interviews and provided information about their test-taking strategies. By analyzing students' reports in the interviews, Farr et al. found that the undergraduate participants engaged in one of four reading strategies. A majority of participants read the passage first (i.e., used a PF strategy) and then alternated between reading the questions and the passage. That is, they immediately returned to the passage to verify their answer after reading each question. The second most popular strategy was a QF strategy, whereby participants read the questions, read the entire passage, and then reread the questions and returned to the passage to find their answers. In the third most popular strategy, students read some of the passage first and then alternated between reading the questions and rereading the text. In the least popular strategy, a single participant immediately skipped to read a question and then searched the passage for a correct answer, continuing to alternate between reading questions and searching for answers for the remainder of the assessment. Interestingly, some participants did not maintain their initially selected strategy as they continued to read. According to Farr et al., these participants reported that when the passage topic was difficult or boring, or when they became fatigued, they shifted away from reading the entire passage to a strategy involving reading the questions first or reading only part of the passage before answering the questions. Of note, participants who started with a QF strategy maintained that strategy throughout all of the readings. All participants reread portions of the passage when answering questions. Similarly, findings from Lewandowski, Gathje, Lovett, and Gordon (2013) suggested that the PF strategy was the most popular strategy in adult readers, followed closely in

popularity by a searching strategy in which participants read each question first and then searched the passage for an answer.

Experimental manipulation of test-taking strategies with skilled readers. In an extension of Farr et al. (1990), Daneman and Hannon (2001) instructed graduate students taking the SAT to use the four reading strategies outlined by Farr et al. so that they practiced each strategy on two passages. Response accuracy and total testing time were measured. Findings indicated that students took the longest time to answer the questions when engaging in the strategy that required them to read the questions and then read the entire passage. However, there were no significant differences in response accuracy based upon strategy employed. Interestingly, reading span correlated highest with participants' performance in the strategy involving questions first/reading the whole passage. That is, Daneman and Hannon found that students with stronger working memory abilities outperformed students with lower working memory abilities when using the strategy that requiring the most reading.

Cerdán et al. (2009) also conducted a study examining the effects of test-taking strategies on skilled readers' test performance. Undergraduate students were asked to either (a) read a passage first and then respond to questions, or (b) read the questions first and then search the passage for answers. Online reading behavior was monitored by Read&Answer software (Martínez & Sellés, 2001), which tracked the time that students engaged with experimental material. Students provided data on their reading and question-answering times by clicking locations in the text to indicate when they were finished reading or answering questions. Results showed that students who read the passage first had better delayed recall and comprehension of the text than students who did not read the passage first. Cerdán et al. noted that, as compared to students who did not read the passage first, students who read the passage first were given the

opportunity to read it in its entirety at least one time. Cerdán et al. hypothesized that reading the passage first allowed students to develop a more complete understanding of the material than students who did not read the passage first. This hypothesis was corroborated by Cerdán et al.'s finding that students who were required to read the passage first allocated more time on the passage than students who read the questions first and searched the passage for answers.

The pattern of results from Cerdán et al. (2009) and Daneman and Hannon (2001) pertaining to the drawbacks of the QF strategy was also observed in an older study conducted with high-school students. Rothkopf and Bisbicos (1967) presented different types of comprehension questions shortly before or shortly after high-school students read a corresponding text. Results indicated that across question types, when students read questions shortly before reading the corresponding text, performance on reading comprehension questions was not improved. Rather, findings suggested that the most beneficial strategy was to present questions shortly after students read the corresponding text.

Test-taking strategies of elementary students. Early studies conducted with elementary students were inconsistent in their findings about whether the QF strategy (described as "prequestioning") improved students' reading comprehension (Wiesendanger, Birlem, & Wollenberg, 1982). Generally, arguments in favor of prequestioning pose that they help the reader find answers within the text more quickly, whereas arguments opposed to prequestioning suggest that students selectively attend to the information relevant to the questions and fail to engage in comprehension on the passage as a whole (Wiesendanger & Wollenberg, 1978). In a review of research from the early to mid-1900's on prequestioning, Wiesendanger et al. examined studies manipulating question placement in reading comprehension assessment. Among the reviewed articles, 12 studies denounced prequestioning and 10 studies supported it.

Wiesendanger et al. discussed possible reasons for disagreement, such as features of the questions, how questions were presented (i.e., orally or written), and the influence of demographic factors. Wiesendanger et al. also noted the importance of a student's reading purpose and how it may impact his or her comprehension of the text.

Although Wiesendanger et al. (1982) introduced an important debate in educational psychology, the review presented with several notable limitations with respect to understanding the effectiveness of prequestioning for elementary students. For example, a majority of the cited sources in Wiesendanger et al. were unpublished dissertation abstracts. Additionally, the review did not describe participants, so it is uncertain whether participants in the cited studies were all elementary students, or if older students and adults were included. Also, some of the studies described research related to improving students' social studies and science text reading performance, making it difficult to compare findings with reading behavior of elementary students who are learning how to engage in reading for understanding.

In one of the only studies manipulating test-taking strategies with elementary students, Bishop and Frisbie (1999) assigned third-, fifth-, and seventh- grade students, by classroom to conditions requiring them to engage in either a QF strategy or a PF strategy. Question type (i.e., literal, inferential, generalization) was also examined. Results indicated that generally, the PF strategy resulted in improved accuracy and more items completed than the QF strategy for all grades and question types. When Bishop and Frisbie standardized completion time across groups, the only grade that benefited from the QF strategy was fifth grade. Interestingly, fifth-grade participants had previously learned to use the QF strategy.

Bishop (2001) subsequently expanded upon the research conducted by Bishop and Frisbie (1999) and assigned third-, fifth-, and seventh-grade students to one of eight experimental

conditions that manipulated whether (a) the questions were presented with or without corresponding text, (b) the questions were presented first or the passage was presented first, and (c) the participants did or did not have an opportunity to practice the procedures associated with their assigned experimental conditions. Participants were administered the *Iowa Tests of Basic Skills* (Hoover, Dunbar, & Frisbie, 2007). Findings revealed that under standard timing conditions, students who read the questions first attained lower scores on the test, as compared to students who read the passage first.

Together, Bishop and Frisbie (1999) and Bishop (2001) denounced the effectiveness of the QF strategy. Both studies highlighted concerns that modifying test conditions so that students engage in unnatural reading behavior during tests reduces the validity and generalizability of scores. Bishop suggested that a within-subjects design may be better able to answer research questions pertaining to the impact of different testing conditions on reading assessment accuracy.

Limitations

There are significant limitations to existing research on test-taking strategies in reading comprehension assessment for elementary students. First, conclusions are largely drawn from studies with adults (e.g., Daneman & Hannon, 2001; Cerdán et al., 2009; Farr et al., 1990; Lewdanowski et al., 2013) even though adults and children have drastically different reading skill levels and generally do not have the same reading goals. Additionally, existing studies rely solely on outcome measures as indicators of comprehension, or use qualitative methods such as cognitive interviews (e.g., Farr et al., 1990), which can disrupt natural reading behavior (Cordón & Day, 1996).

Research on test-taking strategies conducted with elementary students is significantly outdated (e.g., studies mentioned in Wiesendanger et al., 1982). Reading comprehension instruction and testing technology have evolved over time, and students' responses to test-taking strategies may have evolved as well. Furthermore, research on the QF strategy with elementary students utilized outcome measures involving students' accuracy on questions (e.g., Bishop, 2001; Bishop & Frisbie, 1999). Although response accuracy is important, those data do not allow researchers to examine the process by which students arrived at their responses. One way to examine the reading process is through eye tracking procedures, which permit real-time observation of students' eye movements during reading.

Eye Movement Research

Eye movement research in reading is based on the premise that eye movements reflect attention allocation and underlying comprehension processes (Radach & Kennedy, 2013; Rayner et al., 2006). Although the majority of studies in eye movement research on reading comprehension involve skilled adult participants, researchers are increasingly involving elementary students as participants. For example, Vorstius, Radach, Mayer, and Lonigan (2013) conducted a study on comprehension monitoring with elementary students as they read single sentences. Interestingly, results indicated that students made longer fixations on information that was relevant to their purpose for reading. They were also more likely to reread challenging sentences and their accuracy improved when they reread relevant areas of the sentences.

In one of the earliest studies to examine school-aged children's reading behavior on a passage, van der Schoot, Vasbinder, Horsley, and van Lieshout (2008) used eye tracking to examine the reading behavior of fifth- and sixth-grade students. This study examined whether students' reading behavior differed on important and unimportant words and if students reread

previous areas in the text to facilitate interpretation of a word in a later area. Findings revealed that students use differing reading strategies when taking a reading test. Additionally, results suggested that students with higher reading comprehension scores were more efficient in their reading. Specifically, students who had higher comprehension scores were more likely to spend time reading important information as opposed to unimportant information, whereas students with lower comprehension scores were more likely to spend equal amounts of time on important and unimportant information. Additionally, students with higher comprehension scores were able to interpret ambiguities in the text more quickly than students with lower comprehension scores.

In an examination of reading behavior in older students, Solheim and Uppstad (2011) recorded eye movements of 18 seventh-graders as they completed a reading comprehension assessment. Eye movements were recorded on the passage only; participants used a pencil to answer questions presented on a sheet of paper. Participants were permitted to return to the passage when responding to questions and a head-mounted eye tracker continued to record their eye movements on the passage. Solheim and Uppstad examined students' reading strategies by dividing students into groups based on their response accuracy to a particular item and whether they read parts of the text containing the correct response when answering the question. Findings suggested that students who answered the question correctly without returning to the relevant area of the text when responding read that section of the text thoroughly in their initial reading. That is, when these students initially read the passage, they gathered enough information to answer the question, so they did not need to return to the passage and verify their response. Solheim and Uppstad also suggested that differences in prior knowledge may have accounted for differences in their performance. In comparison, readers who did not spend as

much time reading the relevant section of the text were more likely to answer the question incorrectly. Interestingly, some students who returned to relevant areas of the text when answering the question still answered incorrectly. Despite a small sample size, this study demonstrates the utility of eye movement research. Eye-tracking procedures allowed the researchers to observe eye movements of students as they were responding to questions and learn more about the reading behavior of those students who did not answer questions correctly.

Ardoin, Zawoyski, Wagner, Bangs, and Binder (2015) conducted the only study to date examining elementary students' eye movements as they read a passage and answered comprehension questions. Participants were randomly assigned to either a *with-text group* or *without-text group.* Participants in the with-text group viewed the passage and its corresponding multiple-choice (MC) questions together, whereas participants in the without-text group viewed the passage first in isolation, followed by the MC questions in isolation. Findings suggested that accuracy on the MC questions was not impacted by access to text. Accuracy was also not impacted by students' strategy usage, which was coded by examining reading behavior of students in the with-text condition. A majority (i.e., 56.47%) of participants in the with-text condition read the passage first followed by the questions and did not reread any portions of the passage when responding to questions. The QF strategy (i.e., reading a portion of the questions before reading the text) was used by 24% of participants. Interestingly, this percentage is strikingly similar to the percentage of adults in Farr et al. (1990) who used the QF strategy (i.e., 27%).

Of note, only 27% of participants in Ardoin et al. (2015) referenced the text as they were responding to questions. Thus, results suggested that a majority of students did not check the passage to verify their responses to the comprehension questions. According to Perlman (2003),

one possible explanation for this finding is that younger students may not know they are permitted to review the passage when answering questions, particularly if they have not received training in test-taking strategies.

Findings from Ardoin et al. (2015) also have interesting implications for the relationship between eye movements and reading skills, as eye movements differed between students who answered one question correctly and students who answered two questions correctly. Specifically, eye movement dependent measures thought to represent early or lower-level processing (e.g., word recognition and decoding) were longer for students who only answered one question correctly as compared to students who answered both questions correctly. In contrast, students who answered both questions correctly engaged in higher levels of eye movements thought to represent late or higher-level processing (e.g., comprehension). Ultimately, eye movement data from Ardoin et al. provide preliminary evidence that students engage in different reading behavior on comprehension tasks that may be a function of their reading skill level.

Given that Ardoin et al. (2015) conducted the first study to examine elementary students' reading behavior on comprehension questions and code their reading strategies based on their eye movements, more research on this topic is certainly necessary. Ardoin et al. also required students to respond to questions by using a hand signal, which may have impacted their natural reading behavior. Future research should allow students to respond using a more naturalistic method (e.g., by asking students to click their answer choices on the computer screen as they would for a typical computer-based reading assessment).

Summary and Purpose

Research with older students during passage reading (e.g., van der Schoot et al., 2008) and comprehension question responding (e.g., Solheim & Uppstad, 2011) suggests that there is much to be learned from eye movement research with children during reading assessment. Considering that most students have their first standardized reading assessment experiences in third grade, it is important to examine reading behavior of younger students (e.g., third and fourth graders) because it may differ from that of older students who are more proficient readers and test takers. Without a clear understanding of what students do when they engage in test-taking strategies, it is difficult to determine their utility. Some strategies may evoke passage-skimming or answer-searching behavior in students, which might allow students to answer questions without comprehending related text. Eye tracking can help researchers observe students' typical reading behavior and evaluate whether or not recommended test-taking strategies evoke eye movements indicative of reading comprehension. Therefore, the purpose of the current study is to use eye tracking to view students' test-taking behavior and strategy usage under natural circumstances (i.e., in a *control condition*), and to examine whether reading behavior is impacted when students read MC questions prior to (i.e., *QF condition*) or following (i.e., *PF condition*) the presentation of the associated passage.

Research Questions and Hypotheses

The following research questions were evaluated:

1. What reading strategies do participants engage in when taking a reading comprehension assessment with MC items?

Participants' reading behavior on the control passage were coded by whether they read the passage first or whether they read the questions first. Findings were expected to replicate

Ardoin et al. (2015), Farr et al. (1990), and Lewandowski et al. (2013), in that a majority of participants would utilize a PF strategy under natural reading conditions.

2. Do participants' eye movements during a reading comprehension assessment differ as a function of whether the questions (QF) or passage (PF) are read first?

It was expected that participants would engage in greater levels of reading behavior reflective of comprehension on the passage during the PF as compared to the QF condition, based on prior research suggesting that the PF strategy results in a deeper reading of the text than the QF condition (e.g., Cerdán et al., 2009).

3. Does experimental condition (i.e., PF or QF) impact participants' reading behavior on areas of the text containing answers to literal questions?

It was expected that the QF condition would result in more frequent and longer fixations on areas of the passage containing answers to literal questions as compared to the PF condition. Additionally, participants across both conditions were expected to engage in similar eye movements during reading on areas of the passage that did not contain responses to literal questions. These outcomes were anticipated because participants in the QF condition were expected to differentially attend to relevant regions of the text based on their prior reading of the questions, as suggested by studies discussed in Wiesendanger and Wollenberg (1978). Additionally, research suggests that students may engage in different reading behavior on important parts of the text (e.g., van der Schoot et al., 2008) or areas that are significant to their reading purpose (e.g., Vorstius et al., 2013).

4. What experimental condition (i.e., PF or QF) results in greater response accuracy and which condition is the most efficient?

Participants were expected to demonstrate greater response accuracy when reading and responding to questions in the PF condition as compared to QF condition. Additionally, participants were expected to have greater efficiency in answering questions during the PF condition, as evidenced by shorter time required to complete the PF task in comparison to the QF task. This expectation was based on results obtained from studies with children (e.g., Bishop, 2001; Bishop & Frisbie, 1999) and research with adults (e.g., Daneman & Hannon, 2001), suggesting that when participants read the questions before reading the passage, they required a longer time to complete the task without increased benefits in response accuracy as compared to participants who did not use this strategy.

Method

Participants and Setting

The sample size was determined by using power analyses for mixed-design ANOVAs and *t*-tests. A sample size of 56 was originally selected to permit detection of medium effects ($\eta_p^2 = .0588$, $\alpha = .05$, $\beta = .9$) in analyses conducted using mixed-design ANOVAs, according to Cohen's (1988) benchmark *f* value. Additionally, the sample size would allow for detection of large effects for analyses using *t*-tests (Cohen's $d = .8$, $\alpha = .05$, $\beta = .9$). More participants were included in the study that required because it was expected that a portion of eye movement data would not be suitable for analysis due to technical difficulties.

The sample included 43 third-grade students and 41 fourth-grade students from two public elementary schools in the southeastern United States. Exactly 50.00% of the sample was female and the average age was 9 years, 3 months (range = 8 years, 3 months to 10 years, 8 months). Participants identified as the following ethnicities: white (73.81%), Asian (10.71%), Black (3.57%), Hispanic (3.57%), and two or more races (3.57%). Age information was not

obtained for one participant and four participants did not indicate an ethnicity. Students receiving special education or English for Speakers of Other Languages services were excluded from participation. Participating schools were part of a rural-suburban school district comprised of 7,271 students with a predominantly white ethnicity (79.74%). Additionally, students in the district identified as Hispanic (7.33%), Asian (5.12%), Black (4.80%), two or more races (3.11%), American Indian (<1.00%), or Pacific Islander (<1.00%). In the participating school district, 20.52% of students were eligible for free or reduced-price meals during the 2015-2016 school year. State standardized testing results for the 2015-2016 school year indicated that 66.50% of third-grade students and 65.80% of fourth-grade students in the school district demonstrated proficient or distinguished achievement in English/Language Arts.

In addition to providing consent for their participation, participants' parents or guardians also provided consent for a review of records as well as publication and presentation of their children's de-identified data. The records review permitted the experimenter to examine participants' scores from the most recent administration of the Measures of Academic Progress (MAP; Northwest Evaluation Association, [NWEA], 2009) Reading test. Before participating, students provided written assent.

Measures

MAP Reading. MAP Reading is a group-administered computer adaptive test given to all students in the participating schools during the fall, winter, and spring of each school year. MAP Reading measures students' present levels of performance in specific areas of reading, including comprehension. Scores were presented as Rausch Unit (RIT) scores that span across grade levels; an important advantage to the MAP is that participants' reading levels may be compared across grades. Evaluation of the MAP's psychometric properties indicates that its

internal consistency reliability coefficients are excellent, ranging from .940 to .950 for elementary students. Test-retest reliability coefficients are above .800 for third and fourth-grade (NWEA, 2004). The concurrent validity coefficient between the reading portion of the *Stanford Achievement Test - Ninth Edition* and MAP Reading was .870 for third- and fourth-grade in 2001 (NWEA, 2004). The mean RIT score for participants in the study on the MAP reading assessment was 203.90, *SD* = 16.35. This score corresponds closely with the mid-year average score expected for fourth-grade students (*M* = 203.60, *SD* = 14.96; NWEA, 2015). Given that scores were obtained from MAP assessments administered to third- and fourth-grade students at the beginning of the school year, the mean score for the sample suggests that participants demonstrated above average reading achievement.

Eye tracking assessment passages and questions. Participants completed an eye tracking assessment, during which they read four passages from the Gates MacGinite Reading Tests (GMRT), Fourth Edition; Level 3 Form S (MacGinitie, MacGinitie, Maria, & Dreyer, 2000), which is recommended for students ages eight through nine. One passage was selected as a practice passage and one passage was selected as a control passage. The remaining two passages (i.e., Passage A and Passage B) were randomly presented in the experimental conditions (i.e., PF or QF). The practice passage consisted of three corresponding comprehension questions and the two experimental passages each included five corresponding questions. Lexile levels were similar across passages used in the study (control passage Lexile = 680, Passage A Lexile = 700, Passage B Lexile = 670). A measure of the proportion of students within the GMRT norming sample who answered each question correctly (i.e., Easiness Index) was obtained for the questions in the experimental passages. Fall and spring norms are provided due to timing of assessment administration for the current study. The mean Easiness Index for

questions presented in Passage A was .594 for students assessed in the fall and .628 for students assessed in the spring (range = .530-.770 for fall norms and .540-.800 for spring norms). Passage B originally included six questions, but one item was removed to increase consistency with Passage A. The item selected for removal was chosen based on its high Easiness Index relative to the other items (i.e., .780-.810 for fall and spring norms, respectively). For Passage B, the mean Easiness Index of questions selected for the study was .628 for students assessed in the fall and .714 for students assessed in the spring (range = .550-.710 for fall norms and .650-.800 for spring norms).

Texts were presented in black Times New Roman font against a white background. Passages were presented in approximately 15-point font with 1.5 line spacing. Questions were presented in approximately 14-point font. Font size could not be the same for the passage and questions due to spacing concerns. The practice passage contained 49 words with 21-56 characters per line. The control passage contained 110 words with 18-57 characters per line. For the experimental passages, Passage A contained 113 words with 38-55 characters per line and Passage B contained 102 words with 22-55 characters per line. When the passage and the questions were presented together, the passage appeared in the left half of the screen and the questions appeared in the right half of the screen.

Eye movement dependent measures. As participants silently read the passages, the eye-tracking camera recorded data on the locations and durations of their eye movements (i.e., fixations, saccades, and regressions). Researchers examining eye movements during reading define a *fixation* as a pause the eyes make on a word that allows the reader to extract information from the text (Rayner, 1998; Rayner et al., 2006). Between each fixation, a reader makes forward eye movements, known as *saccades,* and backward eye movements, known as

regressions. Different calculations of fixations and regressions yielded the measurements examined in the current study. Eye movements across the passages were calculated and averaged across words (i.e., global analyses) and on areas of the text containing answers to literal questions. For eye movements on the passage, dependent measures included total fixation time and average fixation count. Dependent measures on the questions included total fixation time and total fixation count. Notably, total fixation time measures on the questions were calculated as the sum of all fixation made on all questions and response choices. Eye movement measures used in the current study are thought to represent higher-level processing (Binder, 2003; Binder & Morris, 1995; Foster, Ardoin, & Binder, in press; Hyönä & Niemi, 1990). See Appendix B for definitions of eye movement measures.

Working memory assessment. Participants were administered the *Digit Span* subtest from the *Wechsler Intelligence Scale for Children - Fifth Edition* (WISC-V). This subtest measures participants' ability to recall, repeat, and manipulate sequences of numbers according to a rule. The Digit Span subtest is a technically sound measure of working memory in children and presents with a strong internal consistency coefficient of .89 to .90 for age ranges of participants in the current study (Wechsler, 2014). Across the sample, participants earned a score in the average range on the *Digit Span* subtest (M = 10.63, SD = 2.88). Working memory data were not obtained for two participants.

Oral reading fluency. After reading the passages and responding to questions while having their eye-movements recorded participants orally read all three passages presented on the eye tracker. Curriculum-based measurement (CBM) administration and scoring procedures were implemented with a slight deviation. Rather than allowing participants 3 s to hesitate on a word before examiners provided the word, participants were instead permitted 10 s to read a word

before they were asked to move on to the next word. The purpose of allowing participants 10 s to read each word was to recreate reading conditions on the eye tracker as closely as possible. Additionally, examiners did not provide words that participants had hesitated on for this same purpose. A median oral reading fluency score was calculated from participants' words read correctly in a minute (WRCM) scores for the three passages. The average median WRCM for the sample was 123.09 (S*D* = 31.55). Oral reading data were not obtained for two participants.

Apparatus

Eye movement data were collected with an SR Research EyeLink 1000 system. Its desktop-mounted camera has a resolution of 0.01° of visual angle, a sampling rate of 1000 Hz, and an accuracy range between 0.25° and 0.5°. The gaze tracking range is 32° horizontally and 25° vertically. Stimuli were presented on a flat computer screen, which was either a 19 in. or a 22 in. ViewSonic LCD display monitor. Participants' viewing was binocular, but the camera only recorded movements from one eye. To stabilize their heads during eye tracking, participants positioned their heads on a chin rest. They used a standard USB mouse to transition between stimuli and respond to questions.

Group Contingency

A group contingency intended to reinforce participants' correct responding and careful reading was included in the proposed study. Each participant earned one point for each correct response to the comprehension questions on the eye tracker assessment. At the end of the assessment, the points were added to the participant's class score. The class in each grade with the highest average score on the comprehension questions presented during the eye-tracker assessment earned a leisure event (e.g., pizza party, recess activity, special breakfast). Of note, the average score was used because it was not expected that all students from each class would

participate. In order to ensure consistency across administrations, participants were not informed of their class's progress.

Procedure

Eye tracking assessment. Participants completed one individualized eye-tracking session conducted at their school in a quiet room. First, examiners explained the group contingency by telling participants that they could help their class earn a leisure event by answering comprehension questions correctly on the eye tracking assessment. Participants learned that each question they answered correctly would count towards their class's average score, and the class with the highest average accuracy out of all participating classes in their grade would earn access to the leisure event. Participants were allowed to ask clarification questions before beginning the eye tracking assessment.

At the start of the assessment, participants sat in a chair and placed their heads on a chin rest. The eye-tracking camera and monitor faced the participant. Per tracking recommendations, participants' eyes were 50-55 cm from the camera and approximately 65 or 82 cm from the computer monitor, depending upon which display monitor was used for data collection. Chair height was also adjusted so that participants' forward gaze fell within the upper portion of the screen.

Before beginning the assessment, participants completed a practice trial allowing them to gain experience with reading questions from the computer screen and answering questions using the mouse. Participants read a short passage and answered three corresponding MC questions presented together (i.e., as they were presented in the control condition). To answer questions, participants clicked on circles adjacent to answer choices that resembled answer bubbles on standardized assessments. To move between screens (i.e., when participants were finished

reading the passage and/or responding to questions), participants clicked on a "next" arrow that resembled stimuli presented in standardized assessments.

After the practice trial, participants completed the eye-tracker calibration and validation process. This process ensured that the eye tracking system tracked eye movements correctly. When the process was complete, examiners informed participants that they would read three passages and answer five questions about each passage. Participants were also told that they must read passages silently and that they would not receive help with reading. Then, participants viewed a fixation dot in the center of the screen in order to ensure that participants' reading behavior was not biased by the location of their first fixation. A fixation dot was always presented before each new presentation of information (i.e., passage and/or questions).

Prior to each condition, examiners explained the condition parameters to participants so that they were aware of when they would read each passage and set of questions, as well as whether they would have access to the passage when reading the questions. See Appendix C for directions associated with each condition. The entire assessment was approximately 35-45 min in length. Participants were permitted to take breaks during the assessment if they became fatigued.

Control condition. In order to eliminate the possibility of experimental conditions influencing participants' natural reading behavior, the control condition was always presented first. Prior to the start of this condition, examiners read to participants the instructions for the control condition as specified in Appendix C. Then, participants viewed a passage presented alongside its five corresponding MC comprehension questions. The passage appeared in a column on the left half of the screen and the questions appeared in a column on the right half of the screen. The purpose of the control condition was to determine how students read a passage

and a set of questions as they are typically presented in standardized testing. That is, when students were able to move freely between a passage and corresponding questions.

Questions first (QF condition). Before the QF condition began, examiners explained procedures for the condition as described in Appendix C. During the QF condition, the five questions appeared first. After reading the questions, participants clicked the "next" arrow and the passage was then presented to the left of the same five MC questions (i.e., consistent with how the passage and questions were presented in the control condition).

Passage first (PF condition). Prior to the PF condition, participants were told the procedures for this condition as listed in Appendix C. The stimuli were presented as described to participants; students first read a passage and indicated that they were finished reading by clicking on the "next" arrow. Then, the questions were presented to the right of the passage so that students could respond to the questions while having access to the passage.

Passage and condition combinations. The presentation order of experimental conditions was randomized. Specifically, participants either viewed the QF condition first, followed by the PF condition or participants viewed the PF condition first followed by the QF condition. The two passages (i.e., Passage A and Passage B) presented in each experimental condition were also randomized, creating four possible condition and passage combinations.

Working memory assessment. Following the eye tracking assessment, undergraduate or graduate students trained in standardized testing procedures administered the *Digit Span* subtest of the WISC-V to participants. The assessment was conducted individually in a quiet location within the school. Administration time was approximately 10 min.

Oral reading assessment. In addition to the working memory assessment, participants also completed an oral reading assessment after the eye tracking portion of the study. For this

assessment, participants read aloud the three passages that they read on the eye tracker. Undergraduate and graduate examiners trained in curriculum-based measurement (CBM) procedures administered the assessment. Following participation in all three components of the study, participants selected a small prize.

Data Preparation and Coding

The eye tracking assessment yielded one data file for each participant that included eye movement data for all three conditions in the study. The files were prepared prior to analysis using a three-step process of *cleaning, drift correcting,* and *report running*. This process has been utilized in previous eye movement studies with children (e.g., Ardoin, Binder, Foster, & Zawoyski, 2016; Ardoin et al., 2015; Zawoyski, Ardoin, & Binder, 2015). Individuals who were trained and reliable in eye movement data preparation assisted in preparing the files. For *cleaning*, fixations outside the range of 120 ms and 800 ms were removed, because children are not likely to have fixation lengths that are outside of this range during reading. Some fixations were trimmed if they occurred off-screen (i.e., to the left and right of the stimuli). During *drift correction,* fixations were moved to the area of the passage or questions where participants were most likely to be reading. Drift correction allows for manual correction of fixations that may not accurately reflect the location of the participants' eyes, due to factors such as participants' movement during tracking and poor calibration. Interest areas were drawn around each word in the passage as well as each MC question with its four answer choices. Drift correction on the words in the passage involved moving fixations to the word that the participant was most likely to be reading. Drift correction on the questions involved moving fixations to the interest area surrounding the question and its answer choices on which the fixations were most likely to occur.

Finally, in *report running*, participants' eye movement data files were individually analyzed to obtain reports for all relevant variables, including those calculated on the passage (i.e., total fixation time and average fixation count) as well as on MC questions and answer choices (i.e., total fixation time, total fixation count). In the PF condition, passage data were collapsed across the first presentation (i.e., when the passage was presented without questions) and the second presentation (i.e., when the passage was presented with questions). That is, for each of the eye movement variables, an average was calculated across all words within both passage presentations. For the QF condition, question data were collapsed across the first presentation (i.e., when the questions were presented without the passage) and the second presentation (i.e., when the questions were presented with the passage). However, due to the nature of the interest areas selected for the questions, totals were obtained for all variables.

After data files were prepared, participants' reading behavior on the control passage was coded based on whether participants used a PF strategy or a QF strategy. The PF strategy was coded when participants read the entire passage before reading the questions. The QF strategy was coded when participants made three or more fixations on adjacent or consecutive words within the questions before finishing the passage.

Results

Description of the Sample

Of the 84 participants who completed the study, data for 17 participants were removed (20% of the sample). Reasons for exclusion included technical problems with the eye tracking data (n = 10) and scores that were greater than two standard deviations outside the sample mean for MAP and working memory data (n = 7). Participants were removed due to significantly discrepant working memory and reading achievement scores to ensure that the sample was

generally equivalent in its working memory and reading achievement levels. Lost data were generally equivalent across the four possible passage and condition combinations (i.e., ranging from 13.64% to 26.92% of participants presented with a given combination).

Descriptive statistics are provided in Table 2.1 and include MAP, working memory, and oral reading fluency data for participants in the analyzed sample, as opposed to data for all participants as reported in the Method section. Skewness and kurtosis values were acceptable for all variables (i.e., within positive and negative 2). Prior to analyses, the outlier labeling rule was applied to all variables using a *g* value of 2.2 (Hoaglin & Iglewicz, 1987; how2stats, 2011a, 2011b, 2011c). Specifically, the third quartile of values in the data set was subtracted from the first quartile and multiplied by *g*. The obtained value was subtracted from the first quartile to establish a lower bound for the data set. An upper bound for the data set was established by adding the obtained value to the third quartile. Any values outside of the lower and upper bounds were identified as outliers. Results indicated one significant outlier for average fixation count on answers to literal questions for Passage B. This value was winsorized (how2stats, 2011d, 2011e), resulting in winsorization for 1.49% of the data set.

The final sample size ($N = 67$) permitted detection of large effects, $\eta_p^2 = .250$, $\alpha = .050$, $\beta = .950$, in analyses conducted using mixed-design ANOVAs, according to Cohen's (1988) benchmark *f* value. Additionally, the sample size allowed for detection of large effects for analyses using independent *t*-tests, Cohen's $d = .800$, $\alpha = .050$, $\beta = .944$, and detection of medium effects for analyses using dependent *t*-tests, Cohen's $d = .500$, $\alpha = .050$, $\beta = .981$.

Reading Behavior on the Control Passage

Reading behavior on the control passage was coded to determine whether participants would choose to (1) read the passage first (i.e., use a PF strategy) or (2) read the questions first

(i.e., use a QF strategy) when permitted to read a passage and respond to questions as they normally would. Only 13 of the 67 participants (19.40%) utilized the QF strategy on the control passage. To ensure reliability of coding, 39% of control passages were coded by a second independent rater. Interobserver agreement (IOA) of passage coding was 100.00% for all scored files. Following coding, *z* scores were calculated for the 13 participants who engaged in the QF strategy. These *z* scores suggested that participants who elected the QF strategy did not engage in significantly different reading behavior from participants who used the PF strategy (see Table 2.2). Therefore, their data were included in all analyses reported below.

Global Analyses

A 2 x 2 mixed ANOVA was conducted to examine differences in participants' reading behavior across global eye movement measures of total fixation time of words in the passage, average fixation count across all words within the passage, total fixation time on the questions, and total fixation count on the questions. The between factor was passage (i.e., Passage A or Passage B) and the within factor was condition (i.e., PF, $n = 32$ or QF, $n = 35$). Results from Levene's Test of Equality of Error Variances indicated equal variances for all variables with the exception of the QF condition for total fixation time on the questions (Levene's test; $F(1, 65) = 4.931$, $p = .030$). However, given the robust nature of the ANOVA and cell counts that were greater than 15, findings are considered valid despite this violation of homogeneity of variance.

Results indicated no significant interactions between passage and condition for total fixation time on the passage, $F(1,65) = .011$, $p = .915$, $\eta_p^2 = .000$, average fixation count on the passage, $F(1,65) = .603$, $p = .440$, $\eta_p^2 = .009$, and total fixation time on the questions, $F(1,65) = 3.03$, $p = .086$, $\eta_p^2 = .045$ (see Appendix D). Additionally, there was no significant effect of passage for the following variables: total fixation time on the passage, $F(1,65) = .018$, $p = .894$,

η_p^2 = <.001, average fixation count on the passage, $F(1,65) = .140$, $p = .709$, $\eta_p^2 = .002$, and total fixation count on the questions $F(1,65) = .694$, $p = .408$, $\eta_p^2 = .011$. For all three measures, there was, however, a significant main effect of condition. Analyses indicated that the QF condition resulted in greater total fixation time on the passage, $F(1,65) = 50.41$, $p < .001$, $\eta_p^2 = .437$, higher average fixation count on the passage, $F(1,65) = 120.31$, $p < .001$, $\eta_p^2 = .649$, and greater total fixation time on the questions, $F(1,65) = 214.64$, $p < .001$, $\eta_p^2 = .768$, as compared to the PF condition.

Analyses of total fixation count on the questions revealed the only significant interaction between passage and condition on global measures, $F(1,65) = 6.83$, $p = .011$, $\eta_p^2 = .095$. Specifically, for the PF condition, participants made more fixations on Passage A questions (M = 292.03, SD = 82.74) as compared to Passage B questions (M = 247.60, SD = 66.40) whereas in the QF condition, participants made more fixations on Passage B questions (M = 438.31, SD = 90.76) as compared to Passage A questions (M = 432.91, SD = 115.39). Main effect data indicate that the passage effect was not significant for total fixation count on the questions, $F(1,65) = .969$, $p = .329$, $\eta_p^2 = .015$. Furthermore, the interaction diagram (see Appendix D) indicates that the effects of condition were consistent for both passages, with students in the QF condition making more fixations on the questions than students in the PF condition. This interpretation is supported by the significant main effect of condition, $F(1,65) = 301.44$, $p < .001$, $\eta_p^2 = .823$ and the marginal means for total fixation count on the questions (see Table 2.1).

Eye Movements on Areas of the Passage with and without Answers to Literal Questions

Eye movement data were analyzed separately across areas of each passage that did and did not contain answers to literal questions. Given that these interest areas could not be controlled for equivalence (e.g., character length), data for each passage were analyzed

separately. For each passage, an independent samples *t*-test was conducted examining differences between the QF and PF conditions' total fixation time and average fixation count on areas of the passage that contained answers to literal questions (see Table 2.3 for descriptive data) and areas of the passage that did not contain answers to literal questions (see Table 2.4 for descriptive data).

For both passages, results indicated that on areas of the passage containing answers to literal questions, the QF condition resulted in significantly longer average total fixation times, Passage A, $t(65) = -4.14$, $p < .001$; Passage B, $t(65) = -2.09$, $p = .041$. Similarly, for both passages, the QF condition also resulted in significantly longer total fixation times on areas of the passage that did not contain answers to literal questions, Passage A, $t(65) = -3.00$, $p = .004$; Passage B, $t(50.64) = -3.76$, $p < .001$. Of note, the correction for unequal variances was used for Passage B on total fixation time for areas of the passage not containing answers to literal questions due to a significant Levene's test for Equality of Variances ($F = 7.49$, $p = .008$).

A significant Levene's test for Equality of Variances was obtained for analyses of average fixation count on areas of the passage containing answers to literal questions for Passage A, $F = 15.85$, $p < .001$, and Passage B, $F = 9.37$, $p = .003$. Therefore, the correction for unequal variances was applied. Findings indicated that average fixation count on areas of the passage containing answers to literal questions was significantly higher when participants were reading in the QF condition as compared to the PF condition for both Passage A, $t(53.84) = -5.93$, $p < .001$, and Passage B, $t(47.63) = -4.92$, $p < .001$. Levene's test for Equality of Variances was also significant for analyses of average fixation count on areas of the passage not containing answers to literal questions for Passage A, $F = 9.27$, $p = .003$, and Passage B, $F = 9.61$, $p = .003$. Using the correction for equal variances not assumed, results for areas of the passage that did not

contain answers to literal questions were consistent with results for areas of the passage that did contain answers to literal questions. Specifically, participants in the QF condition made more fixations, on average, than participants in the PF condition when reading Passage A, $t(54.82) = -5.33$, $p < .001$ and Passage B, $t(48.51) = -6.23$, $p < .001$.

Accuracy

Due to a significant interaction between passage and condition for accuracy data, $F(1,65) = 13.64$, $p < .001$, $\eta_p^2 = .173$ (see Appendix D), a follow-up dependent *t*-test was conducted to examine whether participants' accuracy differed based on the passage that was presented. Results indicated that across conditions, participants responded with significantly higher accuracy on Passage B ($M = 3.96$, SD = 1.20) as compared to Passage A ($M = 3.36$, SD = 1.08; $t(66) = -3.72$, $p < .001$). Chi-squared analyses were conducted to further explore accuracy by condition, by passage, and by question (see Table 2.5). Overall, results for Passage A indicated that participants' response accuracy on questions did not differ as a function of condition with the exception of Question 2. Participants in the PF condition responded to this question with significantly poorer accuracy (34.38% responded correctly) as compared to participants in the QF condition (60.00% responded correctly). Chi-squared analyses for Passage B indicated that participants' accuracy was consistent across conditions for all questions.

Efficiency

The total time participants spent in each condition was recorded and compared in order to estimate the efficiency of each condition. Results indicated a significant interaction between passage and condition, $F(1,65) = 12.61$, $p = .001$, $\eta_p^2 = .162$. Specifically, participants in the PF condition spent more time (in ms) completing Passage A ($M = 187134.59$, $SD = 53025.60$) as compared to Passage B ($M = 161877.80$, $SD = 48331.30$). For the QF condition, participants

spent more time (in ms) completing Passage B (M = 214977.23, SD = 50150.55) as compared to Passage A (M = 209504.97, SD = 59639.27). Further inspection of the interaction (see Appendix D) revealed that regardless of passage, the PF condition was more efficient than the QF condition. This interpretation is supported by the significant effect of condition, $F(1,65) = 76.03$, $p < .001$, $\eta_p^2 = .539$ and the lack of significant passage effects, $F(1,65) = .661$, $p = .419$, $\eta_p^2 = .010$.

Procedural Integrity and IOA

Eye tracking assessment. During the eye tracking assessment, researcher behavior was monitored for 38.81% of sessions. A seven-item checklist examining each step in the eye tracking assessment was scored by a second researcher. Results indicated a mean procedural integrity score of 6.96, with a score below 7 occurring on only one occasion. For this occasion, the examiner received a procedural integrity score of 6 because the examiner required a prompt to read the instructions for the practice questions. Given that the instructions were still provided, assessment procedures were wholly implemented for the participant in question. Overall, procedural integrity was strong for administration of eye tracking procedures, suggesting that participants received a consistent and valid assessment.

Working memory. Procedural integrity data were collected by a second independent rater for 33.33% of the sessions. A 73-item checklist was created to ensure that examiners read instructions and items as indicated by the WISC-V manual. The average procedural integrity score was 95.29% (range = 69.86%-100.00%). Only two observations were scored with below 75.00% procedural integrity. Low scores were obtained due to slight examiner error in the speed that digits were read.

IOA data were calculated for 33.33% of the working memory assessment administrations. An independent observer scored 22 working memory assessments from the recordings to determine agreement with scores obtained by the examiner for each portion of the Digit Span subtest. Mean IOA was 99.31% (range = 92.98%-100.00%).

Oral reading fluency. A second rater independently listened to 34.84% of oral reading fluency session recordings to determine whether examiners had implemented procedures correctly. The oral reading assessment was divided into 13 steps encompassing instruction administration, timing, and prompting (i.e., asking participants to move to the next word if 10 s had elapsed). Of the recordings sampled, average procedural integrity was 94.31% (range = 76.67%-100.00%). Imperfect procedural integrity scores were typically obtained when examiners modified instructions or made errors in timing. Instruction modification was generally consistent with the meaning of the original instructions, but was scored as an error due to deviation from the words specified in the protocol. Timing errors occurred when examiners began timing before a participant read the first word of the story. Additionally, a timing error was scored on two occasions when the examiner recorded the entire duration of reading for all passages instead of timing each reading individually.

IOA was calculated for 34.33% of oral reading fluency sessions using recordings that were scored by an independent observer. Agreement was calculated by first taking the difference between the number of errors calculated by the examiner and the number of errors calculated by the independent rater. This difference was subtracted from the total number of words in the passage. Then, the obtained value was divided by the total number of words in the passage and multiplied by 100 to obtain a percentage of agreement. The mean IOA across the 23 assessed sessions was 98.77% (range = 96.16%-100.00%).

Eye tracking data cleaning. Given the potential subjectivity of drift correcting, IOA was calculated for 33% of eye tracking data files to ensure that data cleaners were reliable in fixation placement. Agreement was calculated using the mean count per interval method. For this method, the smallest number of fixations for one interest area was divided by the largest number of fixations on that word or interest area to obtain a percentage of agreement. This calculation was repeated across the remaining interest areas for an entire data file. The average was then calculated and used as the mean count per interval score. IOA for data cleaning was excellent (*M* = 95.35%, range = 86.18%-99.12%).

Discussion

The purpose of the current study was to examine the reading behavior of elementary students during a reading comprehension assessment under control and experimental conditions utilizing eye tracking technology. Experimental conditions included the PF condition, in which participants read a passage before reading its corresponding questions, and the QF condition, in which participants read the questions before reading their corresponding passage. Third- and fourth-grade participants read passages and responded to corresponding MC questions from a computer screen as their eye movements were recorded. Participants' reading behavior was coded on the control passage to identify whether they selected a PF or QF strategy under natural conditions. Analyses for experimental conditions examined participants' reading behavior across the passage and on areas of the text containing and not containing answers to literal questions as well as their accuracy on MC questions and their efficiency in task completion.

Reading Behavior on the Control Passage

Consistent with the hypothesized outcome, evidence from coding data on the control passage revealed that most students utilized a PF strategy (i.e., 80.60%). Third- and fourth-grade

students' testing-taking strategy usage was also consistent with findings from a study with second-grade students (i.e., Ardoin et al., 2015) and research with adult populations (i.e., Farr et al., 1990; Lewandowski et al., 2013). In turn, very few students utilized a QF strategy on the control passage (i.e., 19.40% of the sample). Findings closely replicated the outcomes obtained by Ardoin et al. (2015), in which only 24.00% of participants utilized a QF strategy. Furthermore, analysis of *z*-score data indicated that participants selecting a QF strategy did not significantly differ in their eye movements during reading on the control passage or in their accuracy on the MC questions. Therefore, data for participants electing a PF and QF strategy were combined for analyses on experimental passages.

Eye Movement Data

Global analyses indicated no interaction effects between passage (i.e., Passage A or Passage B) and condition (i.e., PF or QF) across eye movement variables on the passage (i.e., total fixation time, average fixation count) or for total fixation time on the questions. Significant main effects of condition across all eye movement variables indicated that when participants completed the QF condition, they made longer and more frequent fixations on the passage than when participants completed the PF condition. This finding was in conflict with the hypothesis that participants would engage in higher levels of eye movements indicative of comprehension during the PF condition as well as prior research suggesting that adults using a PF strategy gained a more comprehensive understanding of the text than adults who utilized a QF strategy (Cerdán et al., 2009).

There was an interaction between passage and condition for total fixation count on the questions (see Appendix D). Given that participants in the QF condition made more fixations on the questions as compared to participants in the PF condition regardless of whether they were

reading Passage A or Passage B, findings for this global measure are considered generally consistent with findings for other global variables. It is possible that the interaction may be a reflection of question difficulty; perhaps participants returned to re-fixate on questions more frequently for questions in Passage A as compared to questions in Passage B. However, this interpretation cannot be confirmed without an analysis of regressions on the questions.

Similar to global analyses, the analyses on areas of each passage containing and not containing answers to literal questions indicated that participants' total fixation time was longer and their average fixation count was higher in the QF condition as compared to the PF condition. These findings are consistent with the hypothesis that participants would make more frequent and longer fixations on regions of the passage with answers to literal questions, but are in contrast to the hypothesis that participants in both conditions would engage in similar reading behavior on regions of the text that did not contain answers to literal questions. It is likely that participants in the QF condition engaged in longer and more frequent fixations on the passage as a whole as compared to participants in the PF condition because they were searching the passage for answers rather than reading the passage for its meaning.

Accuracy

Participants in the PF condition were expected to attain higher accuracy on MC questions because they were anticipated to read the text more closely, based on prior research (e.g., Cerdán et al., 2009). Mixed ANOVA analyses on MC question accuracy revealed a significant interaction between passage and condition. Although participants across conditions had greater accuracy on Passage B, the interaction revealed that participants reading in the PF condition attained higher accuracy on Passage B and participants reading in the QF condition attained higher accuracy on Passage A. Due to this interaction, analyses were conducted by passage.

Within-passage differences by questions revealed that for Passage B, participants generally responded consistently to each question across conditions. Findings were similar for Passage A, with the exception that participants in the QF condition scored significantly higher on Question 2 as compared to participants in PF. According to the Easiness Index data, Question 2 was the most difficult item presented to students across passages (i.e., Easiness Index score of .530 for students assessed in the fall and .540 for students assessed in the spring). Although it is important to avoid over-interpretation of findings for a single MC question, participants' performance on this item raises interesting questions for future exploration. Notably, Passage A Question 2 was a literal question, with an easily accessible response in the text. However, the distractor items were plausible choices and may have been selected based on students' inferences from their prior knowledge. It is possible that participants in the PF condition were more likely to utilize prior knowledge when responding, whereas participants in the QF condition were primed to locate the correct answer while reading based on their prior exposure to the questions. Additionally, the finding that the QF condition may have facilitated participants' accuracy relative to the PF condition is in contrast to results from Wiesendanger and Wollenberg (1978), which indicated that a QF strategy (i.e., prequestioning) does not promote improved performance on literal questions. Yet, as suggested by Bishop (2001) participants' prior experience with test-taking strategies may have impacted their effectiveness. Overall, results from the current study contribute to previous findings suggesting that additional and possibly individualized factors may influence the effectiveness of test-taking strategies, such as prior knowledge, question type, and opportunities to practice test-taking strategies. Additionally, findings suggested that even though participants in the QF condition spent more time reading the passage than participants in

the PF condition, their attempts were not sufficient to improve their accuracy for most reading comprehension questions.

Efficiency

Although analyses of efficiency data revealed a significant interaction between passage and condition, regardless of the passage presented, participants required more time to read and answer questions under the QF condition as compared to the PF condition. This finding was consistent with expectations for the current study and supports arguments posed by research with children (Bishop, 2001; Bishop & Frisbie, 1999) as well as research with adults (e.g., Cerdán et al., 2009; Daneman & Hannon, 2001), in that individuals across age groups require more time to complete comprehension tests when they engage in a QF strategy.

Limitations and Future Directions

Results should be interpreted with consideration of several limitations. Perhaps the most significant limitation is that the current study evaluated participants' reading behavior and response accuracy on only one passage and question set per condition. Reading comprehension studies typically present multiple passages due to concerns with inter-passage reliability and validity. An additional consideration is that coding on participants' natural reading behavior was conducted on the first passage they read. Given results from Farr et al. (1990) indicating that adult readers often begin by using a PF strategy and switch to a QF strategy if they become uninterested or fatigued during a longer assessment, future studies should examine elementary students' reading strategies during a longer assessment with more passages. Including more passage and question sets in an eye movement study may provide important insight into how students' reading behavior and strategy usage changes throughout a natural testing period.

A second limitation pertains to the questions selected for the experimental conditions. As evident from analyses on accuracy data and information regarding item difficulty, Passage B consisted of easier questions than Passage A. Findings regarding accuracy may have been more comparable if item difficulties were more similar. Unfortunately, due to limited passage and question sets available within the GMRT as well as the need to balance passage length and difficulty, item difficulty across experimental passages was not ideal.

Limitations also arose from restrictions of current eye tracking technology. Due to the size of items presented on the screen, it was not possible to analyze eye movements on individual response choices. This study may have benefitted from analyses on response choices, particularly for Passage A, Question 2, in order to determine participants' reading behavior on distractors. An additional limitation of this study related to eye tracking technology is that participants' early reading behavior (e.g., first fixation duration) was not examined because these measures could not be calculated across multiple presentations of a passage. Data on participants' early reading skills could be valuable in explaining differences in accuracy. For example, participants in Ardoin et al. (2015) with longer early reading measures demonstrated poorer accuracy on comprehension questions. It is possible that participants' accuracy in the current study could have been explained by differences in their early reading behavior. Given that MAP scores were not significantly different across the sample, this possibility was somewhat controlled for. Yet, future eye movement studies should consider examining eye movements thought to represent early reading behavior to examine its influence on accuracy and its relationship to eye movement measures related to higher-level reading behavior.

Furthermore, the test-taking literature can be further extended through a more in-depth analysis of participants' question-responding behavior. That is, future studies may benefit from

item-level analyses of how students respond to questions within a passage and their latency to identifying a correct answer. Such an analysis could also be beneficial in diagnosing students' errors in reading comprehension and providing highly individualized feedback.

A final suggestion for future research pertains to the role of working memory in strategy efficiency. Results from Daneman and Hannon (2001) indicating that working memory capacity was positively correlated with adults' performance when using a QF strategy suggest that perhaps children with stronger working memory abilities are more likely to benefit from the QF strategy, as compared to children with lower working memory abilities. Future studies on reading comprehension should consider students' working memory ability given its exclusive contribution to reading comprehension skills (Andreassen & Bråten, 2010; Cain et al., 2004).

Summary

In summary, results from the current study revealed that although teachers often encourage students to read questions prior to reading their corresponding passage, third- and fourth-grade students participating in the current study did not typically utilize a QF strategy when reading a passage and answering comprehension questions. Interestingly, reading behavior of elementary students in the current study during the QF condition is similar to that of adults (e.g., Cerdán et al., 2009; Daneman & Hannon, 2001), with the QF strategy generally requiring more time than the PF strategy with little to no gains in accuracy. Therefore, in spite of some test preparation materials recommending a QF strategy (See Appendix A), students are electing to utilize a PF strategy. When participants were required to engage in a PF and QF strategy, the QF strategy resulted in less efficient test-taking with limited gains in accuracy. Additionally, participants reading in the QF condition made longer and more frequent fixations across the entire passage and question set as well as on areas of the passage that did and did not contain

answers to literal questions. By examining participants' question-responding behavior, the current study also extended eye movement research conducted by Solheim and Uppstad (2011) to provide a preliminary understanding of underlying test-taking behavior in elementary students.

In general, more research is needed to determine situations when a QF strategy may be more beneficial than a PF strategy. However, based on results from the current study, elementary students should be encouraged to utilize a PF strategy due to its efficiency and general consistency with the QF strategy on outcome measures. Although improved accuracy on the QF condition for Passage A, Question 2 (i.e., the challenging literal question) raises interesting hypotheses regarding the type of questions that may result in improved performance with a QF strategy, perhaps a more efficient solution would be for teachers to encourage students to find support for their answers within the text rather than rely on prior knowledge. Evidence suggests that students may not naturally check their responses. For example, Bishop (2001) proposed that students may be hesitant to return to text if not explicitly instructed to do so. Elementary students' failure to return to text to check their answers has been observed in another eye-movement study (i.e., Ardoin et al., 2015). Overall, strategies recommended by teachers should be consistent with the purpose of reading comprehension assessment and encourage students to read more effectively, rather than search the text for answers.

References

Andreassen, R. & Bråten, I. (2010). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading, 33,* 263-283. doi:10.1111/j.1467-9817.2009.01413.x

Ardoin, S. P., Binder, K. S., Foster, T. E., & Zawoyski, A. M. (2016). A randomized control design study examining the effects of repeated readings on reading achievement and reading behavior. *Journal of School Psychology, 59,* 13-38. doi:10.1016/j.jsp.2016.09.002

Ardoin, S. P., Zawoyski, A. M., Wagner, L., Bangs, K., & Binder, K. S. (2015). *Measuring test-taking behavior: Different behaviors but similar outcomes.* Manuscript submitted for publication.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience, 4,* 829-839. doi:10.1038/nrn1201

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47-89). San Diego, CA: Academic Press.

Barksdale-Ladd, M. A., & Thomas, K. F. (2000). What's at stake in high-stakes testing: Teachers and parents speak out. *Journal of Teacher Education, 51,* 384-397. doi:10.1177/0022487100051005006

Binder, K. S. (2003). Sentential and discourse topic effects on lexical ambiguity processing: An eye movement examination. *Memory and Cognition, 31,* 690-702. doi:10.3758/BF03196108

Binder, K. S., & Morris, R. K. (1995). Eye movements and lexical ambiguity resolution: Effects of prior encounter and discourse topic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1186-1196. doi:10.1037/0278-7393.21.5.1186

Bishop, N. S. (2001, April). *The validity of reading comprehension test scores: Evidence of generalizability across difference test administration conditions.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, Washington.

Bishop, N. S. & Frisbie, D. A. (1999, April). *The effects of different test-taking conditions on reading comprehension test performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Cain, K. (2006). Children's reading comprehension: The role of working memory in normal and impaired development. In S. J. Pickering (Ed.), *Working memory and education* (pp. 61-91). Cambridge, MA: Academic Press.

Cain, K., Oakhill, J. V., & Bryant, P. E. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96,* 31–42. doi:10.1037/0022-0663.96.1.31

Calkins, L., Montgomery, K., & Santman, D. (1999). Helping children master the tricks and avoid the traps of standardized tests. *Practical Assessment, Research, & Evaluation, 6,* 1-3. Retrieved from http://pareonline.net/getvn.asp?v=6&n=8

Cerdán, R., Vidal-Abarca, E., Martínez, T., Gilabert, R., & Gil, L. (2009). Impact of question answering tasks on search processes and reading comprehension. *Learning and Instruction, 19,* 13-27. doi:10.1016/j.learninstruc.2007.12.003

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cordón, L. A. & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology, 88,* 288-295. doi:10.1037/0022-0663.88.2.288

Daneman, M. & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology, 130,* 208-223. doi:10.1037/0096-3445.130.2.208

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27,* 209-226. doi:10.1111/j.1745-3984.1990.tb00744.x

Foster, T. E., Ardoin, S. P., & Binder, K. S. (in press). Reliability and validity of eye movement measures of children's reading. *Reading Research Quarterly.*

Gorin, J. (2006). Test construction and diagnostic testing. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education: Theory and Applications* (pp. 195-198). New York, NY: Cambridge University Press.

Hoaglin, D. C., & Iglewicz, B. (1987). Fine tuning some resistant rules for outlier labeling, *Journal of the American Statistical Association, 82,* 1147-1149. doi:10.1080/01621459.1987.10478551

Holtzman, D. (2017, March 7). Tips for standardized test reading passages. *Chariot Learning.* Retrieved from http://chariotlearning.com/tips-for-standardized-test-reading-passages/

how2stats (2011a, September 8). The right way to detect outliers – Outlier labeling rule (part 1). [Video file]. Retrieved from https://www.youtube.com/watch?v=WSflSmcNRFI

how2stats (2011b, September 8). The right way to detect outliers – Outlier labeling rule (part 2). [Video file]. Retrieved from https://www.youtube.com/watch?v=2HmopqF6V6w&t=80s

how2stats (2011c, September 8). The right way to detect outliers – Outlier labeling rule (part 3). [Video file]. Retrieved from https://www.youtube.com/watch?v=bRdC1u9veg8&t=77s

how2stats (2011d, October 5). Dealing with outliers (part 1). [Video file]. Retrieved from

https://www.youtube.com/watch?v=Ukkcer70r5A&t=1s

how2stats (2011e, October 5). Dealing with outliers (part 2). [Video file]. Retrieved from https://www.youtube.com/watch?annotation_id=annotation_204681&feature=iv&src_vid=Ukkcer70r5A&v=FatA5COFIPU

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2007). *Iowa tests of basic skills (ITBS).* Rolling Meadows, IL: Riverside Publishing.

Hyönä, J., & Niemi, P. (1990). Eye movements during repeated reading of a text. *Acta Psychologica, 73,* 259-280. doi:10.1016/0001-6918(90)90026-C

Kwabi, S., Xu, H., Binder, K. S., Nemier, C., & Ardoin, S. P. (2015). *Eye movements and reading comprehension performance: Examining the relationships among test format, working memory capacity, and reading strategies.* Manuscript submitted for publication.

Lewandowski, L., Gathje, R. A., Lovett, B. J., & Gordon, M. (2013). Test-taking skills in college students with and without ADHD. *Journal of Psychoeducational Assessment, 31,* 41-52. doi:10.1177/0734282912446304

Martínez, T. & Sellés, P. (2001). *Read&Answer: Software for tracking students' question answering behavior.* Unpublished software, University of Valencia, Spain.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie Reading Tests, Fourth Edition, Level 3 Form S.* Rolling Meadows, IL: The Riverside Publishing Company.

Northwest Evaluation Association (2004). *Reliability and validity estimates: NWEA achievement level tests and measures of academic progress.* Lake Oswego, OR: Author.

Northwest Evaluation Association (2009). *Technical manual for Measures of Academic Progress and Measures of Academic Progress for Primary Grades.* Lake Oswego, OR: Author.

Northwest Evaluation Association (2015, August). *NWEA Measures of Academic Progress Normative Data.* Retrieved from: https://www.nwea.org/content/uploads/2015/06/2015-MAP-Normative-Data-AUG15.pdf

Perlman, C. L. (2003). Practice tests and study guides: Do they help? Are the ethical? What is ethical test preparation? In J. E. Wall & G. R. Walz (Eds.), *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators* (pp. 387-396). Greensboro, NC: CAPS Press.

Radach, R., & Kennedy, A. (2013). Eye movements in reading: Some theoretical context. *The Quarterly Journal of Experimental Psychology, 66*, 429-452. doi:10.1080/17470218.2012.750676

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422. doi:10.1037/0033-2909.124.3.372

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading, 10,* 241-255. doi:10.1207/s1532799xssr1003_3

Rothkopf, E. Z., & Bisbicos, E. E. (1967). Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology, 58,* 56-61. doi:10.1037/h0024117

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shape the construct: A cognitive processing perspective. *Language Testing, 23*, 441-474. doi:10.1191/0265532206lt337oa

SAT critical reading prep guidelines (n.d.). Retrieved from http://www-tc.pbs.org/now/classroom/acrobat/lesson02.pdf

SAT reading strategies (n.d.). *Nova Press.* Retrieved from http://novapress.net/sat/sat-strategies/gre-reading-strategies/

Stratakis-Allen, H. G. (2007). Long critical reading passages. In M. Sanders (Ed.), *The complete idiot's guide to acing the GRE.* (p. 82). New York, NY: Penguin Group.

Solheim, O. J. & Uppstad, P. H. (2011). Eye-tracking as a tool in process-oriented reading test validation. *International Electronic Journal of Elementary Education, 4,* 153-168.

Test-taking strategies for three subject areas (n.d.). *Scholastic.* Retrieved from http://www.scholastic.com/teachers/article/test-taking-strategies-three-subject-areas

Top 25 test-taking tips, suggestions, & strategies (n.d.). Retrieved from http://www.lake.k12.fl.us/cms/lib05/FL01000799/Centricity/domain/17/health%20sciences%20blueprints/unit%201/TestTakingTop25.pdf

van der Schoot, M., Vasbinder, A. L., Horsley, T. M., & van Lieshout, E. C. D. M. (2008). The role of two reading strategies in text comprehension: An eye fixation study in primary school children. *Journal of Research in Reading, 31,* 203–223. doi:10.1111/j.14679817.2007.00354.x

Vorstius, C., Radach, R., Mayer, M., & Lonigan, C. (2013). Monitoring local comprehension monitoring in sentence reading. *School Psychology Review, 42,* 191–206.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children – Fifth Edition Technical and Interpretive Manual.* Bloomington, MN: NCS Pearson.

Wiesendanger, K. D., Birlem, E. D., & Wollenberg, J. (1982). A summary of studies related to the effect of question placement on reading comprehension. *Reading Horizons, 23*, 15-21.

Wiesendanger, K. & Wollenberg, J. (1978). Prequestioning inhibits third graders' reading comprehension. *The Reading Teacher, 31*, 892-895.

Zawoyski, A. M., Ardoin, S. P., & Binder, K. S. (2015). Using eye tracking to observe differential effects of repeated readings for second-grade students as a function of achievement level. *Reading Research Quarterly, 50,* 171-184. doi:10.1002/rrq.91

Table 2.1

Descriptive Statistics for Variables in Chapter 3

Measure *(N = 67)*	*M*	*SD*	Skewness	Kurtosis
MAP Score (RIT score)	206.22	13.85	-.522	.013
Working Memory (Scaled score)[+]	10.62	2.41	-.347	-.278
ORF Median (WRCM)[+]	126.17	28.91	-.046	-.425
PF Average Total Fixation Time on Passage (ms)	465.97	110.73	0.966	1.902
QF Average Total Fixation Time on Passage (ms)	582.44	153.13	0.498	-0.161
PF Average Fixation Count on Passage (#)	1.21	.367	0.672	0.192
QF Average Fixation Count on Passage (#)	1.97	.632	0.586	-0.539
PF Total Fixation Time on Questions (ms)	69763.73	21832.07	0.535	-0.205
QF Total Fixation Time on Questions (ms)	112257.28	31840.77	0.173	-0.056
PF Total Fixation Count on Questions (#)	268.85	77.38	0.693	0.421
QF Total Fixation Count on Questions (#)	435.73	102.49	0.203	-0.419
PF Accuracy (#)	3.66	1.10	-.486	-.253
QF Accuracy (#)	3.66	1.26	-.855	.121
PF Efficiency (ms)	173940.75	51822.50	0.537	-0.016
QF Efficiency (ms)	212363.61	54533.19	0.235	-0.566
Passage A Total Fixation Time on Answers to Literal Questions (ms)	562.17	176.76	0.564	0.537
Passage B Total Fixation Time on Answers to Literal Questions (ms)	522.82	170.22	0.850	0.095
Passage A Average Fixation Count on Answers to Literal Questions (#)	1.79	.814	0.829	-0.208
Passage B Average Fixation Count on Answers to Literal Questions (#)	1.43	.700	1.15	1.32
Passage A Total Fixation Time on Non-Literal Question Areas (ms)	518.45	148.15	.893	.455
Passage B Total Fixation Time on Non-Literal Question Areas (ms)	522.56	146.21	.925	1.02
Passage A Average Fixation Count on Non-Literal Question Areas (#)	1.60	.625	1.04	.702
Passage B Average Fixation Count on Non-Literal Question Areas (#)	1.56	.653	.936	.500

[+] N = 66

Table 2.2

Descriptive Statistics for Participants using a QF Strategy on the Control Passage

Measure *(n = 13)*	*M* z score	*SD* z score
MAP score	-.08	1.13
Working memory	.06	.61
ORF Median	-.10	.93
Total fixation time	.38	.70
Average fixation count	.30	.60
Total fixation time on questions	.39	1.15
Total fixation count on questions (control passage)	.41	1.12
Accuracy (control passage)	-.03	.96

Table 2.3

Descriptive Statistics for Eye Movement Data on Answers to Literal Questions

	Total Fixation Time		Average Fixation Count	
Passage/Condition	*M*	*SD*	*M*	*SD*
Passage A/PF (n = 32)	478.40	133.17	1.30	.450
Passage A/QF (n = 35)	638.76	178.44	2.24	.816
Passage B/PF (n = 35)	482.32	147.58	1.08	.416
Passage B/QF (n = 32)	567.11	184.19	1.81	.747

Table 2.4

Descriptive Statistics for Eye Movement Data on Non-Literal Question Areas of the Passage

	Total Fixation Time		Average Fixation Count	
Passage/Condition	*M*	*SD*	*M*	*SD*
Passage A/PF (n = 32)	464.76	126.17	1.25	.366
Passage A/QF (n = 35)	567.54	151.31	1.92	.643
Passage B/PF (n = 35)	462.91	99.61	1.17	.368
Passage B/QF (n = 32)	587.80	161.92	1.98	.640

Table 2.5

Summary of Condition Differences for Performance on Individual MC questions by Passage, Chi-Squared Test of Independence

Passage/Question	% Correct when Passage was PF	% Correct when Passage was QF	X^2	df	*p*
Passage A					
Question 1	96.88%	82.86%	+	+	.108+
Question 2	34.38%	60.00%	4.40	1	.036*
Question 3	62.50%	68.57%	.273	1	.601
Question 4	78.13%	82.86%	.239	1	.625
Question 5	62.50%	42.86%	2.59	1	.108
Passage B					
Question 1	77.14%	87.50%	1.22	1	.269
Question 2	82.86%	71.88%	1.16	1	.281
Question 3	82.86%	78.13%	.239	1	.625
Question 4	88.57%	84.38%	+	+	.727+
Question 5	62.86%	75.00%	1.15	1	.285

+ = Two-tailed Fisher's exact test; expected cell count less than five for two cells
*Significant between-groups differences, p <.05

APPENDIX A

TEST-TAKING STRATEGIES RECOMMENDED BY EDUCATIONAL RESOURCES

Recommendation: Reading the Questions First
Top 25 test-taking tips, suggestions, & strategies (n.d.).: *When faced with a reading passage followed by a series of questions, teach students to read the questions first and then read the passage."*
SAT critical reading prep guidelines (n.d.).: *"Before reading the passage, take time to review the questions associated with the passage so you know what you should be looking for as you read."*
Test-taking strategies for three subject areas (n.d.).: "*When you have a passage to read followed by questions, **read the questions first.** This will give you a good idea of what to look for as you read the passage.*"
Calkins, Montgomery, and Santman (1999).: "*Help children develop scavenger-hunt-type list of things to look for as they read the passages by having them read the questions first."*
Recommendation: Reading the Passage First
Stratakis-Allen (2007). *"Here's what you should do with each passage before going to the questions. (By the way, do this before you look at the questions – don't read the questions first.)"*
Holtzman (2017): "*There is not enough time to read the questions before the passage, read through the passage, and then answer the questions."*
SAT reading strategies (n.d.).: *Many books recommend reading the questions before the passage. But there are two big problems with this method. First, some of the questions are a paragraph long, and reading a question twice can use up precious time."*

APPENDIX B

EYE MOVEMENT DEPENDENT MEAURES IN CHAPTER 2

Measure	Definition	Processing Type (Binder, 2003; Binder & Morris, 1995; Foster et al., in press; Hyönä & Niemi, 1990)
Total Fixation Time	duration (in ms) of all fixations made on a word	late processing
Average Fixation Count	average number of fixations made on each word	late processing
Total Fixation Count	total number of fixations made in an area	late processing

APPENDIX C

INSTRUCTIONS FOR CONDITIONS IN CHAPTER 2

Prior to the start of each condition, participants were read the following instructions:

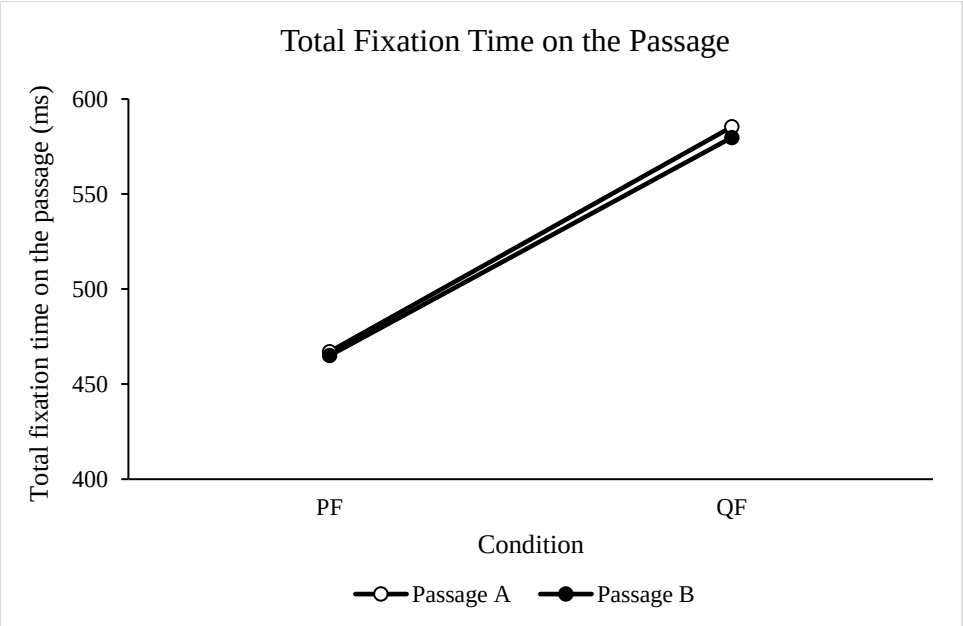
Control: *In a second, you'll see a story and multiple-choice questions. Please read the story and answer the questions just like you always do when you take reading tests. Be sure to answer all the questions. Click on "next" when you're done. Remember, these are just practice questions.*

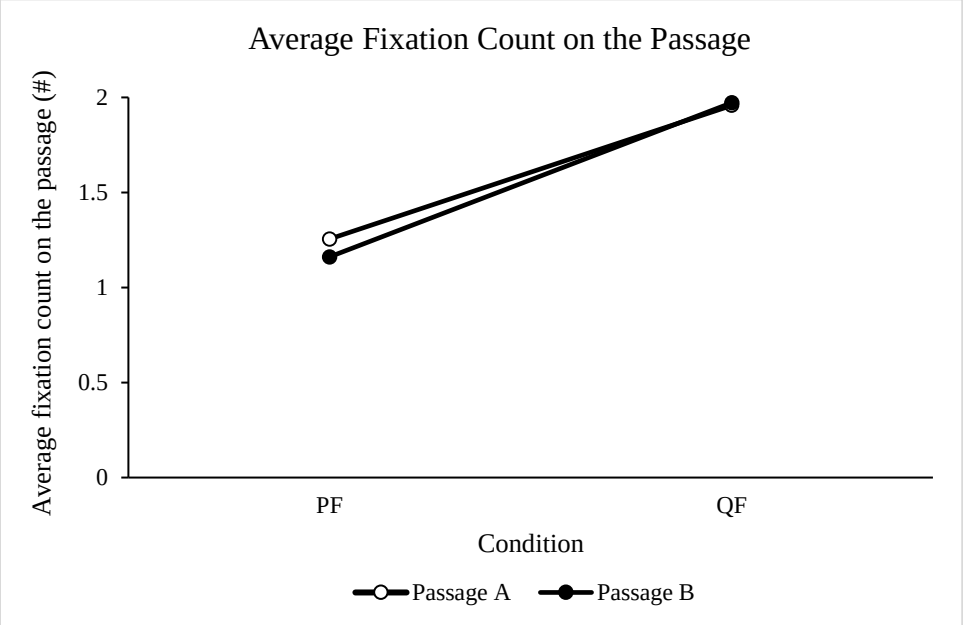
Questions first (QF): *This time, you will start by reading the questions first. Please make sure to read all the questions. After you finish reading all of the questions, we will give you the mouse so you can click "next." Then, you will see the story and the same multiple choice questions you read. Be sure to answer all the questions.*

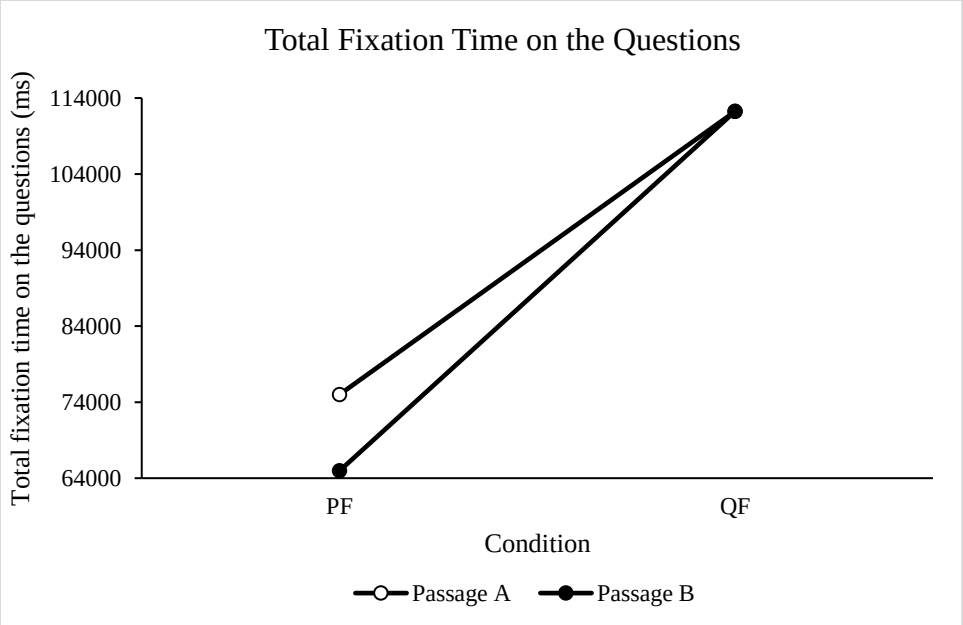
Passage first (PF): *This time, you will start by reading the story first. Please make sure to read the entire story. After you finish reading the story, we will give you the mouse so you can click "next." Then, you will see the same story and the multiple-choice questions for the story. Be sure to answer all the questions.*

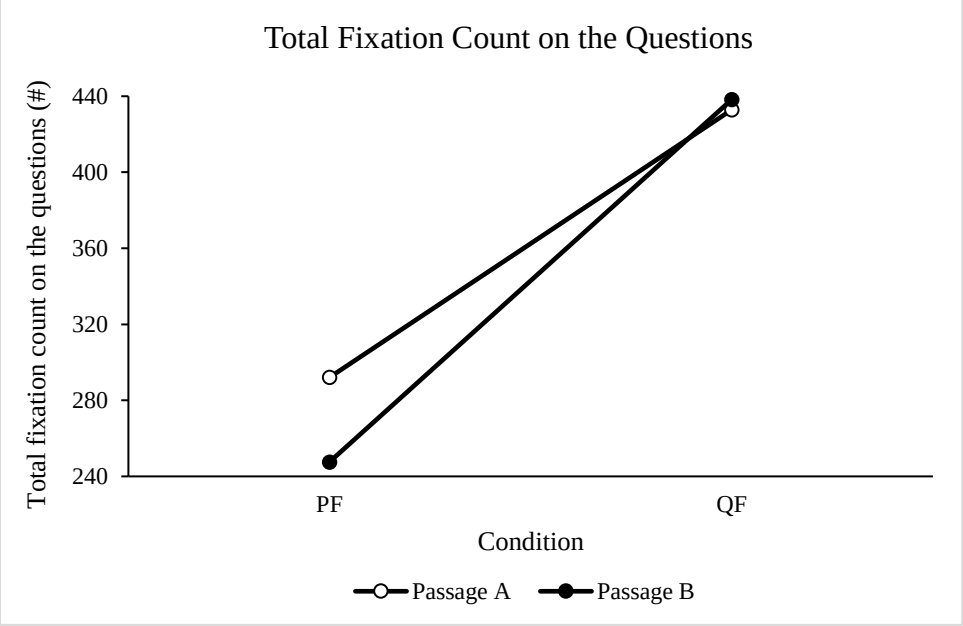
APPENDIX D

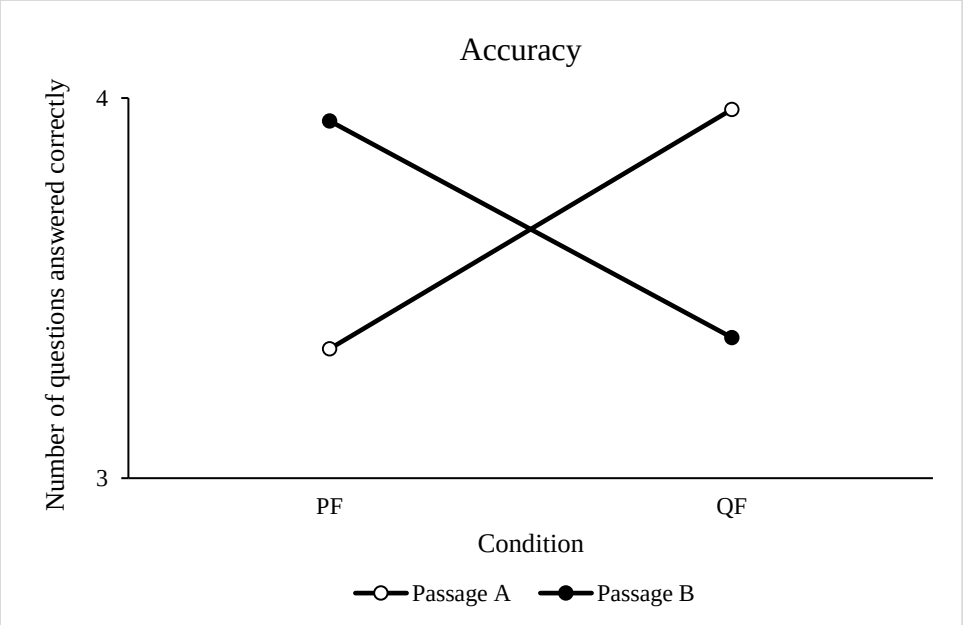
INTERACTION DIAGRAMS FOR GLOBAL ANALYSES

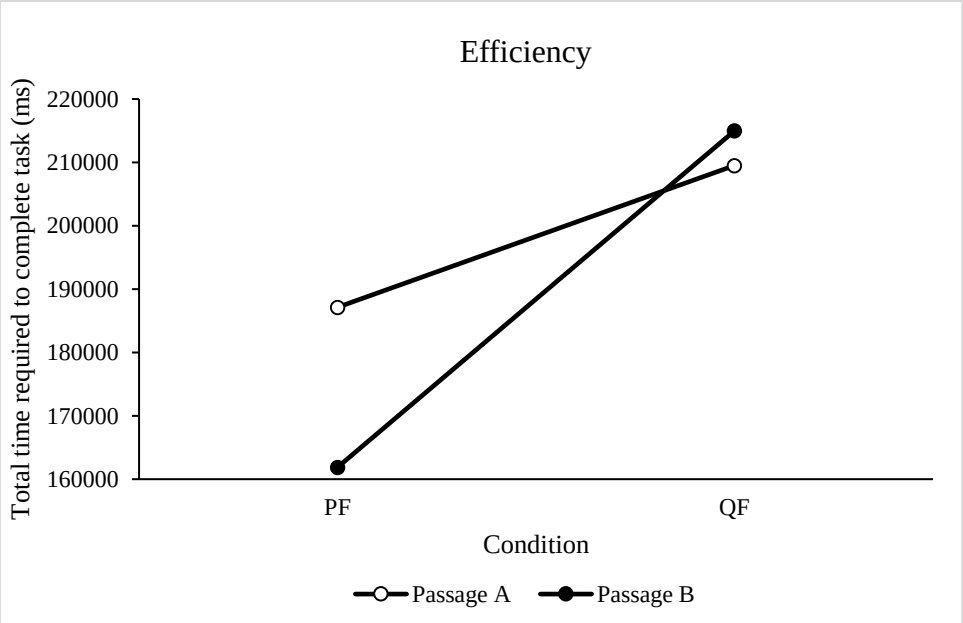





Total Fixation Time on the Questions
Total fixation time on the questions (ms)
114000
104000
94000
84000
74000
64000
PF
QF
Condition
Passage A
Passage B


Total Fixation Count on the Questions
Total fixation count on the questions (#)
440
400
360
320
280
240
PF
QF
Condition
Passage A
Passage B


Accuracy
Number of questions answered correctly
4
3
PF
QF
Condition
Passage A
Passage B


Efficiency
Total time required to complete task (ms)
220000
210000
200000
190000
180000
170000
160000
PF
QF
Condition
Passage A
Passage B

CHAPTER 3

USING EYE-TRACKING TECHNOLOGY TO EXAMINE THE IMPACT OF QUESTION FORMAT ON READING BEHAVIOR IN ELEMENTARY STUDENTS[2]

[2] Zawoyski, A. M. and S. P. Ardoin. To be submitted to *School Psychology Review*.

Abstract

Reading comprehension assessments often include multiple-choice (MC) questions, but some researchers doubt their validity in measuring comprehension (Farr, Pritchard, & Smitten, 1990; Martinez, 1999; Rupp, Ferne, & Choi, 2006). Consequently, new assessments will include more short-answer (SA) items (Polikoff, 2014). The current study contributed to the paucity of research comparing MC and SA items by evaluating the effects of anticipated question format on elementary students' reading behavior. Third- and fourth-grade participants were divided into the MC (n = 43) or the SA condition (n = 44) and expected to answer questions consistent with their group assignment. Eye movements were analyzed across the passage and on areas significant to its meaning. Correlational analyses between eye movements and reading measures as well as multiple linear regressions were conducted. Findings support modification of question format in reading assessments. Implications for test developers, teachers, and eye movement researchers are addressed.

INDEX WORDS: Reading, Elementary students, Eye movements, Reading comprehension, Multiple-choice questions, Short-answer questions, Question format

Introduction

Reading comprehension is a complex process that allows readers to gain meaning from text. For success in a literate society, children must master reading comprehension, as this essential skill allows them to continue learning throughout and well beyond their academic years (National Institute of Child Health and Human Development, [NICHD], 2000). Successful reading comprehension requires orchestration of component skills in areas such as word recognition, fluency, and vocabulary (Rayner, Chace, Slattery, & Ashby, 2006). Students must gain comprehension skills quickly, as performance in higher elementary grades and beyond largely depends on their ability to learn information in all academic subjects through reading comprehension (Hernandez, 2011).

Reading comprehension assessment has potential for identifying students in need of intervention (Nation & Snowling, 1997). Yet, because reading comprehension is not clearly defined or fully understood (NICHD, 2000), the best method for measuring it remains highly debated. There are also concerns that because reading comprehension tests differ in presentation and question format, they may measure different components of the skill (Keenan, Betjemann, & Olson, 2008). One debated area pertains to the formatting of comprehension questions, with the benefits and disadvantages of traditional multiple-choice (MC) questions compared to those of short-answer (SA) formats. Researchers are concerned that different item formats may change students' reading behavior and measure different cognitive processes, including constructs that are extraneous to reading comprehension (Martinez, 1999; Rupp, Ferne, & Choi, 2006).

Question Formats

MC questions. MC items typically include a question, known as a *stem*, as well as a list of three to five answer choices (Haladyna, Downing, & Rodriguez, 2002). The choices include

one correct answer, called a *key*, and a few incorrect choices, called *distractors*. MC testing originated over 100 years ago with Kelly's (1916) development of the Kansas Silent Reading Test, which was devised to measure students' current performance in reading and their progress in skill acquisition. Kelly designed the MC format to ensure that administration and scoring procedures were simple and that the time required for testing and scoring was reasonable. Items were constructed to be objective and test students' ability to gain meaning from the text without influence of vocabulary or prior knowledge (Kelly, 1916). In modern times, MC questions are highly favored for large-scale assessments due to their efficiency; they can include numerous items that assess knowledge across a large range of material (Livingston, 2009). MC tests are also less expensive to administer and score than tests with SA questions (Martinez, 1999).

Despite advantages in testing administration and scoring, it is often suggested that MC questions may not provide the best measurement of students' reading skills. First, MC questions may not necessarily measure the construct of reading comprehension (Martinez, 1999; Rupp et al., 2006). For example, Rupp et al. (2006) suggested that through reading and eliminating answer choices, MC questions evoke problem-solving strategies that are specific to the testing context. This concept was demonstrated when Farr, Pritchard, and Smitten (1990) found that undergraduate students taking an MC reading comprehension test focused on the questions more than the text, and a majority of the strategies they engaged in were test-taking strategies (i.e., looking between the question and the text) as opposed to reading comprehension strategies (i.e., making predictions, thinking about their purpose for reading). Participants also rushed through reading the passage to arrive at the questions and used the questions to facilitate a search through the passage for the correct answer. Daneman and Hannon (2001) argued that replacing comprehension behavior with problem solving remains consistent with the goals of a verbal

assessment because successful problem solving requires high-level reasoning with verbal stimuli. In contrast, Farr et al. questioned the generalizability of the learned problem-solving strategies to actual reading comprehension tasks.

A second major concern about MC questions is the possibility that questions may not have passage dependency, meaning that students may be able to respond correctly to questions without reading the corresponding text (Pearson & Hamm, 2005, pp.24-25). Researchers found that on standardized measures, school-aged children could correctly answer comprehension questions above the chance level without having read corresponding passages (Keenan & Betjemann, 2006; Sparfeldt, Kimmel, Löwenkamp, Steingräber, & Rost, 2012). However, findings from Sparfeldt et al. (2012) suggested that passage qualities may impact the extent to which children can respond to questions correctly without reading, as the fourth-grade participants in their study did not attain above chance levels on narrative fictional passages, but did attain above chance levels on informational fictional passages. Farr et al. (1990) warned that if students are able to answer questions correctly without reading the text, then MC questions may not be a valid method of measuring reading comprehension.

SA questions. SA questions are part of the larger group of test items called constructed responses. In reading comprehension tests, SA questions typically require students to provide a response ranging from several words to a couple of sentences. A major argument in favor of SA questions pertains to their utility in measuring more complex, higher-level thinking than MC questions (Bussis & Chittenden, 1987; Martinez, 1999; Pressley, Ghatala, Woloshyn, & Pirie, 1990). In fact, even when MC questions are written to measure skills beyond rote memory, SA questions are still superior in evaluating higher-level cognitive skills (Martinez, 1999). Additionally, Gay (1980) noted that SA questions are advantageous to MC questions because

they evoke passage recall, rather than MC questions which evoke a combination of recall and recognition. Interestingly, test presentation may impact the extent to which question format evokes different reading behavior on the passage. For example, Ozuru, Best, Bell, Witherspoon, and McNamara (2007) asked undergraduate students to complete MC and SA questions when the corresponding text was and was not available. Findings from Ozuru et al. indicated that when the text was not available, there were significant correlations in participants' performance on MC and SA questions. However, when the text was available, correlations between participants' performance on MC and SA questions were not significant.

The widely-held belief that SA items can sample and evoke higher-level cognitive processes better than MC items is changing current testing practices. For example, more SA items will be included on the Common Core State Standards (CCSS) assessments, due to concerns about the quality of MC assessments and a need for items that can assess the higher-level cognitive skills within the CCSS (Polikoff, 2014). SA items are more appealing than they were in earlier years because of new automated scoring software that can mitigate limitations related to the effort involved in scoring (Educational Testing Service, 2010; Livingston, 2009). Unfortunately, as compared to the literature base on MC questions, far less research exists examining the limitations of SA questions. If test-developers are designing new assessments to include more SA items with the expectation that they will evoke higher-level cognitive skills, research must be conducted to determine if this expectation is valid. Yet, with the majority of studies in educational psychology examining outcomes of reading comprehension (i.e., accuracy scores) rather than the process of reading comprehension, it is difficult to investigate how expectations for different question formats shape an elementary-aged student's reading behavior on a passage.

Eye Movement Research on Reading Comprehension

Eye tracking technology provides a means for examining the process of reading comprehension by allowing for direct and objective observation of students' eye movements during reading. Eye movements are thought to represent attention allocation (Just & Carpenter, 1980) and various processes involved in reading, including comprehension (Rayner, 1998; Rayner et al., 2006). Some eye movement dependent measures are thought to reflect lower-level processes such as decoding and word recognition, whereas other eye movement dependent measures are thought to reflect higher-level processes such as comprehension (e.g., Binder, 2003; Binder & Morris, 1995; Foster, Ardoin, & Binder, in press). By viewing eye movements during reading, researchers can better understand where readers allocate attention during reading and how reading behavior relates to accuracy on different types of comprehension questions. For example, Feng et al. (2012) examined eye movements of undergraduate students as they read and responded to MC questions. All participants were informed that they would read a passage and answer MC questions. Half of the participants were also told that they would need to summarize the passage before answering the MC questions. Findings revealed that when participants were told they would have to write a summary, they were more likely to read the entire passage and they spent less time searching for relevant information. In contrast, readers in the MC condition were more likely to sample the text to search for answers, rather than read the text as a whole. Interestingly, readers in the MC condition spent more time reading unnecessary information than readers in the summary condition. Feng et al. concluded that the readers in the summary condition constructed a better mental model of the passage, which allowed them to answer questions more efficiently. In sum, for adult readers, eye movement research was able to support hypotheses stating that MC questions may not yield the most accurate measurement of

reading comprehension (e.g., Martinez, 1999; Rupp et al., 2006). Additionally, by employing eye tracking procedures to evaluate participants' test-taking behavior, Feng et al. was able to reveal new information about how readers respond to MC questions without disturbing their test-taking behavior.

In one of the few studies examining students' eye movements while they read a passage and responded to questions, Solheim and Uppstad (2011) presented 18 seventh-grade students with a reading comprehension assessment consisting of items in different formats such as MC and SA. Students first read a passage as an eye tracker recorded their eye movements. Then, they answered pencil-and-paper questions. When responding to questions, students were permitted to return to the passage and the eye tracker continued to record their eye movements. Analyses of gross eye movement data (i.e., integrative saccades and gaze duration across regions of the stimulus) indicated that some students engaged in similar reading behavior on the passage but attained different comprehension scores on an item requiring integration of knowledge from a passage and an illustration. Additionally, time spent reading and responding to questions was not related to accuracy. Upon closer examination of reading behavior, Solheim and Uppstad found that participants engaged in different processes while responding to questions that better explained differences in reading comprehension outcomes than merely analyzing attention allocation during reading, such as reading and returning to areas of the text containing answers to the questions. However, Solheim and Uppstad examined fewer eye movement dependent measures than researchers typically analyze, failing to include eye movement measures thought to reflect comprehension processes. Although results may not have fully revealed comprehension processes during question-answering, Solheim and Uppstad were the first to use eye tracking technology to examine question-answering behavior in middle school students.

Another benefit of eye movement research is its ability to determine whether individuals' reading behavior differs on important as compared to unimportant components of the text. Research suggests that eye movements vary across sections of text depending upon the relevance of the text to the reader's goal. For example, when adults were assigned a goal for reading, they made longer fixations on relevant information when it was first encountered, as compared to information that was not relevant to their goal (Kaakinen & Hyönä, 2005). Additionally, readers in Kaakinen and Hyönä (2005) were more likely to reread sentences containing information relevant to their goal, as compared to sentences containing irrelevant information.

Eye movement studies on comprehension with elementary students. Ardoin, Zawoyski, Wagner, Bangs, and Binder (2015) conducted the only study to date that measured students' eye-movements as they read passages and responded to the associated comprehension questions. In this study, second-grade students responded to two MC questions. Interestingly, early reading processes were more effortful for students who answered only one question correctly, as compared with students who answered both questions correctly. This result provides preliminary evidence that eye movement measures may predict reading comprehension performance in elementary students.

Although not a study on passage comprehension, Vorstius, Radach, Mayer, and Lonigan (2013) found interesting patterns in elementary students' eye movements as they read single sentences. The sentences' polarities and clauses were manipulated, and students were asked to determine if sentences were normal or abnormal. Results were similar to those found in comprehension studies involving adults (e.g., Kaakinen & Hyönä, 2005; Rayner et al., 2006), for example, students fixated longer on words that were relevant to their goal for reading. Students were also more likely to reread challenging sentences and their accuracy improved when they

reread relevant areas of the sentences. These findings have important implications for the understanding of reading comprehension processes. For instance, researchers may be able to examine how thoroughly students are reading text by examining whether or not they spend more time reading areas of the text that are central to its meaning.

Working Memory

An important component of reading comprehension that is often overlooked in standardized reading assessment and question formatting is working memory. Working memory is a complex system involved in the maintenance, storage, and processing of information (Baddeley & Hitch, 1974; Baddeley, 2003). Evidence suggests that working memory is distinctly related to reading comprehension (Andreassen & Bråten, 2010; Cain, Oakhill, & Bryant, 2004; Carretti, Borella, Cornoldi, & De Beni, 2009; Just & Carpenter, 1980). When working memory is poor, reading comprehension suffers because there is a diminished capacity for processing new information while integrating it with information that was previously read but still active (Daneman & Hannon, 2001). Even when equivalent in terms of age, vocabulary skills, and word reading skills, students who have difficulty with reading comprehension exhibit distinct deficiencies in working memory (Cain, 2006).

In a study involving fifth-grade Norwegian students, Andreassen and Bråten (2010) found that working memory significantly and distinctly predicted reading comprehension outcomes on MC tests, specifically when passages were longer and unavailable as students responded to questions. Andreassen and Bråten hypothesized that when the passage was unavailable, students needed to rely on working memory to access high-quality representations of the passage in order to successfully answer questions. This hypothesis is consistent with findings from Ozuru et al. (2007), indicating that undergraduate students' responding to MC and

SA questions was similar if the corresponding passage was unavailable when they responded to questions.

Limitations to Research on Reading Comprehension Question Format

A majority of the research on reading comprehension and/or question format has been conducted with skilled adults (e.g., Daneman & Hannon, 2001; Farr et al., 1990; Feng et al., 2012; Kaakinen & Hyönä, 2005; Just & Carpenter, 1980; Ozuru et al., 2007; Rupp et al., 2006) or older students (e.g., Solheim & Uppstad, 2011) rather than with elementary-aged students, who make up a significant portion of the target population for reading comprehension testing. Conducting research directly with the elementary population is important because reading behavior of students learning to read is different from that of skilled adult readers (Blythe & Joseph, 2011). Therefore, the impact of question format may differ between adults or older students and elementary-aged students. Findings from the aforementioned studies are also limited by the use of outcome measures (e.g., Andreassen & Bråten, 2010; Gay, 1980; Ozuru et al., 2007) or subjective methods such as think-aloud protocols or cognitive interviews (e.g., Farr et al., 1990; Rupp et al., 2006) that may influence typical reading behavior (Cordón & Day, 1996). Although Ardoin et al. (2015) provided an interesting analysis of elementary students' eye movements during a comprehension assessment, findings are limited by several factors, such as the low number of questions, the quality of distracter items, and experimental procedures that deviate from typical standardized assessments (i.e., students responded to questions by holding up their fingers). Ultimately, the literature base lacks studies involving direct observation and quantitative analysis of elementary-aged students' reading behavior *during* standardized testing.

Summary and Purpose

In sum, comprehension assessments traditionally include MC questions, which are beneficial because they are objective and easy to score, but disadvantageous because they may measure problem-solving strategies rather than reading comprehension (Martinez, 1999; Rupp et al., 2006). SA questions are considered a viable alternative to MC questions. Proponents of SA questions suggest that this question format allows for better assessment of complex cognitive abilities and a more realistic assessment of reading comprehension (Bussis & Chittenden, 1987; Martinez, 1999; Pressley et al., 1990). Despite the increased use of SA questions within newer standardized assessments (Polikoff, 2014) few studies exist evaluating the impact of question format on elementary students' reading behavior.

The purpose of the current study was to evaluate the impact of two different question formats on third- and fourth-grade students' eye movements during reading. Participants were divided into either the *MC condition* or the *SA condition* and received instructions indicating that they would answer either MC questions or SA questions, depending upon their assigned condition. Then, participants read a passage as their eye movements were recorded and answered both question types on pencil and paper.

Research Questions and Hypotheses

The current study was designed to examine the following research questions:

1. *Do students' eye movements during reading differ as a function of condition assignment (MC vs. SA conditions), both across the passage and on meaningful regions of the text?*

It was hypothesized that participants' reading behavior on the passage would differ depending upon whether they were assigned to the MC or SA condition. Participants in the SA

condition were expected to engage in deeper processing of the text than participants in the MC condition and exhibit reading patterns consistent with those of students in Feng et al. (2012). Specifically, participants in the SA condition were expected to have longer overall reading times than participants in the MC condition. Additionally, participants in the SA condition were expected to engage in more reading behavior indicative of higher-level processing on the passage than participants in the MC condition. That is, they would have longer total fixation times, make more regressions, and have a higher average fixation count than students in the MC condition. This pattern of reading behavior was also expected to occur on meaningful regions of the text.

2. *Does students' accuracy on MC questions differ depending upon whether students believed they would have to answer MC or SA questions?*

Participants in the SA condition were expected to read text more closely and obtain a better understanding of the text than participants in the MC condition, based on behavior of participants in Feng et al. (2012). Consequently, it was expected that participants in the SA condition would be more likely to respond accurately on MC questions than participants in the MC condition.

3. How do eye movement measures correlate with standardized reading assessment scores, oral reading fluency, working memory, and reading outcome measures from the current study (i.e., MC and SA questions)?

Findings from previous eye movement studies (e.g., Binder, 2003; Binder & Morris, 1995; Foster et al., in press) suggest that eye movement measures are representative of different processes in reading (e.g., lower level and higher level). The current study aimed to extend findings from Foster et al. (in press) and evaluate the extent to which eye movement measures correlate with standardized measures of reading. Further evidence that eye movement measures

are related to children's reading outcome measures would strengthen findings from the current study as well as numerous studies evaluating elementary students' eye movements during reading.

4. *What is the relation between participants' eye movements during reading and their accuracy on MC questions? How does condition impact accuracy?*

Ardoin et al. (2015) found that eye movement measures thought to reflect lower-level processing were longer for participants who only answered one comprehension question correctly, as opposed to two. Similarly, findings from the current study were expected to show that eye movement measures are related to accuracy on MC comprehension questions, although analyses focused on examining measures expected to reflect discourse-level reading processes (i.e., total fixation time, average fixation count, and number of inter-word regressions). Findings were expected to show a positive correlation between eye movement measures thought to represent higher-level reading processes and MC question scores, which would suggest that readers are differentially allocating their cognitive resources to higher-level tasks. Given that participants in the SA condition were expected to read text more thoroughly than participants in the MC condition, condition assignment is also expected to predict higher accuracy on MC questions.

Method

Participants and Setting

Participants were selected from two public elementary schools within a combined suburban and rural school district in the southeastern United States. At the time of the study, the district served over 7,000 students, 20.52% of whom were eligible for free or reduced price lunches. Ethnicities represented in this school district were: white (79.74%), Hispanic (7.33%),

Asian (5.12%), Black (4.80%), two or more races (3.11%), American Indian (<1.00%), and Pacific Islander (<1.00%). Students in the school district were typically high-performing in English/Language Arts, as evidenced by results from state-wide standardized testing in the 2015-2016 school year indicating that 66.50% of third graders and 65.80% of fourth graders earned "proficient" or "distinguished" scores.

Parents or guardians of participants provided consent for their children to participate in this study by returning a consent form approved by the author's affiliated university and sent home in students' weekly folders. Additionally, parents gave consent for the schools to provide examiners with their children's most recent scores from the Measures of Academic Progress (MAP; Northwest Evaluation Association, [NWEA], 2009) Reading test. Parents also provided consent for their children's de-identified data to be published and presented at professional events. Students provided assent for participation.

A power analysis was conducted to determine the necessary sample size to conduct *t*-tests for the proposed study. The sample size of 56 was selected to allow for identification of large effects (Cohen's $d = .8$, $\alpha = .05$, $\beta = .9$). Data were collected with more than 56 students because some participants may not yield usable data due to technical problems with the eye tracking system.

The sample was comprised of 87 third- and fourth-grade students (52.87% third graders and 47.13% fourth graders). Participants were typically-developing and native speakers of English. Reported ethnicities were: white (81.61%), Asian (8.05%), Black (4.60%), Hispanic (1.15%), and multi-racial (4.60%). The mean age of the sample was 9 years, 1 month (range = 8 years, 1 month to 10 years, 5 months). All study activities took place at the participating schools.

Participants were randomly assigned to either the MC condition or the SA condition (described below). The MC condition was comprised of 43 students; 55.81% third graders and 44.19% fourth graders. Female students made up 44.19% of the MC condition. The average age of participants in the MC condition was 9 years, 1 month (range = 8 years, 2 months to 10 years, 5 months). Ethnicities represented in the MC condition were white (90.70%), Black (4.65%), and Asian (4.65%). The SA condition included 44 students evenly distributed across third and fourth grade. Participants were 40.90% male and the average age was 9 years, 2 months (range = 8 years, 1 month to 10 years, 1 month). The SA condition included students identifying as white (72.72%), Asian (11.36%), multi-racial (9.10%), Black (4.54%) and Hispanic (2.27%).

Apparatus

Participants viewed the passage on a 19 in. or a 22 in. ViewSonic LCD display monitor. Eye movement data were collected via the desktop-mounted camera of the SR Research EyeLink 1000 system. The camera has an accuracy range of 0.25° to 0.5° and a resolution of 0.01° of visual angle. It was positioned between the participant and the computer monitor at a distance of 50-55 cm from the participant and approximately 10-30 cm in front of the monitor. Participants placed their heads on a chin rest to reduce head movement during tracking. The camera recorded eye movements from one eye, but participants' view was binocular throughout the assessment. Participants communicated that they had completed each task by clicking stimuli presented on the monitor using a standard USB mouse.

Measures

MAP Reading. MAP Reading is a computer adaptive test that measures students' skills across areas of reading including reading comprehension. In the participating schools, MAP Reading is administered to all students three times across the academic year. The MAP yields

Rausch Unit (RIT) scores and can be compared across grade levels. Test-retest reliability coefficients from fall to spring testing were .870 for grade 3 and .900 for grade 4, and the marginal reliability coefficients were excellent, ranging from .940 to .950 across fall and spring administrations for both grades (NWEA, 2004). In addition to good internal validity, the concurrent validity between MAP Reading and other group-administered achievement measures is strong. For example, the concurrent validity coefficient for comparison of the MAP Reading test to the 1999 fall administration of the *Iowa Tests of Basic Skills* was .770 for grade three (NWEA, 2004). MAP testing data indicated that the sample in the current study obtained an average RIT score of 206.22 (*SD* = 16.56), which is consistent with the national average for fourth graders at the end of the school year (*M* = 205.90, *SD* = 14.92; NWEA, 2015). Considering that MAP data presented for participants in the current study were collected at the beginning of the school year, MAP results suggest that the sample was high-achieving.

Eye tracking assessment stimuli. An expository practice passage with three MC questions was selected from the Gates MacGinite Reading Tests, Fourth Edition; Level 3 Form S (GMRT; MacGinitie, MacGinitie, Maria, & Dreyer, 2000). This level is typically administered to eight- and nine-year old children.

The eye tracking assessment included one passage and four MC questions selected from Level 10, Form C of the Iowa Tests of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2007), which falls at a fourth-grade reading level. The passage was 182 words in length and expository in nature. One question was created in order to allow for a wider range in accuracy scores. The experimental passage was presented on one of the monitors described above in black text on a white background in Times New Roman font. The font size was approximately 18-point with 1.5 line spacing. The passage had three paragraphs presented in 13 lines of text with a maximum

of 90 characters on a line. Both types of questions (i.e., MC and SA) were presented on paper and participants provided handwritten responses. The questions were typed in 14-point Times New Roman font, presented in black text.

Identifying meaningful information. Three third-grade teachers and two fourth-grade teachers (two males, three females) reviewed the experimental passage and identified 45 words in the text that they believed were important to the meaning of the story. Teachers were not permitted to view the questions related to the text, but they were given a sample story with meaningful areas underlined as an example. Words in the story with 80% or higher agreement across the five raters were selected as meaningful areas, which were pulled from the global data analyses for a separate evaluation. In total, there were 18 words selected for analysis.

Eye movement dependent measures. The eye-tracker camera recorded participants' eye movements during their silent reading of the ITBS passage. The camera measured basic eye movements known as fixations, saccades, and regressions. A *fixation* during reading occurs when a reader's eyes pause on a word. Fixations permit the reader to extract information from the text. A *saccade* is an eye movement that occurs between fixations. During saccades, vision is briefly suppressed and the reader is unable to extract information from the text. Readers typically make backward saccades, called *regressions*. Regressions often occur when the reader returns to previously read text to gather further information (Just & Carpenter, 1980). This study analyzed five common eye movement dependent measures involving fixations and regressions. The first is called *first fixation duration*, which represents the duration of an initial fixation on a word. *Gaze duration* is calculated as the sum of all fixations made during the initial reading of a word. *Total fixation time* is the total duration of all fixations made on a word. *Number of inter-word regressions* represents the number of regressions occurring between words. *Average*

fixation count per word enumerates the mean amount of fixations made on each word. See Appendix E for a table detailing the eye movement dependent measures analyzed in this study.

Eye movement researchers believe that first fixation duration and gaze duration reflect the initial processes involved in word recognition, known as lower-level, or early processing (Binder, 2003; Binder & Morris, 1995; Foster et al., in press). Total fixation time, number of inter-word regressions, and average fixation count per word are expected to represent higher-level, or discourse-level processing (Binder, 2003; Binder & Morris, 1995; Foster et al., in press).

Working memory assessment. Participants completed the *Wechsler Intelligence Scale for Children – Fifth Edition* (WISC-V) *Digit Span* subtest. The purpose of this assessment was to evaluate whether participants in the MC and SA conditions did not differ significantly in working memory ability. According to the WISC-V Technical and Interpretive Manual (Wechsler, 2014), the Digit Span subtest's internal consistency is .91, providing strong evidence for reliability. Validity for the WISC-V Working Memory Index as a whole is also strong. For example, the correlation coefficient for a similar index on another standardized measure of intelligence, the Kaufman Assessment Battery for Children- Second Edition Sequential Processing scale, is .63. On average, participants' Digit Span scaled score was 10.89 (*SD* = 2.82).

Oral reading assessment. After participants completed the eye tracking assessment and responded to MC and SA questions, they read aloud the experimental eye tracking passage to a trained examiner. The purpose of the oral reading assessment was to examine whether oral reading fluency was consistent across groups. Curriculum-based measurement administration procedures were used, but modified slightly to better represent how participants read the passage on the eye tracker. Specifically, examiners did not provide words to participants after 3 s

hesitations. If participants hesitated on a word for 10 s, examiners asked participants to move on to the next word. Examiners collected data on the number and type of errors participants made, including substitutions, reversals, 10 s hesitations, and omissions. Examiners also recorded the amount of time it took participants to read the entire passage. The number of words read correctly was subtracted from the total number of words and divided by participants' reading time in order to yield a score of words read correctly in a minute (WRCM). Across the sample, participants' average WRCM was 105.89 (*SD* = 29.63, range = 31.40 to 174.10).

Group Contingency

This study incorporated a group contingency in order to encourage participants to put forth their best effort during reading and simulate the pressure of a high-stakes testing setting. The contingency required participating students in each class to collectively attain the highest average number of points on the comprehension questions presented during the eye-tracking assessment. The class in each grade with the highest average received a food-related special event (e.g., pizza party, ice cream party, special breakfast) that was applicable to all students in the class regardless of their participation status. Participants were informed of the group contingency, but they were unaware of their classes' current status at the time of participation in the study.

Procedure

Eye tracking assessment. Each participant completed an individualized eye tracking assessment. The assessment took place in an empty classroom or conference room at each participating school. Two examiners trained in eye tracking procedures conducted the assessment. One examiner operated the eye tracker and the other examiner provided instructions to participants. After obtaining assent for participation, examiners informed participants of the

group contingency. Participants were encouraged to read thoroughly and answer questions accurately because their correct responses contributed towards their class's overall score. Examiners explained that if a class achieved the highest average score in their grade, all students in that class would earn a food-related special event.

To begin, participants were seated in a chair across from the display monitor so that their gaze and the camera position formed the ideal angle for tracking. The height of the chair was adjusted as needed so that participants' forward eye gaze naturally fell within the top quadrant of the monitor screen. Once optimal tracking adjustments were made, examiners described the eye-tracking process and guided participants through completing a practice trial detailing procedures for moving between screens with the USB mouse and answering questions. The type of questions presented in each practice trial depended upon group assignment, which allowed participants to practice reading a passage from the computer screen and answering the type of question they expected to receive. Specifically, participants in the MC condition first read the practice passage from the computer screen and completed three MC questions presented on paper. They were not permitted to access the passage when responding. Participants in the SA condition followed the same procedures as those in the MC condition, with the exception that they provided handwritten responses to three practice SA questions presented on paper.

After the practice trial, examiners conducted the eye-tracker calibration and validation process using a nine-point grid. This process verifies that the system tracks eye movements correctly by locating and validating the eye's position in nine areas of the computer screen. After validation, examiners presented a fixation dot in the area of the screen that the text was presented in to ensure that initial eye placement was consistent across participants. Eye movement data were collected as participants read the passage.

MC condition. After the calibration process, examiners informed participants in the MC condition that they would read a passage silently and answer MC questions without being able to refer to the passage when responding to questions. Then, the passage was presented on the screen. When participants were finished reading, they clicked on a "next" arrow at the bottom of the screen which resulted in a blank screen being presented. The examiner then presented participants with the five MC questions and participants were asked to respond to the questions by circling the answer choice that best answered the question. Finally, participants were presented with a second piece of paper containing the five SA questions and they were asked to provide handwritten responses to the questions.

SA condition. Examiners informed participants in the SA condition that they would read a passage silently and then they would respond to SA questions without access to the passage when responding. Despite the directions they received, presentation of experimental stimuli was identical to that of the MC condition. Specifically, participants in the SA condition first read the passage and then they responded to the five MC questions presented on paper by circling their answer choices. Then, participants provided handwritten responses to the five SA questions presented on paper.

Working memory and oral reading assessments. After the eye tracking assessment, participants completed the WISC-V Digit Span subtest and the Oral Reading Assessment (in no particular order). Both assessments were conducted by undergraduate or graduate students trained in standardized assessment procedures and the aforementioned modified CBM procedures. When participation was complete, participants selected a small prize. Participation time was approximately 35-45 min. Breaks were provided when necessary.

Data Preparation

After eye tracking, participants' data files were prepared for analysis with the assistance of individuals who were trained in eye movement data preparation procedures. Data preparation procedures for the current study were consistent with those employed in previous studies (e.g., Ardoin, Binder, Foster, & Zawoyski, 2016; Foster, Ardoin, & Binder, 2013; Zawoyski, Ardoin, & Binder, 2015). Only eye movement data collected on the stimulus passage were analyzed. First, fixations shorter than 120 ms and longer than 800 ms were removed, as fixations outside of this range are not considered to represent reading behavior in children. Some fixations were trimmed if they occurred off-screen to the left and right of the passage. The remaining fixations were drift corrected, meaning that they were individually moved to the area of the passage that most likely reflected where the participant was reading. This process adjusts for participants' head movement and possible poor calibration of the eye tracking system.

Participants' eye movement data files were individually analyzed using Data Viewer software, which was used to provide reports for all relevant variables, both across the passage (i.e., first fixation duration, gaze duration, total fixation time, average fixation count, number of inter-word regressions) and on meaningful areas of the text (i.e., total fixation time, average fixation count, number of inter-word regressions).

Results

Prior to analysis, data were examined for outliers using the outlier labeling rule (Hoaglin & Iglewicz, 1987; how2stats, 2011a, 2011b, 2011c). In this procedure, the difference between the third and first quartile values in the data set are multiplied by a constant (i.e., 2.2). The obtained value is then subtracted from the first quartile value and added to the third quartile value in order to achieve a lower bound and an upper bound, respectively. Values falling outside

of this range are considered outliers. Applying this labeling rule to the global dependent measures of total fixation time and average fixation resulted in the identification of one significant outlier each. Meaningful area dependent measures of total fixation time and average fixation count had two significant outliers each. Importantly, all of the outliers were found within the SA condition and violated the upper bound of the acceptable range. This result is consistent with hypotheses that the SA condition would make more fixations and longer fixations than the MC condition, particularly on meaningful areas of the text. In order to include these extreme values and meet normality requirements for statistical analyses, the outliers were winsorized (how2stats, 2011d, 2011e). Specifically, the values were replaced with the value of the upper bound determined by the outlier labeling procedure. For global measures of total fixation time and average fixation count, 1.23% of the data were winsorized. For meaningful area measures of total fixation time and average fixation count, 2.47% of the data were winsorized. After winsorization, data for all dependent measures fell in a normal distribution, with skewness values between -2 and 2 (i.e., -.242 to 1.29) and kurtosis values between -2 and 2 (i.e., -.562 to 1.72). Levene's Test for Equality of Variances indicated homogeneity of variance across measures with the exception of the global eye movement data for average fixation count ($F = 5.654$, $p = .02$). For this measure, a *t*-test allowing for unequal variances was conducted.

Prior to analyses, 6 of the 87 participants' data were excluded. Three participants' data were excluded due to either their reading achievement or working memory scores falling above or below two standard deviations from the grand mean of the sample (i.e., two participants were excluded for MAP scores outside the acceptable range and one participant was excluded for a working memory score that was outside the acceptable range). An additional two participants' data were excluded due to technical problems with the eye tracker and one participant failed to

complete study requirements. The final sample included 39 participants in the MC condition and 42 participants in the SA condition.

A post hoc power analysis indicated a high amount of power for detecting large effects in a one-tailed *t*-test (Cohen's $d = .8$, $\alpha = .05$, $\beta = .972$). Additionally, power to detect large effects was greater than .8 at the $\alpha = .05$ level for two-tailed *t*-tests and multiple linear regression analyses.

Two-tailed independent samples *t*-tests were conducted to ensure that participants in the MC and SA conditions were not significantly different in their reading achievement, working memory, and oral reading fluency. Results indicated no significant differences between the MC and SA conditions in their MAP reading RIT scores (MC condition: $M = 208.85$, $SD = 12.69$; SA condition: $M = 204.98$, $SD = 15.00$; $t(79) = 1.25$, $p = .215$), or their WISC-V Digit Span scaled scores (MC condition: $M = 11.03$, $SD = 3.00$; SA condition: $M = 10.67$, $SD = 2.43$; $t(79) = .594$, $p = .554$). Additionally, participants in the MC condition ($M = 105.77$, $SD = 27.74$) and the SA condition ($M = 106.16$, $SD = 30.83$) did not differ significantly in their oral reading fluency, $t(79) = -.060$, $p = .952$, measured in words read correctly per minute (WRCM).

Global Eye Movement Measures

In order to examine whether participants in the SA condition would engage in greater levels of reading behavior indicative of higher-level processing, one-tailed independent samples *t*-tests were conducted for first fixation duration, gaze duration, total fixation time, number of inter-word regressions, and average fixation count. Results indicated that MC and SA conditions did not differ in measures of first fixation duration, $t(79) = -.874$, $p = .193$, gaze duration, $t(79) = -1.22$, $p = .113$, or number of inter-word regressions $t(79) = -.628$, $p = .266$. However, the SA condition's total fixation time was significantly longer than that of the MC condition, $t(79) = .-$

2.11, $p = .019$. Additionally, the SA condition had a higher average fixation count across the text as compared to the MC condition, $t(71.71) = -2.016$, $p = .024$. (See Table 3.1 for descriptive statistics).

Analysis of Eye Movements on Meaningful Areas

Eye movement data were averaged across the 18 meaningful areas (i.e., words). Then, one-tailed *t*-tests were conducted to compare the MC and SA conditions' total fixation time, number of inter-word regressions, and average fixation count on the meaningful areas. Results indicated no significant between-groups differences for number of inter-word regressions, $t(79) = -1.02$, $p = .156$, and average fixation count $t(79) = -1.21$, $p = .115$. Analysis of total fixation time revealed that the SA condition spent significantly more time fixating on meaningful areas than the MC condition $t(79) = -1.81$, $p = .037$. (See Table 3.1 for descriptive statistics).

Performance on MC and SA Questions

MC questions. Immediately after reading the eye tracker passage, all participants responded to five MC questions. The questions were scored to yield the number of questions answered correctly for each participant. A one-tailed *t*-test revealed that the MC condition ($M = 2.72$, $SD = 1.30$) and the SA condition ($M = 2.64$, $SD = 1.08$), did not differ significantly in their performance on MC questions, $t(79) = .284$, $p = .389$. Further analyses were conducted on MC questions to assess for potential between-groups differences on individual questions. A chi-squared test of independence was conducted for each question to determine whether group assignment impacted correct responding on MC questions. Results indicated no significant differences between the MC and SA conditions in their performance on individual MC questions (See Table 3.2).

To examine the extent to which eye movement measures predicted accuracy on MC questions, multiple linear regression analyses were conducted. Originally, the purpose of this analysis was also to examine the predictive power of condition assignment on MC question accuracy, but condition was removed from the analysis because a *t*-test revealed no significant between-groups difference on MC question performance. First, a multiple linear regression was conducted using only first fixation duration and gaze duration as predictors for MC question accuracy. Together, these variables explained a significant portion of the variance in MC question accuracy, Adjusted $R^2 = .233$, $F(2, 78) = 11.84$, $p < .001$. When variables expected to reflect reading comprehension were added to the model (i.e., total fixation time, number of inter-word regressions, and average fixation count), a greater portion of variance in MC question accuracy was explained above and beyond that of first fixation duration and gaze duration alone, Adjusted $R^2 = .373$, $F(5, 75) = 8.92$, $p < .001$. Further analysis revealed that significant predictors in the second model were total fixation time and number of inter-word regressions. Total fixation time was negatively correlated with accuracy and number of inter-word regressions was positively correlated with accuracy. That is, across groups, for every 1 ms increase in total fixation time, MC question accuracy decreases by .007 of a point. For number of inter-word regressions, as the number of regressions increases by one, MC accuracy increases by 5.55 points. (See Table 3.3 for significance values of coefficients in the model).

SA questions. SA questions were scored after the assessment. Participants earned between zero and two points for each of the five questions. Questions were rated by two independent raters trained in scoring procedures using an experimenter-developed rubric. Scores from each rater were averaged to determine participants' final score, out of 10 possible points. Findings revealed that the MC condition ($M = 3.06$, $SD = 1.68$) and the SA condition ($M = 3.18$,

SD = 1.95) did not differ significantly in their performance on SA questions, $t(79) = -.282$, $p = .390$.

Correlational Analyses

Correlational analyses were conducted to examine correlations among reading/standardized measures (i.e., MAP scores, WISC-V Digit Span scores, oral reading fluency), outcome measures (i.e., MC and SA questions), and eye movement measures (See Table 3.4). Findings revealed significant negative correlations ($p < .001$) between MAP reading scores and three of the global eye movement measures (i.e., first fixation duration, gaze duration, total fixation time) as well as one of the eye movement measures on meaningful areas (i.e., total fixation time). A significant negative correlation ($p = .039$) was also observed between MAP reading scores and average fixation count. Working memory scores correlated negatively and significantly ($p < .001$) with measures of first fixation duration and gaze duration. As with correlations found with MAP reading scores, there were significant negative correlations ($p < .001$) between oral reading fluency scores and three of the global eye movement measures (i.e., first fixation duration, gaze duration, total fixation time) as well as total fixation time on meaningful areas. In addition, a significant negative correlation ($p = .044$) was found between oral reading fluency and average fixation count.

Correlational analyses conducted with outcome measures revealed that both the number of MC questions answered correctly and the number of points scored on SA questions correlated positively ($p < .001$) with MAP and oral reading fluency scores as well as working memory scores ($p = .033$ for MC questions and $p = .011$ for SA questions). With respect to eye movement measures, significant negative correlations were found between the number of MC questions answered correctly and global measures of first fixation duration, gaze duration, and

total fixation time (all, $p < .001$) as well as the meaningful area analysis of total fixation time ($p = .027$). Also significant were positive correlations between the number of MC questions answered correctly and both the global and meaningful area analyses for number of inter-word regressions ($p = .018$ and $p = .047$, respectively.)

Procedural Integrity and Interobserver Agreement (IOA)

Eye tracking sessions. Procedural integrity checks were conducted during eye-tracking sessions to ensure that the examiner who was administering instructions followed procedures correctly. These checks were conducted by the second examiner (i.e., the one operating the eye-tracker) during a randomly selected 38.27% of eye tracking sessions. Data were collected using a six-item checklist allotting one point to the examiner for: providing initial directions, describing the group contingency, giving condition-specific instructions for the practice passage, presenting condition-specific instructions and stimuli for the practice questions, providing condition-specific instructions for the experimental passage, and administering multiple-choice questions after reading. Initially, providing participants with feedback about their performance on the was a procedural integrity item, but this step was removed from the protocol soon after the start of the study when it was apparent that many students were performing poorly on the questions. Overall, the mean procedural integrity for eye tracking procedures was 100%.

Eye movement data. IOA was calculated for the drift correction process of eye movement data preparation on a randomly selected 35.80% of the sample using the mean-count-per-interval method for fixation count. Specifically, the smallest number of fixations in each interest area was divided by the largest number of fixations in each interest area and multiplied by 100. Then, an average was calculated across all interest areas in the passage. Mean IOA was

98.20% (range = 87.00%-100%). With the exception of one file with IOA at 87.00%, scores for all other examined files were above 95.00%.

Working memory. For the administration of the WISC-V Digit Span subtest, a second examiner listened to 37.04% of files to conduct procedural integrity and inter-observer agreement checks. The procedural integrity checklist was divided into 73 steps involving instruction delivery and item administration. The mean procedural integrity score was 99.57% (range = 97.26%-100.00%). Inter-observer agreement was determined by having the second examiner score the session recording and calculate the percentage of items that were scored consistently across both examiners (i.e., both raters indicated that an item deserved one or zero points). The mean inter-observer agreement score was 99.52% (range = 96.49%-100.00%).

Oral reading. For the oral reading fluency assessment, a five-step procedural integrity checklist was used to ensure that examiners implemented the assessment correctly. The mean procedural integrity score across 35.80% of the data was 97.93% (range = 80.00%-100.00%). Only three of the examined files earned a procedural integrity rating of 80.00%, and in all three cases, it was because examiners started their timers a few seconds before students said the first word of the passage, rather than precisely when students said the first word of the passage. The mean IOA for the oral reading assessment across the selected files was 99.34% (range = 95.60%-100.00%). IOA was calculated by obtaining the difference between the highest number of errors and the lowest number of errors scored by both observers for each selected file. This difference was subtracted from the total number of words in the passage and then divided by the total number of words in the passage. A percentage of agreement was obtained from this value.

SA question scoring. Reliability on SA questions was calculated by first finding the absolute difference between the two raters' final scores for each participant, out of 10 total

points. The number of participants with an absolute difference score greater than one was subtracted from and then divided by the total number of participants to obtain a percentage of agreement within one point. There was a two-point margin of disagreement for nine participants and a three-point margin of disagreement for two participants. Thus, the initial percentage of agreement within one point was 86.42%. Following the initial reliability calculation, data for individual items, which could be awarded zero, one, or two points, were reviewed. Raters were asked to reach a consensus about scores for six individual items with a two-point margin of disagreement. After consensus scoring, agreement across all items was recalculated. The final percentage of agreement within one point was 91.36%. Seven participants remained who had a two-point margin of disagreement; there were no participants with a three-point margin of disagreement or greater. SA question scores obtained after consensus were used in analyses.

Discussion

MC questions are frequently used as a means of quickly, easily, and inexpensively measuring students' reading comprehension skills. There are, however, concerns associated with the use of MC questions as it is possible for students to employ skills other than their reading comprehension skills to correctly answer the MC questions associated with a text. Given these concerns as well as the desire to measure deeper levels of comprehension, many tests are now employing SA questions. Unfortunately, to date there exists only one study (i.e., Feng et al., 2012) examining how question format might impact students' reading behavior. That study was conducted with college students whose reading skills and behaviors differ from that of elementary students who are still learning to read. The current study was developed to address this issue by measuring students' eye movements while they read a passage and expected to answer either MC or SA questions.

Global eye movement analyses revealed no differences between the MC and SA conditions in measures of first fixation duration and gaze duration. Given that these measures are thought to represent lower-level processing and a reader's initial word reading, these findings are as expected because students' reading skills as measured by the MAP and oral reading fluency data were consistent across conditions. There were also no differences between conditions in the number of regressions to previously read words. This lack of difference can possibly be explained by several factors. First, although regressions are thought to represent times in which readers return to re-read previously read text, regressions can also occur when readers are correcting for overshooting eye movements (Rayner et al., 2006). Second, largely due to difficulties triggering and manipulating them, regressions are poorly understood in comparison to other eye movements (Rayner, 1998). Furthermore, the passage may not have been difficult enough to evoke numerous regressions in the high-performing third- and fourth-grade readers who participated in the study.

Despite no differences in lower-level processing and regressions, as hypothesized based upon prior research (e.g., Feng et al., 2012), analyses of eye movements on the experimental passage revealed that the SA condition had longer total fixation times as well as higher average fixation counts on words as compared to participants in the MC condition. Likewise, in the areas identified as contributing the greatest amount of meaning to the text, students in the SA condition had longer total fixation times than students in the MC condition. The between-groups differences on total fixation time and average fixation count suggest that similar to college students in Feng et al. (2012), when the elementary students expected to complete a more challenging comprehension task, they made greater efforts to develop a more comprehensive understanding of the passage. This outcome is likely due to participants' prior experience with

responding to SA questions; SA questions may have acquired stimulus control for more effortful reading behavior. In comparison, MC questions require less effortful reading because participants must only recognize a correct choice from the provided options.

Despite findings from eye movement analyses indicating that participants in the SA condition engaged in higher levels of eye movements related to discourse-processing on the text, there was no difference between conditions in their performance on either the MC or SA questions. For both types of questions, requiring students to respond without access to the text may have impacted their performance by making the task more challenging. Additionally, it is possible that the SA questions were too difficult for students, particularly given that students had to answer the questions without access to the text, as on average, students earned only 3 out of 10 points. Unfortunately, these questions were developed for the purpose of the current study and thus had not been validated

In addition to examining students' response accuracy, data were collected on the amount of time participants needed to complete the MC and SA questions. Descriptive data for 42 participants indicated that the average time to complete SA questions (M = 8 min 13 s; SD = 4 min 47 s) was nearly five times longer than their time to complete MC questions (M = 1 min 33 s, SD = 33 s). Although not one of the original research questions for this study, the amount of time students needed to complete SA questions was another important finding, because it exposes a significant limitation to SA questions that was not discussed in extant literature.

Correlational analysis. Correlational analyses supported prior research regarding the relationship between eye movement variables and reading outcomes (i.e., Foster et al., in press) and provided interesting implications regarding the variables and participants in the current study. First, findings indicating a significant relationship between MAP reading scores and

students' accuracy on the MC and SA comprehension questions provides some evidence that the questions used in the current study were associated with reading comprehension. Second, and perhaps most importantly, findings indicating significant relationships between eye movement measures and MAP scores, oral reading fluency scores, and working memory scores provide evidence for the association between elementary students' eye movements during reading and their performance on reading comprehension outcome measures. Particularly, MAP reading scores and oral reading scores were significantly and negatively correlated with first fixation duration, gaze duration, total fixation time, and average fixation count, as well as total fixation time on meaningful areas of the text. More specifically, greater efficiency in eye movements was correlated with higher reading scores, which is consistent with hypotheses made in the current study. A third important finding from the correlational analyses was the significant negative correlation between working memory and eye movement measures of first fixation duration and gaze duration. Although further research is necessary to investigate and confirm this relationship, preliminary findings suggest that greater efficiency in initial word processing may be associated with better working memory scores.

Multiple linear regressions. Multiple linear regressions were conducted to identify eye movement variables that significantly predicted variance in students' MC question response accuracy. It was hypothesized that higher-level processing eye movement measures would be significant predictors of MC question accuracy**.** The first model included only first fixation duration and gaze duration and indicated that those variables significantly predicted the variance in MC question accuracy. Adding total fixation time, number of inter-word regressions, and average fixation count to the model explained a greater portion of the variance in MC question accuracy. Total fixation time and number of inter-word regressions were the only significant

predictors in this model. As expected, number of inter-word regressions positively correlated with MC question accuracy. The negative correlation between total fixation time and accuracy on MC questions was unexpected, but likely suggests that overall efficiency in reading is related to better reading comprehension. Taken together, findings from the multiple regression analyses suggest that spending less time on words but returning to them more frequently is significantly associated with improved responding on MC questions.

Limitations and Future Directions

Results of this study should be evaluated in the context of several limitations. First, implications are based on findings from a limited amount of stimuli; participants read only one passage and answered only five MC and SA questions. Future studies should investigate how expectation of question type impacts eye movements during reading across multiple passages, perhaps in a full standardized reading assessment with a greater number of questions.

A second limitation to this study is the potential impact of participants' inability to access the text when responding for MC and SA questions. Results indicated a lack of differences in performance between groups on the comprehension questions. Across groups, participants struggled to answer SA questions correctly, earning a mean of approximately 3 points out of 10, which may have been higher if students were able to access text during responding. For example, Solheim and Uppstad (2011) found that some readers were able to gather all the necessary information from the text on their first reading, whereas other readers required additional readings and text referencing in order to answer questions correctly. Therefore, methods in the current study may not have allowed readers who are accustomed to referencing the text while responding to achieve typical scores. Additionally, according to Ozuru et al. (2007), skilled readers obtained similar scores on MC and SA questions when they were not able

to access the text during responding. For the elementary students in the current study, knowing that they had to answer questions without text may have encouraged students in the MC condition to attend to the text more closely than usual.

Participants' struggle to answer comprehension questions after one reading of the text illuminates an important concern with the structure of reading comprehension assessments. That is, reading comprehension assessments may not adequately evaluate students' proficiency in practical applications of reading comprehension. Although reading comprehension assessments typically contain mostly MC questions and allow test-takers to reference the text while responding, this practice is not always consistent with reading comprehension requirements in daily life. For example, an adult may be expected to read a manual detailing job duties and then recall them while performing the job, without being able to reference the manual. Similar concerns have been raised in the literature for decades; Farr et al. (1990) also questioned the generalizability of reading comprehension assessment (specifically MC questions) and its relevance to the range of reading tasks individuals may be asked to complete. Despite concerns, reading comprehension assessments have not changed. Unfortunately, it appears that reading assessments do not encourage students to prioritize comprehension of the passage from a single reading because they have learned to expect access to the text when responding to questions. Findings from the current study support encouraging test developers to create more practical reading assessments. Additionally, future reading instruction should be aimed at teaching students how to gather and retain important information from the text during their first reading.

Another future direction based on findings from this study is related to the appropriate usage of SA questions. Between-groups comparisons revealed no significant differences in performance on SA questions. Taken into consideration with the amount of time students

needed to answer these questions, SA questions initially appear inefficient. However, with only the expectation of question type as the difference between groups, eye movement analyses revealed that participants in the SA condition engaged in more higher-level reading behavior than participants in the MC condition. Consistent with recommendations from Martinez (1999), findings from the current study support developing assessments that include both MC and SA questions. Future eye movement studies should examine the proper ratio of MC to SA questions that would balance efficiency with an increased likelihood that students would engage in higher-level processing on the text.

Summary

Polikoff (2014) reported that standardized assessments are shifting to include more SA questions, yet few studies had been conducted on whether this shift would actually cause students to engage in higher-level processing on text. Results from the current study provide preliminary evidence to support this shift, making this the first eye movement study to suggest that eye movements indicative of higher-level processing may be evoked by informing students that they will need to answer SA comprehension questions. Consistent with findings from extant research (e.g., Kaakienen & Hyönä, 2005; Rayner et al., 2006), results from the current study indicated that elementary students' reading behavior could be changed by manipulating their goal for reading. This was evident across the passage as a whole and on meaningful areas of the text. In turn, findings also suggest that when expecting only MC questions, students are less likely to engage in higher-level processing.

Unlike eye movement analyses, differences were not observed between groups' performance on MC and SA questions, perhaps because participants could not access text when responding. Despite engaging in more thorough reading, participants in the SA condition could

not outperform participants in the MC condition, suggesting that learned reading strategies were not sufficient in aiding participants to complete a reading task that is more likely to represent real-life demands of reading and expectations of comprehension. Developers of both reading comprehension assessment and reading instruction should view the findings from this study as a call for improvement and strive to shape reading comprehension education so that skills may generalize to tasks outside of the classroom.

References

Andreassen, R. & Bråten, I. (2010). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading, 33,* 263-283. doi: 10.1111/j.1467-9817.2009.01413.x

Ardoin, S. P., Binder, K. S., Foster, T. E., & Zawoyski, A. M. (2016). A randomized control design study examining the effects of repeated readings on reading achievement and reading behavior. *Journal of School Psychology, 59,* 13-38. doi:10.1016/j.jsp.2016.09.002

Ardoin, S. P., Zawoyski, A. M., Wagner, L., Bangs, K., & Binder, K. S. (2015). *Measuring test-taking behavior: Different behaviors but similar outcomes.* Manuscript submitted for publication.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, *4*, 829-839. doi:10.1038/nrn1201

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47-89). San Diego, CA: Academic Press.

Binder, K.S. (2003). Sentential and discourse topic effects on lexical ambiguity processing: An eye movement examination. *Memory and Cognition, 31,* 690-702. doi:10.3758/BF03196108

Binder, K.S., & Morris, R.K. (1995). Eye movements and lexical ambiguity resolution: Effects of prior encounter and discourse topic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1186-1196. doi:10.1037/0278-7393.21.5.1186

Blythe, H. I., & Joseph, H. S. S. L. (2011). Children's eye movements during reading. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 643–662). Oxford, UK: Oxford University Press.

Bussis, A. M., & Chittenden, E. A. (1987). Research currents: What the reading tests neglect. *Language Arts, 64,* 302-308.

Cain, K. (2006). Children's reading comprehension: The role of working memory in normal and impaired development. In S. J. Pickering (Ed.), *Working memory and education* (pp. 61-91). Cambridge, MA: Academic Press.

Cain, K., Oakhill, J. V., & Bryant, P. E. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96,* 31–42. doi:10.1037/0022-0663.96.1.31

Carretti, B., Borella, E., Cornoldi, C., & De Beni, R. (2009). Role of working memory in explaining the performance of individuals with specific reading comprehension difficulties: A meta-analysis. *Learning and Individual Differences, 19,* 246-251. doi:10.1016/j.lindif.2008.10.002

Cordón, L. A. & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology, 88,* 288-295. doi:10.1037/0022-0663.88.2.288

Daneman, M. & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology, 130,* 208-223. doi:10.1037/0096-3445.130.2.208

Educational Testing Service (2010). *ETS automated scoring and NLP technologies.* http://www.ets.org/Media/Home/pdf/AutomatedScoring.pdf

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27,* 209-226. doi:10.1111/j.1745-3984.1990.tb00744.x

Feng, G., Gorin, J., Sabatini, J., O'Reilly, T., Wall, C., & Bruce, K. (2012, July). Reading for understanding: How comprehension facilitates answering questions, and what questions enhance understanding. In G. Feng (Chair), *Higher order literacy skills.* Symposium conducted at the meeting of the Society for the Scientific Study of Reading, Montreal, Canada.

Foster, T. E., Ardoin, S. P., & Binder, K. S. (2013). Underlying changes in repeated reading: An eye movement study. *School Psychology Review, 42,* 140-156.

Foster, T. E., Ardoin, S. P., & Binder, K. S. (in press). Reliability and validity of eye movement measures of children's reading. *Reading Research Quarterly.*

Gay, L. R. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement, 17,* 45-50. doi:10.1111/j.1745-3984.1980.tb00813.x

Haladyna, T M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309-334. doi:10.1207/S15324818AME1503_5

Hernandez, D. J. (2011). *Double Jeopardy: How Third-Grade Reading Skills and Poverty Influence High School Graduation.* Annie E. Casey Foundation.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2007). *Iowa tests of basic skills (ITBS).* Rolling Meadows, IL: Riverside Publishing.

how2stats (2011a, September 8). The right way to detect outliers – Outlier labeling rule (part 1). [Video file]. Retrieved from https://www.youtube.com/watch?v=WSflSmcNRFI

how2stats (2011b, September 8). The right way to detect outliers – Outlier labeling rule (part 2). [Video file]. Retrieved from https://www.youtube.com/watch?v=2HmopqF6V6w&t=80s

how2stats (2011c, September 8). The right way to detect outliers – Outlier labeling rule (part 3). [Video file]. Retrieved from https://www.youtube.com/watch?v=bRdC1u9veg8&t=77s

how2stats (2011d, October 5). Dealing with outliers (part 1). [Video file]. Retrieved from https://www.youtube.com/watch?v=Ukkcer70r5A&t=1s

how2stats (2011e, October 5). Dealing with outliers (part 2). [Video file]. Retrieved from https://www.youtube.com/watch?annotation_id=annotation_204681&feature=iv&src_vid=Ukkcer70r5A&v=FatA5COFIPU

Just, M. A. & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 4,* 329-354. doi:10.1037/0033-295X.87.4.329

Kaakinen, J. K., & Hyöna, J. (2005). Perspective effects on expository text comprehension: Evidence from think-aloud protocols, eye tracking, and recall. *Discourse Processes, 40,* 239-257. doi:10.1207/s15326950dp4003_4

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading, 10,* 363-380.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12,* 281-300. doi:10.1207/s1532799xssr1004_2

Kelly, F. J. (1916). The Kansas silent reading tests. *Journal of Educational Psychology, 7,* 63-80. doi:10.1037/h0073542

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; How we score them. *R & D Connections, 11,* 1-8.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie Reading Tests, Fourth Edition, Level 3 Form S.* Rolling Meadows, IL: The Riverside Publishing Company.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*, 207–218. doi:10.1207/s15326985ep3404_2

Nation, K. & Snowling, M. (1997). Assessing reading difficulties: the validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67,* 359-370. doi:10.1111/j.2044-8279.1997.tb01250.x

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: Reports of the subgroups* (NIH Publication No. 00-4754). Washington, DC: U.S. Government Printing Office.

Northwest Evaluation Association (2004). *Reliability and validity estimates: NWEA achievement level tests and measures of academic progress.* Lake Oswego, OR: Author.

Northwest Evaluation Association. (2009). *Technical manual for Measures of Academic Progress and Measures of Academic Progress for Primary Grades.* Lake Oswego, OR: Author.

Northwest Evaluation Association (2015, August). *NWEA Measures of Academic Progress Normative Data.* Retrieved from: https://www.nwea.org/content/uploads/2015/06/2015-MAP-Normative-Data-AUG15.pdf

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction, 25,* 399-438. doi:10.1080/07370000701632371

Pearson, P. D. & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices- past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 24-25). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Polikoff, M. S. (2014). *Common core state standards assessments: Challenges and opportunities*. Retrieved from: https://cdn.americanprogress.org/wpcontent/uploads/2014/04/CCCAssessments-report.pdf

Pressley, M., Ghatala, E. S., Woloshyn, V., & Pirie, J. (1990). Sometimes adults miss the main ideas and do not realize it: Confidence in responses to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly, 25,* 232-249. doi:10.2307/748004

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422. doi:10.1037/0033-2909.124.3.372

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading, 10,* 241-255. doi:10.1207/s1532799xssr1003_3

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shape the construct: A cognitive processing perspective. *Language Testing, 23*, 441-474. doi:10.1191/0265532206lt337oa

Solheim, O. J. & Uppstad, P. H. (2011). Eye-tracking as a tool in process-oriented reading test validation. *International Electronic Journal of Elementary Education, 4,* 153-168.

Sparfeldt, J. R., Kimmel, R., Löwenkamp, L., Steingräber, A., & Rost, D. H. (2012). Not Read, but Nevertheless Solved? Three Experiments on PIRLS Multiple Choice Reading Comprehension Test Items. *Educational Assessment, 17,* 214-232. doi:10.1080/10627197.2012.735921

Vorstius, C., Radach, R., Mayer, M., & Lonigan, C. (2013). Monitoring local comprehension monitoring in sentence reading. *School Psychology Review, 42,* 191–206.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children – Fifth Edition Technical and Interpretive Manual.* Bloomington, MN: NCS Pearson.

Zawoyski, A. M., Ardoin, S. P., & Binder, K. S. (2015). Using eye tracking to observe differential effects of repeated readings for second-grade students as a function of achievement level. *Reading Research Quarterly, 50,* 171-184. doi:10.1002/rrq.91

Table 3.1

Means and Standard Deviations for Eye Movement Measures in Chapter 3

Dependent Measures	MC Group M *(SD)*	SA Group M *(SD)*
Global Eye Movements		
First fixation duration (ms)	252.69 (44.93)	260.74 (37.86)
Gaze duration (ms)	320.59 (87.91)	343.40 (79.82)
Total fixation time (ms)*	470.65 (128.88)	545.95 (184.49)
Number of inter-word regressions (#)	.341 (.128)	.361 (.147)
Average fixation count (#)*	1.54 (.409)	1.77 (.616)
Meaningful Area Eye Movements		
Total Fixation Time (ms)*	525.03 (152.36)	601.60 (219.93)
Number of Inter-Word Regressions (#)	.297 (.175)	.341 (.208)
Average Fixation Count (#)	1.89 (.619)	2.08 (.801)

*Significant between-groups differences, $p < .05$

Table 3.2

Summary of Between-Groups Differences by Question, Chi-Squared Test of Independence

Question	Correct MC Condition	Correct SA Condition	X^2	df	*p*
Question 1	43.59%	64.29%	3.49	1	.062
Question 2	51.28%	40.48%	.952	1	.329
Question 3	64.10%	50.00%	1.64	1	.200
Question 4	58.97%	61.90%	.073	1	.787
Question 5	53.85%	50.00%	.120	1	.729

*Significant between-groups differences, p <.05

Table 3.3

Multiple Regression of Predictors for MC Question Accuracy

	B	SE B	β	*t*	*p*
Constant	3.51	1.246		2.82	.006
First Fixation Duration (ms)	-.008	.007	-.268	-1.08	.282
Gaze Duration (ms)	.006	.004	.456	1.65	.103
Total Fixation Time (ms)	-.007	.003	-.990	-2.38	.020*
Number of Inter-Word Regressions (#)	5.55	1.36	.646	4.08	.000*
Average Fixation Count (#)	.421	.649	.191	.648	.519

*Significant predictor of MC question accuracy, $p < .05$

Table 3.4

Correlation Matrix between Reading, Working Memory, Eye Movement, and Outcome Measures

	MAP	WM	ORF	MC_Q	SA_Q	FFD	GD	TFT	IWR	AFC	MA_ TFT	MA_ IWR	MA_ AFC
MAP	———												
WM	.436**	———											
ORF	.701**	.460**	———										
MC_Q	.469**	.238*	.328**	———									
SA_Q	.633**	.280*	.573**	.482**	———								
FFD	-.549**	-.335**	-.569**	-.481**	-.451**	———							
GD	-.634**	-.402**	-.693**	-.447**	-.487**	.888**	———						
TFT	-.401**	-.186	-.408**	-.310**	-.226*	.600**	.738**	———					
IWR	.142	.025	.115	.261*	.268*	-.116	-.042	.518**	———				
AFC	-.229*	-.087	-.224*	-.148	-.077	.269*	.486**	.890**	.608**	———			
MA_ TFT	-.353**	-.168	-.308**	-.246*	-.165	.451**	.596**	.929**	.589**	.905**	———		
MA_ IWR	.046	-.104	-.072	.222*	.208	-.129	.003	.316**	.662**	.420**	.466*	———	
MA_ AFC	-.078	-.082	-.086	.002	.078	.022	.238*	.705**	.637**	.890**	.851**	.606**	———

* $p < .05$
** $p < .001$

Note. MAP = Measures of Academic Progress RIT scores, WM = WISC-V Digit Span scaled scores, ORF = oral reading fluency (words read correctly in a minute), MC_Q = number of multiple-choice questions answered correctly (out of 5), SA_Q = number of short-answer question points (out of 10), FFD = first fixation duration (ms), GD = gaze duration (ms), TFT = total fixation time (ms), IWR = number of inter-word regressions, AFC = average fixation count, MA_TFT = total fixation time on meaningful areas (ms), MA_IWR = number of inter-word regressions on meaningful areas, MA_AFC = average fixation count on meaningful area

APPENDIX E

EYE MOVEMENT DEPENDENT MEASURES IN CHAPTER 3

Measure	Definition	Processing Type (Binder, 2003; Binder & Morris, 1995; Foster et al., in press)
First Fixation Duration	duration (in ms) of the initial fixation on a word	early processing
Gaze Duration	sum (in ms) of all fixations on the initial viewing of a word	early processing
Total Fixation Time	duration (in ms) of all fixations made on a word	late processing
Number of Inter-word Regressions	number of regressions made between words	late processing
Average Fixation Count	average number of fixations made on each word	late processing

CHAPTER 4

GENERAL DISCUSSION

This two-study dissertation extended eye movement research and reading research by utilizing techniques across educational, cognitive, and school psychology fields of research in order to better understanding underlying reading comprehension processes in elementary students. Both studies built upon prior research in educational psychology examining test-taking strategies (e.g., Bishop, 2001; Bishop & Frisbie, 1999; Cerdán, Vidal-Abarca, Martínez, Gilabert, & Gil, 2009; Daneman & Hannon, 2001; Farr, Pritchard, & Smitten; 1990; Wiesendanger & Wollenberg, 1978; Wiesendanger, Birlem, & Wollenberg, 1982) and question format (Martinez, 1999; Ozuru, Best, Bell, Witherspoon, & McNamara, 2007) as well as eye movement research evaluating reading behavior in adults (e.g., Feng et al., 2012) and school-aged children (e.g., Ardoin, Zawoyski, Wagner, Bangs, & Binder, 2015; Solheim & Uppstad, 2011; van der Schoot, Vasbinder, Horsley, & van Lieshout; 2008; Vorstius, Radach, Mayer, & Lonigan, 2013). Chapters 2 and 3 implemented a group contingency to simulate a high-stakes testing experience. Additionally, both studies ensured that participants were not significant different in their working memory abilities due to the close relationship between working memory and reading comprehension (Cain, 2006).

Review of Chapter 2

Chapter 2 presented the first study, aimed at examining natural test-taking behavior in third- and fourth-grade students and evaluating the impact of two test-taking strategies on students' eye movements during reading. Specifically, Chapter 2 compared the questions-first

strategy (QF), in which students read a set of questions prior to reading their corresponding passage and the passage-first strategy (PF), in which students read a passage prior to reading their corresponding questions. An important study conducted by Farr et al. (1990) illuminated the concept that individuals may engage in different strategies during a reading comprehension assessment. Results obtained from eye movement research with second grade students also found that children selected different reading strategies, including the PF and QF strategies (Ardoin et al., 2015). Extant research was unclear regarding the benefits and disadvantages of the QF strategy, with arguments suggesting that it may encourage more efficient responding (Wiesendanger & Wollenberg, 1978) and other studies suggesting that it is less efficient than the PF strategy (Bishop 2001; Bishop & Frisbie, 1999; Cerdán et al., 2009; Daneman & Hannon, 2001).

Results from Chapter 2 indicated that third- and fourth-grade readers elected a QF strategy infrequently, with only 13 out of 67 participants engaging in this behavior under natural conditions. Analyses investigated the impact of experimental condition (i.e., PF or QF) on eye movements during reading across the passage, on the questions, and on areas of the passage containing and not containing answers to literal questions. Findings indicated that generally, the QF condition resulted in longer, more frequent fixations and less efficient test-taking. Analyses examining participants' accuracy on MC questions suggested that question difficulty interacted with condition, such that participants utilizing a QF strategy outperformed participants utilizing a PF strategy on more challenging questions, whereas participants utilizing a PF strategy outperformed participants utilizing a QF strategy on easier questions. Further inspection of the interaction suggested that the interaction was a result of one difficult item. Without over-interpreting findings from one question, results from this item cautiously raise important

questions for future studies. The difficult item was a literal question with the answer clearly available in the text, however, distractor items may have been easily selected based on their plausibility. Therefore, participants in the QF condition may have been primed to search for the correct answer based on their previous experience with reading the questions, whereas participants in the PF condition may have relied on prior knowledge to answer this deceptive question.

Ultimately, findings from the study described in Chapter 2 suggest that further research should be conducted to examine the conditions under which a QF strategy would be beneficial. Until such findings are obtained, teachers were cautioned to avoid recommending the QF strategy due to its generally poor efficiency without significant benefits in accuracy.

Review of Chapter 3

Chapter 3 presented findings from a study evaluating the impact of anticipated question format on elementary students' eye movements during reading. Specifically, Chapter 3 examined the benefits and disadvantages of multiple-choice (MC) question formats and short-answer (SA) question formats. Proponents of MC questions suggest that they are efficient with respect to test-taking time and monetary value (Livingston, 2009; Martinez, 1999). In contrast, proponents of SA questions suggest that MC questions may modify the reading comprehension construct (Martinez, 1999; Rupp, Ferne, & Choi, 2006). Feng et al. (2012) conducted an eye movement study examining the reading behavior of adult participants who expected to answer MC questions or summarize a passage. Results indicated that adult participants read more thoroughly when expecting to summarize a passage. Chapter 3 describes a study involving 87 third- and fourth-grade students who practiced reading a passage and responding to questions that were consistent with their assigned experimental condition (i.e., MC or SA questions). They

were then informed that they would read a passage and respond to the type of questions associated with their experimental condition, although they actually responded to both types of questions.

Results from Chapter 3 suggested that the SA condition resulted in more in-depth reading of the text as a whole in addition to meaningful areas of the text. This finding was supported by similarities between the MC and SA conditions in measures of early reading behavior, but differences between the MC and SA conditions in measures of higher-level reading behavior, with longer measures evidenced for the SA condition. This result is consistent with previous findings from research with adults (Feng et al., 2012). Additionally, a correlational analysis between eye movement variables, reading measures, and accuracy revealed that most eye movement measures were significantly and negative correlated with reading measures, suggesting that efficiency in eye movement measures is related to improved reading skill level and supporting findings from a recent review conducted by Foster, Ardoin, and Binder (in press). Notably, the SA questions required significantly more time to complete than the MC questions. Furthermore, findings from multiple linear regressions examining the relation between eye movement measures and accuracy on MC questions indicated that reduced time spent reading words and increased regressions to previous areas led to improved outcomes. Considered alongside findings from Ozuru et al. (2007), further research is required on the impact of access to the text while responding to questions.

Overall, results from Chapter 3 revealed that elementary students' eye movements during reading could be influenced based on their expectation for question format. Test developers were encouraged to consider findings from Chapter 3 in development of reading comprehension

assessments and include a mixture of MC and SA questions to encourage both in-depth reading as well as efficiency in testing.

Contributions of Both Studies

Together, the studies presented in Chapter 2 and Chapter 3 were the first to examine underlying reading behavior in third- and fourth-grade students as test-taking strategies and were manipulated. Perhaps the most important contributions of this research are the applied implications for teachers and test developers. Specifically, results from Chapter 2 and Chapter 3 suggest that teachers can evoke efficient, in-depth reading behavior by encouraging students to use a PF strategy when taking tests and by including more SA items in reading comprehension assessments. Teachers should also prepare students to perform well on tests by encouraging them to find support for their answers within the text rather than attempt to use "short-cuts" such as the QF strategy, which are actually more time-consuming. Additionally, test developers may consider utilizing more SA items on elementary-level reading assessments in order to evoke higher-level comprehension. Developers of test-preparation materials should also heed the limitations of the QF strategy discussed in Chapter 3, as it is less efficient than the traditional PF strategy.

Both studies also highlight the benefits of an interdisciplinary approach; this research was founded in cognitive psychology, but informed by principles of behaviorism from educational and school psychology fields. Notably, both studies included the first examination of inter-observer agreement (IOA) for eye movement data cleaning using a moderately conservative method. Results indicated strong IOA across both studies. Gaining information about reliability in data cleaning increases the believability of results. With continued collaboration between eye

movement research and school psychology research, the understanding of reading comprehension and education will hopefully be improved.

Limitations and Future Directions

The primary limitation across both studies was that each experimental condition only included one passage. Therefore, results should be evaluated with consideration that idiosyncrasies of individual passages may have impacted outcomes. Replication with multiple passages and different samples is essential to increase the plausibility of conclusions for both studies. Future eye movement studies should include longer assessment periods in order to examine students' reading behavior throughout a typical reading comprehension assessment period. Additionally, although both studies ensured that participants were not significantly different in their working memory abilities, working memory was not included in analyses with experimental conditions. Future studies should examine the impact of working memory abilities more closely to assess for interactions with question format and test-taking strategies.

Ultimately, the impact of high-stakes testing is too important to leave uninvestigated. Future studies building on the research conducted in this two-study dissertation could lead to the development of evidence-based test-taking strategies and assessment formats, with the goal of measuring and evaluating reading comprehension as accurately as possible.

References

Ardoin, S. P., Zawoyski, A. M., Wagner, L., Bangs, K., & Binder, K. S. (2015). *Measuring test-taking behavior: Different behaviors but similar outcomes.* Manuscript submitted for publication.

Bishop, N. S. (2001, April). *The validity of reading comprehension test scores: Evidence of generalizability across difference test administration conditions.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, Washington.

Bishop, N. S. & Frisbie, D. A. (1999, April). *The effects of different test-taking conditions on reading comprehension test performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Cain, K. (2006). Children's reading comprehension: The role of working memory in normal and impaired development. In S. J. Pickering (Ed.), *Working memory and education* (pp. 61-91). Cambridge, MA: Academic Press.

Cerdán, R., Vidal-Abarca, E., Martínez, T., Gilabert, R., & Gil, L. (2009). Impact of question answering tasks on search processes and reading comprehension. *Learning and Instruction, 19,* 13-27. doi:10.1016/j.learninstruc.2007.12.003

Daneman, M. & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology, 130,* 208-223. doi:10.1037/0096-3445.130.2.208

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27,* 209-226. doi:10.1111/j.1745-3984.1990.tb00744.x

Feng, G., Gorin, J., Sabatini, J., O'Reilly, T., Wall, C., & Bruce, K. (2012, July). Reading for understanding: How comprehension facilitates answering questions, and what questions enhance understanding. In G. Feng (Chair), *Higher order literacy skills.* Symposium conducted at the meeting of the Society for the Scientific Study of Reading, Montreal, Canada.

Foster, T. E., Ardoin, S. P., & Binder, K. S. (in press). Reliability and validity of eye movement measures of children's reading. *Reading Research Quarterly.*

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; How we score them. *R & D Connections, 11,* 1-8.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist, 34,* 207–218. doi:10.1207/s15326985ep3404_2

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction, 25,* 399-438. doi:10.1080/07370000701632371

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shape the construct: A cognitive processing perspective. *Language Testing, 23,* 441-474. doi:10.1191/0265532206lt337oa

Solheim, O.J. & Uppstad, P. H. (2011). Eye-tracking as a tool in process-oriented reading test validation. *International Electronic Journal of Elementary Education, 4,* 153-168.

van der Schoot, M., Vasbinder, A. L., Horsley, T. M., & van Lieshout, E. C. D. M. (2008). The role of two reading strategies in text comprehension: An eye fixation study in primary school children. *Journal of Research in Reading, 31,* 203–223. doi:10.1111/j.14679817.2007.00354.x

Vorstius, C., Radach, R., & Lonigan, C. J. (2014). Eye movements in developing readers: A comparison of silent and oral sentence reading. *Visual Cognition, 22,* 458-485. doi:10.1080/13506285.2014.881445

Wiesendanger, K. D., Birlem, E. D., & Wollenberg, J. (1982). A summary of studies related to the effect of question placement on reading comprehension. *Reading Horizons, 23,* 15-21.

Wiesendanger, K. & Wollenberg, J. (1978). Prequestioning inhibits third graders' reading comprehension. *The Reading Teacher, 31,* 892-895.