Learning Disabilities: A Contemporary Journal 12(1), 43-68, 2014

Effects of Note-taking and Extended Writing on Expository Text Comprehension: Who benefits?

Michael Hebert*
University of Nebraska—Lincoln

Steve Graham
Arizona State University

Hope Rigby-Wills
Houston Public Schools

Katie Ganson
University of Nebraska—Lincoln

Writing may be an especially useful tool for improving the reading comprehension of lower performing readers and students with disabilities. However, it is reasonable to expect that students with poor writing skills in particular, may actually be less adept at using writing to improve their reading skills, and may not be able to do so without instruction. The purposes of this study were to examine (1) the extent to which writing about text (i.e., taking notes or writing an extended response) would enhance reading comprehension, (2) whether note taking was more effective than writing extended responses for improving reading comprehension for fourth grade students across three measures, and (3) whether the effects of the two writing tasks were moderated by student writing ability, indicating a minimum level of writing proficiency needed to take advantage of writing to improve reading. Students were randomly assigned to a note taking condition in which they took notes about an expository text, an extended writing condition in which they compared and contrasted ideas from the text with their own experiences, or a read and study control condition in which they studied the important ideas from the text. Minimal instruction was provided to the students in each treatment group during a single 45-minute session, primarily to ensure they understood their assigned task. The students then met for another 45-minute session, during which they were asked to read an expository passage and complete their assigned task. Students' reading comprehension was tested using three measures. Students in the combined writing treatments made significantly greater gains than students in the read and study condition on a multiple choice inference measure. No other statistically significant differences were found between the treatment groups, and no moderator effects were found. Implications for future research are framed in terms of the limitations of the study.

Keywords: Reading comprehension, note-taking, extended writing, fourth grade students.

*Please send correspondence to: Michael Herbert, Department of Special Education and Communication Disorders, University of Nebraska, 318Q Barkley Memorial Center, Lincoln, NE 68583-0738, michael.hebert@unl.edu.

Despite large-scale efforts to improve students' reading in policy endeavors such as *No Child Left Behind* and *Reading First,* a large number of students in this country are not particularly good readers. The 2009 National Assessment of Educational Progress (NAEP; National Center for Educational Statistics, 2010) reported that only 38% of 12th grade students performed at or above the "proficient" level in reading (defined as competent academic performance). In terms of younger students, only 33% of 4th graders and 32% of 8th graders performed at the proficient level or above (National Center for Educational Statistics, 2009). In contrast, 34%, 43%, and 36% of 4th, 8th, and 12th grade students, respectively, scored at the "basic" level, denoting only partial mastery of the literacy skills needed at their grade-level. The rest of the tested students' scores were below this basic level.

In the dawn of the adoption of Common Core State Standards Initiative in many states, more emphasis is being placed on reading informational text. This may be concerning for students with learning disabilities (LD), who have particular difficulties with expository text. Research suggests that students with LD lack awareness of organizational components of expository text (Dickson, Simmons, & Kameenui, 1998; Gersten, Fuchs, Williams, & Baker, 2001). Due to this lack of awareness of text organization, students with LD do not approach text with a plan of action, and often retrieve information in a random way (Gersten et al., 2001).

Writing as a Potential Solution

A potential tool for improving expository text comprehension is writing. Some theorists have argued that writing about information enhances learning or causes new learning to occur (Klein, 1999; Newell, 2007). Systematic reviews of experimental and quasi-experimental literature found writing activities to be effective for improving content area learning (Graham & Perin, 2007), academic outcomes (Bangert-Drowns, Hurley, & Wilkinson, 2004), and reading outcomes, including word reading, reading fluency, and reading comprehension (Graham & Hebert, 2010, 2011).

Klein (1999) argued that writing may facilitate learning in four ways: 1) writing fosters explicitness and structured thinking through semantic and syntactic choices, 2) it creates a permanent product that can be reviewed and transformed when contradictions arise, 3) it encourages authors to construct relationships among ideas, and 4) it may help writers to generate and revise goals for the audience based on new content and ideas. It has further been suggested that the cognitive processes involved in writing correspond to general modes of learning that can be actively applied through metacognitive and self-regulation strategies by writers to improve their learning (Bangert-Drowns et al., 2004). Students who actively engage in thinking about their own thinking during writing are more likely to monitor, evaluate and adapt the strategies they use to elaborate ideas, build conceptual frameworks, and synthesize knowledge.

Writing may be an especially useful tool for improving the reading comprehension of lower-performing readers and students with disabilities. Graham and Hebert (2010, 2011) found a statistically significant average weighted effect size of 0.63 across studies in which lower-achieving students wrote about text using various writing tasks (i.e., note-taking, summary writing, question generation, question

answering, and extended writing tasks), which was larger than the statistically significant average weighted effect size found for studies involving all students (ES = 0.51). Although the effect sizes presented by Graham and Hebert do not represent a direct comparison between lower- and higher-achieving students in the same studies, they suggest that students with learning difficulties may actually profit more from writing than their normally achieving peers. One potential explanation is that students with stronger reading skills may not need to rely on another skill to augment their comprehension of text, while students with weak reading skills benefit from additional support to improve their understanding.

Alternatively, it is reasonable to expect that students with poor writing skills may not be able to use writing to improve their reading skills without instruction. An average weighted effect size for writing about text was equal to zero when lower-achieving students did not receive instruction in how to use the targeted writing activities (Graham & Hebert, 2010, 2011). As the authors indicated, "Students who do not develop strong writing skills may not be able to take full advantage of the power of writing as a tool to strengthen reading" (Graham & Hebert, 2010, p. 29). Students with writing disabilities may be at a particular disadvantage, as they may struggle with handwriting, spelling, ideation, and organization of writing, causing them to write less and work with fewer ideas than their typically developing peers.

Different Writing Tasks May Yield Differential Results

Due to the various challenges struggling writers and students with disabilities face when writing, it may be that success with using writing to improve reading depends on the writing task. For instance, the length and complexity of extended writing tasks may make it difficult for students with writing disabilities to take advantage of these tasks to learn from reading. Conversely, note taking may be better suited to the writing skills of weaker writers, as it allows for brevity. Notes also may not require much organization or elaboration, as students can write single words or short phrases to represent big ideas, instead of complete sentences.

Complicating this, some writing-to-read tasks may facilitate distinct types of learning and thinking. Langer and Applebee (1987) argued that tasks such as note-taking, summary writing, and answering questions focus students' attention on the text as a whole and lead to superficial manipulation of the content. Conversely, extended writing tasks require students to reformulate and extend ideas (Kiewra, 1989). In other words, various writing tasks may cause students to perform differently on assessments of reading comprehension based on the focus of the writing task and the information tapped by an assessment. Some support for this contention was provided in a meta-analysis conducted by Hebert, Gillespie, and Graham (2013), who found that specific writing tasks were sometimes more effective when comprehension was assessed using *treatment-inherent* measures. Treatment-inherent measures were defined as measures that are highly similar to the intervention task used to promote better comprehension, as opposed to *treatment-independent* measures that are not so tightly tied to the intervention task (Slavin, 2008a, 2008b). For example, note-taking may be more aligned with factual recall than with inference making.

Purposes of the Current Study

A primary reason for conducting this study was to determine whether writing ability moderates the effectiveness of writing for improving reading. Because the complexity of writing tasks may influence whether students complete the task adequately, two writing tasks of theoretically different levels of complexity were examined in this study: *note-taking* (NT) and *extended writing* (EW). This led to three purposes for the study.

Purpose one. The first purpose of this study was to establish whether NT and EW were effective for improving the reading comprehension of fourth grade writers, after controlling for initial writing ability. Although both NT and EW have previously been shown to be effective for improving reading (see Graham & Hebert, 2010, 2011), these treatments have not previously been examined with fourth grade students using expository text. Therefore, it was important to establish the effectiveness of the writing tasks in the current study, prior to comparing the two writing tasks. To this end, a control group was included in this study in which students read and studied (RS) text without writing. In order to keep the comparisons in the study orthogonal, we combined the NT and EW groups for the comparison to the RS control group. This resulted in an overall test of whether writing (in general) was effective for improving reading comprehension in the current study.

Purpose two. The second purpose of this study was to compare whether NT was more effective than EW for improving reading comprehension. As previously suggested, NT and EW activities may result in students attending to different aspects of the text, which, in turn, should lead to differential effects on reading comprehension. Kiewra (1989) suggested note taking is an effective tool for writing about text because it serves as an encoding function that increases attention to text and allows for surface organization of information. Langer and Applebee (1987) argued that note-taking focuses student attention on passage specific ideas and allows students to read in small segments, but typically results in little integration ideas. On the other hand, extended writing activities focus attention on generating, integrating, evaluating, combining, and recombining ideas, resulting in a deeper level of processing (Langer & Applebee, 1987).

Purpose three. The final, but driving purpose of this study was to examine whether the treatment comparisons in purposes one and two were moderated by students' writing ability. That is, the homogeneity of regression lines assumption was examined to determine whether there was consistency in the effects for treatment across different levels of student writing ability. To investigate this, students' initial writing ability was assessed prior to the experiment and included as a covariate in the final regression models. Interactions were then created between writing ability and the treatment comparison variables. The interaction acted as a test to determine whether the effects of treatment were consistent across levels of student writing ability. This analysis was directed at our interest in determining if struggling writers, including those who have learning difficulties, benefit from writing about their reading.

Rationale for including students of all ability levels. A primary reason for conducting this study was to determine whether students with writing problems have enough skills to take advantage of writing as a tool to improve reading. Nevertheless, all fourth grade students from schools in the study were included in the analyses.

This decision was based on the notion that many students are identified with learning disabilities late in their education career (Torgesen, 2005). For example, Catts, Compton, Tomblin, and Bridges (2012) used latent transition analysis to model changes in reading classification, and estimated that 13.4% of children could be classified as late-emerging poor readers. Thus, using only students with already identified disabilities may have excluded an important subset of students. Further, analysis of the interaction term was expected to estimate the minimum level of writing ability necessary to take advantage of writing to improve reading, which may not have been possible if only students with disabilities were included.

Research Questions

This study was designed in terms of its three purposes to answer three research questions: (1) Are the combined writing treatments (CW) more effective than reading and studying for improving expository text comprehension for fourth grade students, after controlling for initial writing ability? (2) Is note-taking more effective than extended writing for improving the expository text comprehension of fourth grade students, after controlling for initial writing ability? (3) Does writing ability moderate the effects of the treatment conditions for questions one and two?

Hypotheses

A true experiment across multiple sites was used to examine the research questions, with students randomly assigned to treatments. We hypothesized that CW would outperform the RS condition on all comprehension measures (Research Question 1), as both writing groups would benefit from manipulating the ideas in the text and making explicit choices about which ideas to include in their written products. We further hypothesized that students in the NT condition would significantly outperform students in the EW condition on the recall of factual information, but that students in the EW condition would significantly outperform students in NT when asked to apply the information to a new situation in an extended writing task (Research Question 2).

Writing by treatment interactions. In the model comparing CW to RS (Research Question 1), a significant interaction between writing ability and treatment (Research Question 3) was expected. In this comparison, it was expected that weaker writers would benefit from CW as compared to RS. However, we hypothesized that stronger writers may also be stronger readers, and they would not necessarily need to use writing as a tool to augment their reading comprehension. Therefore, the effect for CW may be smaller for stronger writers than weaker writers. This was expected across all measures of reading comprehension.

A significant writing ability by treatment interaction (Research Question 3) was also expected in the model comparing NT to EW (Research Question 2). The NT task involved writing words and short phrases instead of connected text, and the relationships between ideas could be organized by physical arrangement on the page, rather than through text descriptions. On the other hand, the EW task required students to generate, integrate, evaluate, combine, and recombine ideas in connected text. For these reasons, it was hypothesized that the NT would be an easier writing task than EW. Thus, the reading comprehension of weaker writers was expected to

improve more from NT than EW. Alternatively, it was expected that stronger writers would be better able to complete the EW task as intended. Although the stronger writers were also expected to complete the NT task without issue, it was expected that the stronger writers would benefit more from the deeper processing the EW task was expected to elicit. Therefore, it was predicted that the stronger writers would benefit more from EW than NT.

Method

Participants

Fourth grade was selected as the ideal grade level for comparing the writing tasks. Two reasons influenced this decision. One, it is widely agreed that fourth grade is the grade-level at which students make the transition from learning to read to reading to learn (Chall, 1983, 1996). Consequently, students at this grade level are expected to read more expository text and often demonstrate comprehension through writing. Two, previous systematic reviews have not identified studies that have examined the effects of these tasks on the expository text comprehension of fourth grade students or younger (see Graham & Hebert, 2010, 2011). Thus, this study extends the literature on the effectiveness of these tasks for younger students.

Participants for the study included 209 students from 13 fourth-grade classrooms across three schools from a school district in the south that serves rural and suburban schools. All fourth-grade students were eligible for inclusion in the study. Using a person-randomized, multi-site design, students were randomly assigned (within-classroom) to one of three treatment conditions: (a) reading and studying with no writing ($n = 69$), (b) note-taking ($n = 70$), and (c) extended writing ($n = 70$). During the course of the study, 12 students were lost to attrition due to lack of attendance, four students could not be included due to failure to complete posttests, and one student moved. Consequently, 192 students (88 boys, and 104 girls) completed the study: 61 in the NT group, 67 in the EW group, and 64 in the RS group. Consistent with the schools' populations, the majority of students who completed the study were Caucasian ($n = 158$; 81.9%). Students ranged in age from 9.51 to 11.56 years ($M = 10.26$; $SD = 0.38$). Twenty-six students (13.5%) received special education services. Fifty-four students (28.1%) received free or reduced lunch. Demographic information summarized by treatment group can be found in Table 1. After randomization, categorical data were examined for potential relationships between the demographic variables and treatment groups using the chi-squared test for independence. A statistically significant chi-squared value was found for the relationship between gender and treatment, suggesting that a disproportionate number of boys and girls were assigned to each condition ($\chi^2 = 7.09$, $p = .029$). Follow-up analyses showed that the NT group had a disproportionate number of males (59%), while the EW group had a disproportionate number of females (64.2%). Chi-square analyses contrasting Group X Race [$\chi^2 = 4.64$, $p = 0.79$] and Group X Special Education Status [$\chi^2 = 0.03$, $p = 0.99$], were not statistically significant.

Initial writing performance was measured using the third edition of the Wechsler Individual Achievement Test (WIAT-III, Breaux, 2010; described later); the average standard score on the WIAT-III was 107.76 ($SD = 13.96$). A one-way ANOVA

was used to analyze whether there were differences between the treatment groups on the pretest writing measure. No statistically significant difference was found between the groups, $F(2, 189) = 0.45$, $p = .638$.

Table 1. ***Demographic Information of Students by Treatment Condition***

	Read and Study (n = 64)	Note Taking (n = 61)	Extended Writing (n = 67)	Total (n = 192)
Age				
Mean	10.23	10.29	10.25	10.26
SD	(0.36)	(0.38)	(0.40)	(0.38)
Gender				
Males	28 (43.8%)	36 (59.0%)	24 (35.8%)	88 (45.8%)
Females	36 (56.2%)	25 (41.0%)	43 (64.2%)	104 (54.2%)
Race				
White	53 (82.8%)	52 (85.2%)	52 (77.6%)	158 (81.9%)
Black	6 (9.4%)	5 (8.2%)	7 (10.4%)	18 (9.3%)
Asian	1 (1.6%)	0	2 (3.0%)	3 (1.6%)
Hispanic	3 (4.7%)	4 (6.6%)	5 (7.5%)	12 (6.2%)
Other	1 (1.6%)	0	0	1 (0.5%)
Unknown	0	0	1 (1.5%)	1 (0.5%)
Primary Language				
English	62 (96.9%)	59 (96.7%)	65 (97.0%)	186 (96.9%)
Spanish	1 (1.6%)	1 (1.6%)	2 (3.0%)	4 (2.1%)
Amharic	1 (1.6%)	0	0	1 (0.5%)
Unknown	0	1 (1.6%)	0	1 (0.5%)
Students with Disabilities				
Yes	9 (14.1%)	8 (13.1%)	9 (13.4%)	26 (13.5%)
No	55 (85.9%)	53 (86.9%)	58 (86.6%)	166 (86.5%)
Writing Pretest (WIAT-III)				
Mean	109.00	106.64	107.60	107.76
SD	(14.53)	(14.30)	(13.17)	(13.96)

Note. WIAT-III = Wechsler Individual Achievement Test, 3rd Edition, paragraph writing subtest.

Reading Passages Used in the Experiment

The reading passages used for this study were informational texts previously used by the National Assessment of Educational Progress (NAEP) to test the reading comprehension skills of fourth grade students (National Center for Education Statistics, 2012). These passages were chosen because they were considered to be fourth-grade level appropriate informational passages by the National Assessment Governing Board (National Center for Education Statistics, 2011). The first passage, "Daddy Day Care," contained information about how penguins care for their young,

and was used on Day 2 as the example passage for which each treatment group task was modeled. The second passage, "A Brick to Cuddle Up to," provided information about strategies colonial Americans used to stay warm in the winter, and was used as the experimental passage.

Treatment Conditions

The study took place over four consecutive days. All study activities were completed on the same days at all three schools. Students took the pretest measure on Day 1 of the study, and were then randomly assigned to treatment conditions. On Day 2, the "instructors" (the first author and two graduate assistants) modeled and demonstrated how to carry out the experimental tasks to students in each condition. Each instructor taught all three conditions in a counterbalanced order, across the three schools on the same day.

The treatment conditions were carefully designed to include similar elements, instructions, and examples, so that the conditions differed only in the treatment task. Instructors modeled the task and provided students with time to practice it during Day 2, and gave students a sheet of "tips" to remind them of their task on Day 3. Students then read a text and completed the treatment task on Day 3. The instructors provided only "minimal instruction" to ensure students knew what was expected, but did not teach students how to identify or use important information in each condition. The intent was to examine whether students would benefit from the tasks if they were simply assigned; if students benefit from the tasks with minimal instruction, it may not be necessary to spend valuable instructional time on these skills.Students were given the posttest measures on Day 4. Time was controlled across condition (45 minutes each on Days 1 – 3, and 60 minutes on Day 4).

Read and Study (RS). Students in the RS condition received instructions to read a passage and study the important ideas. The instructors used an interactive think-aloud to model an example of one way to study text after reading. Emphasis was placed on identifying important information, using single words and short phrases to represent big ideas, and repeating the information to aid in memory. Instructors did not tell students *how* to choose what was important, but told students they could study the text any way they chose, as long as it did not involve writing.

On Day 3, students in the RS group were asked to read a new passage and study the important information. No writing tools or paper were provided to the students. The instructors monitored the students to ensure that they did not write.

Note-taking (NT). Students assigned to the NT group were instructed to take notes on important information. The instructors modeled paraphrasing main ideas and details in note form on Day 2. Examples were written in single words and short phrases grouped together in unconnected text. Students were not told *how* to choose what was important, nor how to organize their notes. Instead, the instructors emphasized that choices about important information, and the organization of notes, were up to the individual students.

On Day 3, students were asked to read a new passage and take notes on the important information. The instructors provided students with pencils and lined paper for taking notes.

Extended-Writing (EW). Students assigned to the EW condition wrote compare and contrast essays to connect information between the text and their prior knowledge; this task was chosen because research suggests reading comprehension improves when students make such connections (National Institute of Child Health and Human Development, 2000). To ensure they knew what was expected, the instructors provided an example of a good compare and contrast essay to the students on Day 2. The example essay included four paragraphs comparing and contrasting how penguins take care of their young (from the text) with how people take care of their young (something the students should have some prior knowledge about). The example included an introduction, a paragraph about the similarities, a paragraph about the differences, and a concluding paragraph. It also included words and phrases indicating whether a comparison or contrast was made (e.g., similarity, same, alike, different, difference, dissimilar), and these words and phrases were highlighted during instruction. However, students were not given instruction on how to identify which information was relevant, nor how to organize or order the ideas they chose to write about. Instead, the instructors emphasized that choices about the ideas used for comparisons and contrasts were up to the individual students.On Day 3, students were asked to read a new passage and write a compare and contrast essay. Students were given the following prompt:

"Compare and contrast how people in colonial times stayed warm in winter with how people stay warm in winter today."

The instructors provided students with pencils and lined paper on which to write.

Measures

Four measures were used in the experiment, one pretest measure and three posttest measures. The writing pretest is described below, followed by the three outcome measures.

Pretest measure: Wechsler Individual Achievement Test, 3rd Edition (WIAT-III). On Day 1 of the study, students were pretested for initial writing ability using the expository paragraph-writing subtest of the Wechsler Individual Achievement Test, Third Edition (WIAT-III, Breaux, 2010). The test publishers report a test-retest reliability of .82 for scoring of the Theme Development and Text Organization of the paragraph-writing subtest for grade 4. The test was administered to all of the students in each of the 13 classrooms by the first author and two research assistants. Students were given 10 minutes to write about their favorite game and give at least three reasons why it was their favorite. All compositions were scored by a graduate student researcher, with a random sample of the essays (33%) scored by the first author. The scores of both raters were correlated to obtain interrater reliability; Interrater reliability of the scoring was .91. Only the scores of the graduate student researcher were used in the analyses.

Outcome measures: Topic Knowledge, Multiple Choice Inference, & Application Essay. Outcome assessments were given on Day 4 of the experiment, one day following the treatment. Because the two writing treatments may have effects on different aspects of reading comprehension (see the Introduction), three measures were used. Two of the measures were designed to be "treatment-inherent," one

aligned with the NT treatment and one aligned with the EW treatment, and one measure was designed to be a "treatment-independent" measure that did not align with either treatment (Slavin, 2008a, 2008b).

Topic Knowledge (Aligned with NT). A measure of passage specific knowledge adapted from Langer and Applebee (1987) was used to measure students' memory of factual information explicitly presented in the text. Students were asked to write free-association responses to four key topics from the passage: 1) The center of family life in the colonial home, 2) foot-stoves, 3) bathing in colonial times, and 4) keeping warm at bedtime. Students were instructed to write everything they could remember about each topic using single words, short phrases, or complete sentences. To ensure the students understood how to respond, the instructor modeled an example response using an unrelated topic prior to distributing the assessment.

The *Topic Knowledge* measure was considered *"treatment inherent"* because it was designed to align with the NT treatment. Like note taking, the free association response allowed students to write short words or phrases, and the four topics required students to recall factual information across the whole text.Prior to scoring, each topic was reduced to independent facts introduced in the text. Each fact was listed on a scoring sheet, by topic. The student responses were then parsed into propositions, and each proposition was compared with facts on the scoring sheet. Based on the comparison, each of the students' propositions were placed into one of the following categories adapted from scoring systems used by Hayes (1987) and Konopak, Martin and Martin (1990): a) text reproductions; b) incorrect information; or c) irrelevant information. *Text Reproductions* were defined as a match between a proposition in the student response and a proposition in the text, although they were not required to match verbatim. Propositions that conflicted with information in the text were classified as *Incorrect Information. Irrelevant Information* was broadly defined as information not directly referenced in the passage, regardless of whether it was true, untrue, fact, or opinion.Responses from students in all conditions were de-identified, typed, and double-scored in random order. Two raters (the first author and a graduate student research assistant) parsed each response into idea units (defined as clausal units) and categorized the propositions. Interrater reliability for categorizing propositions was .93.After parsing and categorizing the propositions, two scores were created, a "Total Correct" and a "Proportion Score." Each instance of a *text reproduction* was totaled across all categories for Total Correct. The Total Correct scores of both raters were averaged. The Total Correct score captured all of the information students remembered specifically from the passage.However, some students' responses included long lists of irrelevant and or incorrect information with only sporadic correct answers. In those instances, it appeared students might have stumbled across a correct answer in their response. Therefore, a "Proportion Score" was calculated by dividing the number of *Text Reproductions* by the total number of propositions (i.e., *text reproductions* plus *incorrect* and *irrelevant* propositions).

Application Essay (Aligned with EW treatment). The *Application* measure was designed to align with the EW treatment. Similar to the compare and contrast writing done in the EW treatment, the essay measure required students to process the ideas presented in text, analyze how those ideas relate to another situation, and elaborate on the ideas in an extended response. The question was also specifically designed

to elicit responses related to the ideas about staying warm, the same ideas compared and contrasted by students in the EW treatment on Day 3.The assessment required students to write an extended response to a question asking them to apply concepts presented in the text to a new situation. The question read:

> "Imagine that it is a very cold winter. After a bad snowstorm, the electricity goes out in the whole city and it is going to take about a week to fix it. Because of that, you will have no heat in your house. Describe what you and your family could do to stay warm at home and elsewhere?"

Students were given paper, pencils, and copies of the question. They were provided 20 minutes to construct their response. The instructions were read aloud to the students and they were given an opportunity to ask questions. After students' questions were answered, they were instructed to begin writing.

Two raters (the first author and a graduate student) scored the responses. The essays were scored on two dimensions: 1) *Application* of the concepts presented in the text, and 2) *Elaboration* on the ideas in the text. Scores were summed to create a total score. Interrater reliability, calculated as the correlation of the *total score* between the raters, was .93.

Application of Concepts. Raters scored the essays by comparing idea units in the essay with idea units in the passage. Each idea unit was scored by awarding points based on following scale: Zero points for ideas not included in the original text; 1 point for each idea partially representing an idea from the original text; and 2 points for each idea fully representing an idea from the original text, including the correct verbiage. One bonus point was awarded to the essay if the student referenced colonial times, history, or the original text anywhere in the essay.

Elaboration on Ideas. Each essay was also rated on the extent to which it included a new innovation for concepts described in the text, a realistic or sensible reason for wanting to use each strategy, or an appropriate elaboration on how it might be used. Points could only be scored for elaborations if they were connected to the application score.

Multiple Choice Inference Measure (Treatment-independent). This measure included 15 multiple choice questions, developed by the authors, that required students to make inferences based on information provided in the reading passage. Each item had four possible answers for students to choose from, consisting of one correct answer and three distractors. Each question was scored as either correct or incorrect, and the number of correct answers was summed to create a total score for the measure. A total score of 15 points was possible.

The multiple choice measure was considered to be independent of the treatments for two reasons. First, the items did not require written responses, which might have favored one or both of the writing treatments due to the mode of response. Second, the multiple choice items required students to make inferences from text using clues from the content presented, which did not align with any of the treatment tasks. That is, students in the RS and NT conditions completed tasks requiring them to study or take notes on information explicitly presented in the text, but not beyond the text. Conversely, students in the EW group were asked to complete a task requiring them to examine how the ideas in the text related to prior knowledge. Although

the EW task required deeper processing and reorganization of the ideas presented in the text, it did not require students specifically to make inferences about information in the text. A graduate student scored the multiple-choice measure, with 30 percent of the items scored by the first author for reliability purposes. As expected, reliability of scoring was high (r = 0.97), with only three errors found due to mistakes in coding.

Study Implementation

The study included four sessions and took place over four consecutive school days. The instructors (the first author and two graduate students) conducted the experimental procedures.

Day 1 - Pretesting. On Day 1, the students were provided an overview of the study schedule and procedures, excluding details about differences in the treatment conditions. The instructors then administered the WIAT-III paragraph-writing subtest to assess student writing ability. Students were provided 10 minutes to complete the writing test. For students who missed Day 1 (n = 6), there was no time for a make-up test prior to implementing the study, because the intervention occurred over four consecutive days. However, because growth on standardized measures was expected to be minimal over that time frame, a make-up day was included following the study for those students. (Note: Students were also given a standardized pretest reading measure, but differences in administration across the groups invalidated the results and necessitated that the measure be dropped from the study).

Day 2 – Examples and modeling for each of the treatment groups. The modeling occurred in separate classrooms for each treatment, reducing the possibility of treatment contamination. Students were randomly assigned to treatment groups within classrooms, requiring them to be regrouped and moved to appropriate classrooms for instruction.

The first author counterbalanced the instructors across conditions and classrooms to control for potential teacher effects. Specifically, each instructor conducted the modeling portion of each treatment condition at least once, and one of the treatment conditions twice (assigned randomly). Instructors followed a written script, which included modeling and think aloud examples aimed at helping students understand the task. During training, the instructors read the scripts word for word to become familiarized with the protocol for each treatment. For each condition, the purpose of the modeled task was discussed with students, followed by an example of one way to complete the task. Students were given opportunities to ask questions. The instructors explained that the demonstration represented only an example of how students might complete the assigned task.

Treatment Integrity. The instructional steps included in the modeling and examples provided on Day 2 of the interventions were examined for implementation fidelity. All instructional sessions were tape recorded and reviewed by a graduate assistant who was not involved with the intervention and was blind to the hypotheses. The graduate assistant checked for fidelity using the same checklist used by the instructors, and marked off the steps completed. The sessions included four lessons for

each treatment groups, twelve lessons overall. Treatment fidelity was high, with more than 90% of the steps completed as intended in all three of the instructional conditions, including a mean score of 96.00% (*SD* = 3.28) for RS, 93.27% (*SD* = 3.68.) for NT, and 96.67% (*SD* = 1.28) for EW. A one-way ANOVA was conducted to determine whether there were differences between the conditions. No statistically significant difference was found between the groups, $F(2, 11) = 1.50$, $p = .274$.

Day 3 – Students complete their assigned writing or studying tasks. On Day 3, students were again grouped by treatment condition to complete the task for their treatment condition. For consistency, the instructors worked with the same groups for which they modeled the tasks on Day 2.Instructors gave students a sheet of paper with written instructions, along with the tips for completing their assigned task. The instructors read the instructions and tips aloud to the students, and then asked them to read the passage and complete the task. Students in the NT and EW conditions were provided with writing materials (i.e., pencils, erasers, and paper). Students in the RS condition were not provided with writing materials, and instructors monitored them to ensure they did not write during the session.Students read the experimental passage titled "A Brick to Cuddle Up to." They were told they could ask the instructor to read single words to them if they got stuck, but the instructor would not read phrases or sentences to them. Student completed their NT, EW, or RS task immediately following the reading. The instructors monitored the students to ensure they completed the assigned task, prompting students to keep working if they were off task.

Day 4 - Posttests. The order of the posttests was counterbalanced to control for any potential order effects. The instructors were randomly assigned to give the assessment in six counterbalanced orders. Students were randomly assigned, within each treatment condition, to take the assessments with one of the three instructors. The students were then regrouped so that the instructions for each of the test orders could be given to the entire group at once. Students were not given the opportunity to review the reading passage prior to taking the tests.The instructions and items for the assessments were read aloud to the students to reduce the possibility of differences in the outcome due to students' ability to read the test. For the *Topic Knowledge* measure, an interactive example was provided and completed orally as a class (see measure description earlier in the Method section). The instructors then read the prompt for each item and provided the students with 15 minutes to complete the test.

For the *Multiple Choice Inference* measure, the instructors read each of the test items and four possible answers for each item, repeating each question and answer before moving on to the next question. The multiple choice measure took about 10 minutes to complete.For the *Application* essay, the instructors read the instructions aloud to the students, and then read the question and provided students with 20 minutes to construct a response. The instructors repeated the question and directions to students as necessary.

Analysis

As a starting point for each of the analyses, an unconditional two-level mixed-effects model was examined to determine the portion of variance due to class-

room differences, as compared to individual differences. The interclass correlation coefficients (ICCs) were calculated for each outcome measure based on the following model for student *i* in classroom *j*:

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij}$$

$$\beta_{0j} = \gamma_{0j} + \delta_{ij}$$

where β_{0j} is the mean score of each school, and γ_{0j} is the grand mean. The models were estimated using Stata's xtmixed command, using the following syntax with the multiple choice measure used as an example:

```
xtmixed mc || teacher:, var
```

where *mc* was the multiple choice outcome. The ICCs calculated for each of the outcome measures indicated that two percent or less of the variance was attributable to classrooms for all of the measures, indicating that a multilevel analysis was not necessary. However, Roberts (2007) cautioned against assuming no group dependence based on a small ICC, arguing that the degree of dependence may actually depend on the covariates included in the model. Therefore, the first author estimated the full model for each of the outcome measures, including all of the covariates and interactions chosen for the analyses, and then recalculated the ICCs. The ICCs dropped to zero in all of the models. Table 2 shows the ICCs calculated for each of the outcome measures in the unconditional model and fully defined models.

Table 2. ***Intra-class correlations for the unconditional and fully defined multilevel models***

Outcome Measure	ICC (Unconditional Model)	ICC (Full Model)
Multiple Choice	0.01	0.000
Essay (Concept Application and Elaboration)	0.01	0.000
Topic Knowledge (total correct)	0.02	0.000
Topic Knowledge (proportion correct)	0.00	0.000

Note. ICC = Intraclass Correlation Coefficient

Likelihood ratio tests comparing the multilevel models to simple linear regression models were also statistically non-significant in each instance, indicating that a simple linear regression was appropriate for all three outcomes. Therefore, single level regression analyses were conducted to examine the effects of treatment.

Data Modifications. Data were examined prior to and during the analyses to be sure that the models met the regression assumptions. During this process, it was necessary to modify the data due to missing values and non-normal data patterns. The data and regression models were also examined for potential outliers.

Missing data. Despite the pretest make-up session, two participants who completed all other aspects of the study were not able to complete the pretest. To avoid losing the participants to attrition, values for their pretest writing scores were imputed using the *mi impute mvn* procedure in STATA/SE 11. The *mi impute mvn* employed a Markov Chain Monte Carlo (MCMC) method using data augmentation to generate missing values, assuming a multivariate normal model (StataCorp, 2009). Ten imputations were produced. An average of the ten imputations was calculated and substituted for the missing values in the two missing cases.

While the MCMC method assumes multivariate normality, the inferences made based on multiple imputations using MCMC are robust if the amounts of missing data are not large (Yuan, 1990). In this case, the amount of missing data imputed was only 1.03% of the pretest writing data, and less than 0.1% of the overall data used in the regression models.

Data Transformations. The assumption of normality was checked for each of the regression models prior to making inferences. The models for each of the outcome variables were constructed with all of the variables in their original metric. Heteroskedasticity was examined using the Breusch-Pagan/Cook-Weisburg test. The models returned Chi-square values of 0.06 ($p = 0.80$), 0.03 ($p = 0.87$), 9.91 ($p = .002$), and 5.50 ($p = .02$) for the multiple choice, essay, topic knowledge (total correct), and topic knowledge (proportion correct) outcome, respectively. These results indicated that heteroskedasticity was not a concern for the multiple choice and essay outcomes. However, there was statistically significant heteroskedasticity in the models for both of the topic knowledge outcomes.

Further examination revealed scores for the Application Essay, Topic Knowledge Total Correct, and Topic Knowledge Proportion Correct outcomes were not normally distributed. Box and Cox (1964) suggested that transformation of the dependent variable may be desirable for satisfying the assumptions of multiple regression, and to produce the simplest possible regression model. Further, fitting a linear model to transformed variables often leads to a clearer analysis than positing a non-linear model (Singer & Willett, 2003). Likelihood-ratio tests of Box-Cox regression models for both outcomes allowed for the rejection of the null hypothesis that no transformation was needed, Topic Knowledge Total Correct ($\chi^2 = 30.03$, $p < .001$) and Topic Knowledge Proportion Correct ($\chi^2 = 9.37$, $p = .002$). Examination of quantile-normal plots based on the ladder of powers indicated that taking the square root was the most appropriate transformation for both variables.

Following the transformations, recalculated Breusch-Pagan/Cook-Weisburg tests of heteroskedasticity were not statistically significant for Topic Knowledge Total Correct ($\chi^2 = 0.05$, $p = .83$) or Topic Knowledge Proportion Correct ($\chi^2 = 0.47$, $p = .49$) measures, indicating that the transformations were successful in eliminating the heteroskedasticity in these models.

Potential Outliers. Casewise diagnostics were obtained to identify possible outliers in each of the regression models. Several potential outliers were identified in each model. All models were run with and without the potential outliers included. Elimination of the outliers did not result in significant changes to the models, nor interpretations of any of the results. Therefore, all potential outliers were included in each of the final models.

Results

The results for the treatment comparisons are organized by outcome measure, with the research questions addressed for each measure. For the topic knowledge measure, two scores were created and analyzed as outcomes in separate models: 1) the total number of correct propositions, and 2) the proportion of correct answers to the total number of propositions. Thus, although there were only three outcome measures, four regression models were created to accommodate the two scores for the Topic Knowledge measure. The means and standard deviations for the measures are presented for each of the treatment groups in Table 3.

All of the research questions were addressed using one single-level regression model per outcome, with the same independent variables used in the examination of each of the measures. First, students' pretest writing scores from the WIAT-III were included as a covariate, as more skilled writers were expected to perform better on the outcomes. Gender was also included as a covariate due to the disproportionate number of males and females in the NT and EW conditions, coupled with the tendency of girls to be better writers than boys (Berninger & Fuller, 1992; Graham, 2006). To examine the effects of treatment, contrast coding was used to make orthogonal comparisons for Research Question 1 (the comparison of CW to RS) and Research Question 2 (the comparison of NT to EW). Two interaction terms were included in the models to examine potential heterogeneity of the effects of treatment across different levels of student writing ability (Research Question 3) for each of the comparisons: 1) Contrast 1 – [(CW versus RS) X WIAT-III], and 2) Contrast 2 – [(NT versus EW) X WIAT_III].

For each outcome measure, we report the results of the model and identify the results that pertain to each research question.

Table 3. ***Means and Standard Deviations for each Treatment Condition on Three Outcome Measures***

Outcome Measure	Read & Study (n = 64)	Note-taking (n = 61)	Extended Writing (n = 67)
Multiple Choice	8.00 (2.12)	8.74 (2.41)	8.37 (2.18)
Essay	3.57 (2.26)	3.86 (3.49)	3.26 (2.59)
Topic Knowledge			
Total Correct	5.88 (3.37)	5.83 (3.61)	5.06 (2.81)
Proportion	0.51 (0.25)	0.53 (0.26)	0.46 (0.25)

Note. Scores for the essay measure are a sum of scores on the application and elaboration rubrics.

Multiple-Choice Outcome

Results of the regression model for the multiple-choice outcome can be found in Table 4, columns 2-4. All variables were entered simultaneously. The model results revealed that the variables explained 9% of the variance in the outcome, $F(6, 185) = 2.94$, $p < .001$. Student writing ability was a statistically significant predictor of scores on the multiple-choice measure ($t = 3.36$, $p < .001$). The coefficient was 0.04, indicating a 10-point standard score increase on the writing pretest was associated with an increase of 0.4 questions answered correctly when controlling for gender, treatment group, and treatment by writing skill interactions.

Students in the combined writing (CW) conditions outperformed students in the RS condition when controlling for initial writing ability and gender, resulting in a statistically significant main effect for treatment, as predicted (*Research Question 1*). The coefficient for the CW to RS contrast was significant ($t = 2.20$, $p = .029$) and positive ($B = 0.57$), indicating that students who wrote scored an average of .57 points higher on the 15 question measure, or had 3.8% more correct answers than students who read and studied the text without writing. This represents an effect size of 0.34 favoring the writing treatments. However, The coefficient for the interaction of writing ability with Contrast 1 was not statistically significant ($t = -1.14$, $p = .254$; *Research Question 3*). Therefore, the null hypothesis that the slopes were homogeneous for the writing treatment groups when compared to the RS treatment cannot be rejected. In other words, the positive effect of the writing was not significantly different across levels of student ability.

There were no main effects for the second contrast included in the model (*Research Question 2: NT vs. EW*). The coefficient was not statistically significant ($t = 0.47$, $p = .636$). Therefore, the null hypothesis that there is no difference between the scores for these treatment groups cannot be rejected. There were also no statistically significant interactions between the treatment comparisons and writing ability. The coefficient for the interaction of writing ability with Contrast 2 was also not statistically significant ($t = 0.60$, $p = .548$), indicating the null hypothesis that the slopes are homogeneous for the NT and EW groups cannot be rejected (*Research Question 3*).

Application Essay

Results of the regression model for the application essay can be found in columns 5-7 of Table 4. The variables included in the model explained 12% of the variance in the application essay outcome, $F(6, 185) = 4.21$, $p < .001$. Gender was not a statistically significant predictor of the essay outcome ($t = 0.53$, $p = .594$). However, student writing ability was a statistically significant predictor of the essay scores ($t = 3.79$, $p = .001$). The coefficient was 0.02, indicating a 10 point standard score increase on the writing pretest was associated with an increase of 0.2 increase in the essay score when controlling for gender, treatment group, and treatment by writing skill interactions.

Table 4. ***Summary of Regression Analyses for the Effects of Treatment and Writing Skill on Four Reading Outcomes***

	Multiple Choice			Application Essay			Topic Knowledge					
							(Total Correct)			(Proportion)		
Variable	*B*	SE_B	*t(185)*	*B*	SE_B	*t(185)*	*B*	SE_B	*t(185)*	*B*	SE_B	*t(185)*
Intercept (B_0)	8.19	0.24	34.65**	1.53	0.08	17.98**	2.02	0.07	27.32**	0.48	0.03	18.14**
Gender[d]	-0.28	0.35	-0.79	0.09	0.13	0.74	0.13	0.11	1.14	-0.01	0.04	-0.14
Writing Ability[a]	0.04	0.01	3.36**	0.02	0.004	3.79**	0.02	0.003	5.28**	0.002	0.001	1.58
CW vs RS[b]	0.57	0.26	2.20*	0.01	0.09	-1.21	-0.08	0.08	-0.97	-0.03	0.03	-1.01
NT vs. EW[c]	0.10	0.22	0.47	0.06	0.08	0.72	0.08	0.07	1.15	0.03	0.02	1.13
(CW vs. RS) X Writing Ability	-0.18	0.02	-1.14	0.01	0.01	1.43	0.01	0.003	1.06	0.002	0.002	1.17
(NT vs. EW) X Writing Ability	0.01	0.01	0.60	0.01	0.08	1.08	0.004	0.003	0.86	0.001	0.002	0.78
Model Fit	R^2 = .09, = .06			R^2 = .12, = .09			R^2 = .19, = .17			R^2 = .04, = .01		
Omnibus Test	*F*(6, 185) = 2.94**			*F*(6, 185) = 4.21**			*F*(6, 185) = 7.38**			*F*(6, 185) = 1.30*ns*		

Note. CW = Combined Writing Treatments. RS = Read and Study. NT = Note-Taking. EW = Extended Writing.

Due to contrast coding and the inclusion of Treatment by Writing Skill interaction terms, the regression coefficients for Writing Skill represent the gain in the outcome associated with an increase in pretest writing ability when controlling for gender and treatment.

[a] Centered. [b] Contrast Coded Treatments: Note-taking = .5, Extended Writing = .5, Read and Study = -1. [c] Contrast Coded Treatments: Note-Taking = 1, Extended Writing = -1, Read and Study = 0.

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

Learning Disabilities: A Contemporary Journal 12(1), 43-68, 2014

Contrary to predictions there were no statistically significant main effects for treatment on this outcome. We were unable to reject the null hypotheses that there were no differences in the scores between CW and the RS condition (t = -1.21, p = .228; *Research Question 1*), and that there were no differences between the scores of students in the NT and EW conditions (t = 0.72, p = .474; *Research Question 2*). Additionally, there were no statistically significant differences for the variables included to examine whether there were interactions between writing ability and Contrast 1 (t = 1.08, p = 0.281) or writing ability and Contrast 2 (t = 1.43, p = 0.156), indicating no interaction between the treatments and student writing ability (*Research Question 3*). In other words, the null hypothesis of similar slopes for each treatment across levels of writing ability cannot be rejected, meaning that any differences between the treatments, or lack thereof, are expected across all levels of student ability.

Topic Knowledge-Total Correct

Results of the regression model for topic knowledge outcome can be found in columns 8-10 of Table 4. The model explained 19% of the variance in the Topic Knowledge-Total Correct outcome, $F(6, 185) = 7.38$, $p < .001$. Gender was not a statistically significant predictor of this outcome measure ($t = 1.14$, $p = .255$). However, student writing ability was a statistically significant predictor of the essay scores ($t = 5.28$, $p < .001$). The coefficient for writing ability was 0.02, indicating a 10 point standard score increase on the writing pretest was associated with a 0.2 increase in the square root of the number of correct propositions included in the student responses to this measure (or a 0.04 point increase in the number of propositions when accounting for the variable transformation) when controlling for gender, treatment group, and treatment by writing skill interactions.

There were no statistically significant main effects for treatment on this outcome. Therefore, we were unable to reject the null hypotheses that there were no differences between CW and RS (t = -0.97, p = 0.335; *Research Question 1*), and that there were no between the scores of students in the NT and EW conditions (t = 1.15, p = .252; *Research Question 2*). The coefficients were also not statistically significant for interactions between writing ability and the treatments included in the first contrast (t = 1.06, p = 0.290) and writing ability and the treatment comparison included in the second contrast (t = 0.86, p = 0.388), indicating no interaction between the treatments and student writing ability (*Research Question 3*). Thus, the lack of differences would be expected across all levels of student writing ability.

Topic Knowledge-Proportion Correct

The regression model results for the proportion score of the topic knowledge outcome can be found in Table 4, columns 11-13. Contrary to predictions, the model did not explain a statistically significant amount of variance for this outcome, $F(6, 185) = 1.30$, $p = .260$. There were also no statistically significant predictors of the outcome variable included in the model. Therefore, the null hypotheses for the three research questions were not rejected for this measure.

Discussion

Writing has been shown to improve learning (Bangert-Drowns et al., 2004; Graham & Perin, 2007) and reading comprehension outcomes (Graham & Hebert, 2010, 2011). Consequently, more attention is being paid to writing as an essential element of reading instruction (Duke, Pearson, Strachan, & Billman, 2011). However, insufficient research has been conducted to determine if weaker writers and students with learning disabilities can take advantage of writing to improve reading, or how factors such as gender, writing task, or type of reading comprehension measure may impact the effects of writing on reading for different writing tasks. In the present study, we examined whether note-taking and extended writing tasks were effective for improving the expository text comprehension of fourth grade students on three reading comprehension outcomes. We also examined whether the two writing tasks were differentially effective across measures, and whether student writing ability moderated the effectiveness of these writing tasks. It is important to restate that only minimal instruction was provided to students on how to complete the writing tasks, in order to test whether instruction is necessary for these tasks, as many classroom teachers may simply assign such writing tasks without providing instruction. All results must be interpreted with this in mind.

Is writing more effective than reading and studying for improving the expository text comprehension of fourth graders?

The comparison of the combined writing treatments to the read and study treatment across three outcome measures in this study yielded mixed results. There were no statistically significant differences between groups on the application essay or the topic knowledge measures. However, the CW groups statistically outperformed the RS group on the multiple-choice outcome measure, identifying more correct inferences. With an ES of 0.34, the results indicated that students who wrote about the text scored an average of 3.8% higher on the outcome than students who read and studied without writing.

Based on these findings, the act of writing seemed to allow students who wrote to solidify the information in a way that caused them to be able to identify correct inferences more often than students who did not write. One potential explanation is that fourth grade students who wrote about text remembered the information more readily than students who did not write, freeing up cognitive resources that were used to think beyond the text. While we need to replicate this finding before any strong conclusions can be drawn, such a finding may provide an important step toward improving the theory about how writing influences text comprehension.

It is also important to explore potential reasons for the lack of findings on the Essay and Topic Knowledge measures, as these results run contrary to expected findings based on prior research (see Graham & Hebert, 2010, 2011). An important consideration for the lack of findings here is the strength of the control condition. Instead of using a no-treatment or business-as-usual (BAU) control condition, we opted to use a read-and study treatment as our comparison condition. The RS treatment may be stronger than control conditions used in previous studies, as instructors provided modeling and tips for how to study the information, time was controlled for, and the instructors made sure students stayed on task by repeatedly suggesting

students reread or study more to students who appeared off-task. This may not be typical of instruction in schools, where teachers may allow students to read other texts or complete unfinished assignments when they are finished studying. Moreover, only minimal instruction in writing was provided to the treatment groups, and additional instruction in writing may have strengthened the treatments. Nevertheless, the lack of findings on these measures is important, because it suggests there may be other options for improving comprehension beyond writing, depending on the type of comprehension expected of students for a particular reading assignment.

Is note-taking more effective than extended writing for improving the expository text comprehension of fourth grade students, after controlling for initial writing ability?

Based on the results, the null hypothesis that there was no difference between the two writing treatment groups for improving reading comprehension outcomes cannot be rejected. There were three potential reasons for these findings: 1) lack of power, 2) students' inability to complete the writing tasks as intended, and/or 3) potential treatment contamination.

Lack of Power. First, the study may not have been sufficiently powered. With power set to .80, an alpha level of .05, and 128 participants, this comparison had the power to detect an effect size of $d = 0.36$ or greater. This difference was reasonable to expect based on effect sizes found for these two writing tasks in prior research (Graham & Hebert, 2010, 2011). However, the differences between the groups in this study were smaller, resulting in statistically non-significant effect sizes for the multiple choice (ES = 0.16, *ns*), topic knowledge (ES = 0.24, *ns*), and essay outcomes (ES = 0.19, *ns*), respectively. All three statistically non-significant effect sizes favored note-taking over extended writing. The consistency of these effects favoring the note-taking treatment indicates a potentially meaningful difference between the two groups, suggesting it may be more beneficial to have fourth grade students take notes about text than to write extended responses, if the goal is to improve reading comprehension.

Students' inability to complete the writing tasks as intended. Some students in the two writing groups had difficulty completing the NT and EW writing tasks as intended, which may have led to a less than perfect comparison of the treatments. Examination of the notes and essays revealed commonly recurring difficulties with the writing tasks. This may simply mean that fourth grade students do not yet have sufficient writing and/or reading skills to take advantage of these tasks to augment their reading comprehension without additional instruction.

Problems with NT. The most common characteristics of note-taking difficulties were:Notes were sparse, including little to none of the information from the passage

1. Notes represented only one aspect of the passage, ignoring complete sections entirely
2. Notes included superfluous information not included in the passage
3. Notes resembled connected text instead of identifying important information
4. Notes included all or most of the information in the passage.

The first three characteristics on this list revealed that some students did not complete the note-taking task as intended. Although this does not necessarily mean these students did not engage in the thinking involved in identifying and remembering the important information (i.e., the writing may reflect fluency or mechanics issues rather than problems with ideation), it is possible that at least some of these students did not fully engage the text as anticipated. On the other hand, the last two characteristics on the list revealed that some of the students had a difficult time distinguishing important information from unimportant information. Thus, an unsystematic approach to taking notes may have led them to remember information arbitrarily.

Problems with EW. Problems with the compare and contrast writing activity were:

1. Inclusion of comparisons or contrasts beyond the scope of the prompt
2. Listing ways colonists stayed warm, making no comparisons or contrasts
3. Limited amount writing, not completing the task
4. Focus on only one aspect of the passage
5. Improperly characterizing similarities as differences, and vice versa.

The first four problems noted above revealed that some students in EW did not make comparisons or contrasts that allowed them to attend to or reorganize the ideas as intended. Problem number five revealed misunderstandings about the content and, in turn, poor comprehension of the text that writing about the ideas was not likely to resolve.

Treatment contamination. An additional potential problem was that treatment contamination might have occurred inadvertently. Some students in the note-taking group included comparisons and contrasts in their notes, much like the EW group, while some students in the EW group simply listed facts about the passage, more closely resembling the NT task. These observations revealed that the two treatments were not always executed as intended by all students, with some students applying procedures intended for the competing treatment.

Did writing ability moderate the effects of writing the treatment conditions for questions one and two?

The third and driving purpose of the experiment was to examine potential interactions between the writing treatments and students' writing ability. We assumed that students of different writing abilities would differentially benefit from one type of writing over another, and that students with writing difficulties would not gain as much of an advantage from writing overall. These interactions were not realized in the current experiment, as all of the interactions between writing ability and the treatment comparisons were statistically non-significant in every model. There are two potential interpretations of these findings, both of which need to be approached judiciously.

One potential and obvious interpretation is that treatment effects of note-taking and extended writing, or lack of effects, were not moderated by student writing ability. However, this interpretation is tenuous at best, due to previous concerns raised about lack of power. Cohen, Cohen, West and Aiken (2003) indicated "the

power to detect an interaction is reduced, relative to first order effects" (pp. 297). Because the study only had the power to detect main effect sizes equal to or larger than $d = 0.36$ and $d = 0.29$ for questions 1 and 2, respectively, it may be that the study simply did not have the power to detect potential interaction effects.

A second possibility is that no moderator effects were found because the fourth grade students included in the study, regardless of initial writing ability, did not write well enough to sufficiently differentiate the writing tasks. If the writing tasks were simply too difficult for the fourth grade students (of any ability level) to complete effectively, then it does not stand to reason that there would be differential effects for tasks by students by measures. Although this point admittedly requires considerable supposition, it is important to accentuate that these interactions could potentially emerge as students become more skilled, or if they were provided more instruction in how to employ the writing tasks.

Conclusion

The findings of this study provide partial support for the theory that writing about text improves reading comprehension for fourth grade students. A significant difference was found between CW and the RS control condition on the multiple choice assessment. However, it is important and prudent to translate the effect size into a more interpretable form to get a clearer picture of what this effect might mean (Lipsey et al., 2012). Although the effect size was small to moderate (ES = 0.34), it suggests only a very slight increase in the number of comprehension questions students answered correctly. Students in the treatment group scored an average of 3.8% higher than students in the control group, or a half-point higher on a 15-point test. Because it is not possible to score a half-point on the multiple choice test, these findings suggests that writing may have been effective for improving the scores of some students, but not others. Therefore, this finding is limited, and should be interpreted cautiously.

Based on the limited findings of the study, it may be that "minimal instruction" is not enough to produce practically significant results for writing about text with fourth grade students. Because these students are just transitioning from "learning to read" to "reading to learn" (Chall, 1983, 1996), it may be that students have not developed sufficient writing skills to take advantage of the writing tasks they were asked to use in this study. Although students may have some experience taking notes or writing extended responses, these experiences may be limited or used primarily with narrative text prior to fourth grade. This is evident in the students' inability to complete some of the writing tasks as intended. Future research should be targeted at providing more instruction for these tasks to fourth grade student, to examine whether instruction makes a difference.

Limitations

There are also some aspects of this study that can provide valuable insight for designing future research in this area, which are best examined through the limitations of the study. First, the study lacked power to detect differences in treatments for effect sizes smaller than 0.29 and 0.36 for the two primary research questions, respectively. Although larger effect sizes have previously been found for these treat-

ment comparisons, this study included students in an earlier grade level than in past research. The data from this study showed that smaller effects may indeed be evident for fourth grade, but there was not sufficient power to obtain statistical significance. This lack of power may also have led to difficulty identifying potential interaction effects in the study. Although smaller, these effects may still be practically significant and important to identify. Future research studies should be designed with smaller effects in mind.

Second, some students had difficulty completing the writing tasks as intended, which likely influenced the effectiveness of the writing approaches studied here. This is an especially important finding, as one of the goals of this study was to determine whether fourth grade students could take advantage of these tasks with minimal instruction. It appears that this was not the case for all students. However, the lack of findings does not suggest that students would not benefit from more intensive instruction in this area, or that minimal instruction (as was applied here) would not be appropriate for students in higher grade levels. Future research on this topic conducted with students in fourth grade and earlier should almost certainly include an instructional component, while studies examining minimal instruction should be conducted with students in later grades and/or designed to look for smaller effect sizes.

Third, the essay writing measure may not have been designed well enough to elicit responses that were indicative of the knowledge students gained through writing. Although the writing prompt was related to the writing done by the EW condition during treatment, the essay question did not directly ask students to apply the information from the passage, nor mention the passage directly. Careful framing outcome measure questions for direct application of knowledge should be an important consideration in future research.

A final consideration is that we did not include student reading ability as a potential covariate. Reading ability would almost certainly have accounted for variability in reading outcomes, and it may be a potential moderator the effects of the treatment tasks. Researchers conducting studies in this area in the future may wish to control for reading ability and examine potential interactions between reading ability and the writing task comparisons.

References

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based Writing-to-Learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74,* 29-58.

Berninger, V., & Fuller, F. (1992). Gender differences in orthographic, verbal, and compositional fluency; Implications for assessing writing disabilities in primary grade children. . *Journal of School Psychology, 30*, 363-382.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211-252.

Breaux, K. C. (2010). *Wechsler Individual Achievement Test – Third Edition (WIATT-III): Technical Manual.* Bloomington, MN: Pearson.

Catts, H. W., Compton, D., Tomblin, J. B., and Bridges, M. S. (2012). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology, 104*, 166-181.

Chall, J.S. (1983). *Stages of reading development*, New York, NY: McGraw-Hill.

Chall, J. S. (1996). *Stages of reading development* (2nd ed.). Fort Worth, Tex.: Harcourt Brace.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (Third ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Dickson, S. V., Simmons, D. C., & Kameenui, E. J. (1998). Text Organization: Research Bases. In D. C. Simmons & E. J. Kameenui (Eds.), *What reading research tells us bout children with diverse learning needs* (pp. 279-294). Mahwah, NJ: Erlbaum.

Duke, N. K., Pearson, P. D., Strachan, S. L., & Billman, A. K. (2011). Essential elements of fostering and teaching reading comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What Research Has to Say about Reading Instruction (4th ed.)*. Newark, DE: International Reading Association.

Gersten, R., Fuchs, L. S., Williams, J. P., & Bakers, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research, 71*, 279 - 320.

Graham, S. (2006). Writing. In P. Alexander & P. Winne (Eds.), *Handbook of Educational Psychology*. Mahwah, NJ: Erlbaum.

Graham, S., & Hebert, M. (2011). Writing-to-read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review, 84*, 710-744.

Graham, S., & Hebert, M. (2010). *Writing to read: An evidence base for how writing can improve reading. A Time to Act Report.* Washington, DC: Alliance for Excellent Education.

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high school. A report to Carnegie Corporation of New York*: Alliance for Excellent Education.

Hebert, M., Gillespie, A., & Graham, S. (2013). Comparing effects of different writing activities on reading comprehension: A meta-analysis. *Reading and Writing, 26*, 111-138.

Hayes, D. A. (1987). The potential for directing study in combined reading and writing activity. *Journal of Reading Behavior, 19*, 333-352.

Kiewra, K. A. (1989). A review of note-taking: The encoding-storage paradigm and beyond. *Educational Psychology Review*, 1, 147-172.

Klein, P. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review, 11*, 203-270.

Konopak, B. C., Martin, S. H., & Martin, M. A. (1990). Using a writing strategy to enhance sixth-grade students' comprehension of content material. *Journal of Reading Behavior, 22*, 19-37.

Langer, J., & Applebee, A. (1987). *How writing shapes thinking: A study of teaching and learning.* Urbana, IL: National Council of Teachers of English.

Lipsey, M., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., Busick, M. D. (2012). *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. (NCSER 2013-3000).* Washington, DC: U.S. Government Printing Office.

National Center for Educational Statistics (2009). *Reading 2009: National assessment of educational progress at grades 4 and 8.* Washington, DC: Institute of Educational Sciences.

National Center for Educational Statistics (2010). *Grade 12 reading and math 2009 national and pilot state results.* Washington, DC: Institute of Educational Sciences.

National Center for Education Statistics. (2011). *What does the NAEP reading assessment measure?* Retrieved March 4, 2012, from http://nces.ed.gov/nationsreportcard/reading/whatmeasure.asp.

National Center for Education Statistics. (2012). *NAEP Question Tool.* Retrieved February 23, 2012, from http://nces.ed.gov/nationsreportcard/itmrlsx/search.aspx?subject=reading.

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* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Newell, G. E. (2007). Writing to learn: How alternative theories of school writing account for student performance. In C. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of Writing Research* (pp. 235-247). New York, NY: The Guilford Press.

Roberts, J. K. (2007). *Group dependency in the presence of small intraclass correlation coefficients: An argument in favor of not interpreting the ICC.* Paper presented at the annual meeting of the American Educational Research Association.Singer, J. D., & Willet, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press.

Slavin, R. E. (2008a). What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*, 5-14.

Slavin, R. E. (2008b). Evidence-based reform in education: Which evidence counts? *Educational Researcher, 37*, 47-50.

StataCorp. (2009). *Stata Multiple-Imputation Reference Manual: Release 11*. College Station, TX: Stata Press.

Torgesen, J. K. (2005). Recent discoveries from research on remedial interventions for children with dyslexia. In M. Snowling & C. Hulme (Eds.), *The science of reading* (pp. 521–537). Oxford, UK: Blackwell.

Yuan, Y. C. (1990). Multiple imputation for missing data: concepts and new development; Proceedings of the 25th Annual SAS Users Group International Conference. Retrieved March 4, 2012, from http://www2.sas.com/proceedings/sugi25/25/st/25p267.pdf.