GRAIN SIZE OF RETRIEVAL PRACTICE FOR LENGTHY TEXT MATERIAL:
FRAGILE AND MYSTERIOUS EFFECTS ON MEMORY

A thesis submitted
To Kent State University in partial
Fulfillment of the requirements for the
Degree of Masters of Arts

By

Kathryn Taylor Wissman

December 2013
Thesis written by
Kathryn T. Wissman
B.A., Kent State University, 2009
M.A., Kent State University, 2013

Approved by
Katherine A. Rawson, Ph.D. ____________, Advisor
Maria S. Zaragoza, Ph.D. ____________, Chair, Department of Psychology
James Blank, Ph.D. ____________, Dean, College of Arts and Science
# TABLE OF CONTENTS

LIST OF FIGURES ........................................................................... v

LIST OF TABLES ............................................................................ vii

INTRODUCTION ............................................................................. 1

EXPERIMENT 1 .............................................................................. 7
   Method .................................................................................... 7
   Participants and Design ......................................................... 7
   Materials and Procedure ....................................................... 7
   Scoring ................................................................................... 8
   Results .................................................................................... 8

EXPERIMENT 2 .............................................................................. 10
   Method .................................................................................... 10
   Participants and Design ......................................................... 10
   Materials and Procedure ....................................................... 11
   Results .................................................................................... 11

EXPERIMENT 3 .............................................................................. 13
   Method .................................................................................... 14
   Participants and Design ......................................................... 14
   Materials and Procedure ....................................................... 14
   Results .................................................................................... 15

EXPERIMENTS 1-3: RELATIVE MEMORABILITY ANALYSIS ...... 17

EXPERIMENT 4 .............................................................................. 22
   Method .................................................................................... 23
   Participants and Design ......................................................... 23
   Materials and Procedure ....................................................... 23
   Results .................................................................................... 24

EXPERIMENT 5 .............................................................................. 27
LIST OF FIGURES

1 Mean number of idea units correctly recalled during practice and final test for each group, Experiment 1. Error bars represent standard errors.

2 Mean number of idea units correctly recalled during practice and final test for each group, Experiment 2. Error bars represent standard errors.

3 Mean number of idea units correctly recalled during practice and final test for each group, Experiment 3. Error bars represent standard errors.

4 Hypothetical outcomes for relative memorability analyses evaluating the claims of the DUDD hypothesis. DUDD hypothesis predicts the pattern shown in Panel A (in which the recall advantage for the section group over whole text group during practice is primarily due to elevated recall of details) and in Panel D (in which the decline in recall in the section group from practice to final test primarily reflects the loss of details). Panels B-C and E-F illustrate other logical possibilities. See text for further details about these predictions.

5 Mean number of idea units correctly recalled during practice and final test for each group, Experiment 4. Error bars represent standard errors.

6 Mean percent of idea units recalled from target sentences during practice and final test for each group, Experiment 4. The left panel represents recall of main ideas,
the middle panel represents recall of important details, and the right panel represents recall of unimportant details. Error bars represent standard errors.

Mean number of idea units correctly recalled during practice and final cued recall test for each group, Experiment 5. Error bars represent standard errors.

Mean number of idea units correctly recalled during practice and final cued recall test for each group, Experiment 6. Error bars represent standard errors.
LIST OF TABLES

1  S = study, T = free recall test, FT = final test, M = 2 minutes of math problems.
   Numbers refer to text section that was studied or recalled, and italics denote
   length of retention interval (m = minute, d = day).

2  Experiment 2 values are collapsed across math groups. Experiment 3 values are
   collapsed across delay groups. Experiment 6 values are collapsed across
   connection groups.

3  Mean values reflect number of idea units recalled. SE = standard error of the
   mean. Near refers to recall of information from the same section of the text that
   contained the prompt sentence. Distant refers to recall of information from a
   section other than the one containing the prompt sentence. Experiment 6 values
   are collapsed across connection groups.
Introduction

A wealth of previous research has established that testing enhances subsequent learning and memory (for reviews, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Rawson & Dunlosky, 2011; Roediger & Butler, 2011). Although various formats of practice tests have been shown to facilitate learning (e.g., multiple choice, cued recall, short answer; Butler, Karpicke, & Roediger, 2007; Barber, Rajaram, & Marsh, 2008; Chan, 2010) the effects of retrieval practice are of particular interest in the current research. Retrieval practice has been shown to produce greater learning and retention compared to alternate test formats (e.g., Butler & Roediger, 2007; Carpenter & DeLosh, 2006; Glover, 1989; Kang, McDermott, & Roediger, 2007), and students report using retrieval practice as a learning technique (Hartwig & Dunlosky, 2012; Karpicke, Butler, & Roediger, 2009; Wissman, Rawson, & Pyc, 2012). Furthermore, retrieval practice has been shown to facilitate the learning of various kinds of materials, including word lists (Zaromb & Roediger, 2010), key-term definitions (Rawson & Dunlosky, 2011), and face-name pairs (Carpenter & DeLosh, 2005).

Of greatest interest here, students are often required to learn from lengthy text material in educational contexts. An important question concerns the grain size at which students should engage in retrieval practice for lengthy expository text. One approach is for students to read the entire text and then try to recall as much information as they can
remember. Another approach is for students to segment the text into smaller sections and attempt to recall information after each section. Hereafter, we refer to these two grain sizes as whole text recall versus section recall. The goal of the current research was to evaluate the extent to which grain size affects recall during practice and subsequent memory for lengthy text material.

Specifically, the current research evaluated the grain size hypothesis, which states that a smaller versus larger grain size will increase retrieval success during practice that in turn will enhance subsequent memory for lengthy text material. The a priori predictions of this hypothesis are based on empirical evidence regarding the effects of retrieval success during practice and the interim test effect. Concerning the effect of retrieval success during practice, previous research has shown that retention is enhanced to a greater extent when retrieval practice is successful versus unsuccessful (Carrier & Pashler, 1992; Butler, 2010; Butler & Roediger, 2008; Butler et al., 2007). Importantly, prior research supports the expectation that retrieval success during practice will be greater for smaller sections versus whole text. Studies directly comparing retention for longer versus shorter texts have shown that learners remember more after reading shorter texts (Mayer & Jackson, 2005; Rothkopf & Billington, 1983). For example, Rothkopf and Billington (1983) presented learners with five expository texts and asked them to remember as much of the information as they could about the passages. Half of the learners were presented with the abridged versions of text passages (approximately 215 words each) whereas the other half of learners were presented with the full versions of text passages (approximately 540 words each, which included the content of the abridged
version plus additional factual information). One day later, all learners took a final completion test over information included in the abridged versions of the text passages. Performance was significantly greater for learners who read abridged passages compared to learners who read full passages. Note that in this study, learners were only tested at one time point (as opposed to retrieval practice followed by a retention test as in the current research) and the total amount of text material that was studied differed (as opposed to learners all studying identical text), so their study did not involve a grain size manipulation per se. Nonetheless, their outcomes provide evidence that retrieval success is greater for a shorter unit of text than for a longer unit of text.

Additionally, Wissman, Rawson, and Pyc (2011, Experiment 2) presented learners with an expository text divided into four sections. Learners were prompted to recall either after studying each text section or after all sections had been studied. Recall was significantly greater for learners who engaged in retrieval practice directly after studying each section compared to learners who engaged in retrieval practice after all sections had been studied. Wissman et al. (2011) did not include a subsequent final test, so the extent to which the level of retrieval success during practice affected retention is unknown. Nonetheless, given that engaging in recall after each section versus after all sections yielded greater retrieval success during practice and that successful retrieval has been shown to enhance subsequent memory, a straightforward prediction is that engaging in section recall versus whole text recall during practice will in turn increase subsequent memory for lengthy text material.
Additional support for this prediction comes from recent research on the interim test effect. In brief, research has shown that taking a test over previously studied material facilitates the learning of subsequent new material (Szpunar, McDermott, & Roediger, 2008; Weinstein, McDermott, & Szpunar, 2011; Wissman, Rawson, & Pyc, 2011). For example, Szpunar et al. (2008) presented learners with five lists of interrelated nouns. After studying each list, learners completed one minute of math problems followed by either a test over the previously studied list or another minute of math problems. After studying List 5, all learners took a practice test over List 5. Recall of List 5 was significantly greater for learners who received interim tests following Lists 1-4 (54%) versus no interim tests (24%). Thus, engaging in retrieval practice after each list facilitated the learning of subsequent lists. Furthermore, the advantage persisted on a final test administered 30 minutes later, such that recall of List 5 was significantly greater for learners who previously engaged in interim testing (56%) versus no interim testing (26%).

Wissman et al. (2011, Experiment 1a) extended these outcomes to show that taking interim tests also enhances the learning of text material. Learners were presented with an expository test divided into three sections. In the interim test group, learners completed free recall for Sections 1 and 2 prior to studying Section 3, whereas learners in the no interim-test group did not. After studying Section 3, all learners took a free recall test over Section 3 (cf. List 5 recall in Szpunar et al., 2008). Recall of Section 3 was significantly greater for learners who received interim tests (21%) versus no interim tests (11%). Not surprisingly, overall recall was lower than in the study by Szpunar et al.
(2008) given the increased difficulty of recalling text material versus word lists. Nonetheless, learners recalled almost twice as much information from Section 3 when preceding sections had been tested versus not tested. The same qualitative pattern of recall emerged in three follow-up experiments. However, none of the experiments reported by Wissman et al. (2011) included a final test, so the extent to which the learning advantage of taking interim tests during practice persists across a delay for text material is unknown.

Although previously discussed outcomes are suggestive, no prior research has directly explored how the grain size of retrieval practice affects subsequent memory for lengthy text material. Accordingly, the current research tested the grain size hypothesis, which predicts that section recall will yield greater retrieval success during practice that in turn will enhance subsequent memory for text material. To overview, in all six experiments, participants studied an expository text that was divided into four sections. Learners were prompted to recall directly after studying each section (hereafter referred to as section group) or after all sections had been studied (hereafter referred to as whole text group). All learners completed a final test after a delay (Table 1 provides a schematic of the method for each experiment). The basic claim of the grain size hypothesis is that engaging in section recall will increase recall during practice, which in turn will enhance memory on the final test. Thus, the grain size hypothesis predicted a main effect of grain size across the two time points, such that recall will be greater for the section group versus the whole text group during practice and on the final test.
To foreshadow, Experiment 1 revealed a significant interaction that partially disconfirmed the prediction of the grain size hypothesis. Although the section group versus the whole text group recalled substantially more information during practice, the sizable effect did not persist on the final test. Experiments 2-3 explored potential methodological explanations for why section recall significantly enhanced memory during practice but not on the final test. Experiments 4-6 explored two theoretical explanations for why the significant grain size effect during practice was attenuated across the delay.
Experiment 1

Method

Participants and design. Undergraduates who participated for course credit were randomly assigned to one of two groups (section or whole text, $ns = 28$ and $25$), with test (practice versus final) manipulated within subjects.

Materials and procedure. Materials included an expository text on government intervention in the labor market (779 words, 12.0 Flesch grade level) based on materials developed by Rawson and Kintsch (2002, 2004). The text was divided into four sections, with each section describing a different form of government intervention. Participants in both groups were given five minutes to study each text section. In the section group, immediately after studying Section 1 participants were shown an empty text field and were given five minutes to recall everything they could remember from that section. Participants then studied Section 2 followed by recall of Section 2, and so on until all four sections had been studied and recalled. In the whole text group, participants studied all four text sections before any recall occurred. Immediately after studying Section 4, participants were given 20 minutes to recall everything they could remember from all four sections. After the practice phase, both groups completed an unrelated filler task for 20 minutes. All participants then completed a final free recall test in which they were asked to recall as much information as they could remember from all text sections. Following the free recall test all participants completed a final cued recall test (from
Rawson & Kintsch, 2002, 2004). Each prompt included part of a sentence extracted from the text and participants were asked to provide the missing part of the sentence. The cued recall test included 16 prompts, with four prompts corresponding to each of the four sections.

**Scoring.** We describe the free recall scoring procedure for all six experiments here. All free recall responses generated during practice and final test were scored. For scoring, text sections were parsed into idea units approximately corresponding to the content of a phrase. Credit was assigned for verbatim responses or correct paraphrases. Across the six experiments, three different expository texts were used. Given the substantial number of free recall responses to be hand-scored, multiple raters were trained to complete the scoring. For each expository text, we chose a random sample of protocols to serve as the training set. Raters were given the training set to score and the reliability of his or her scores was checked against other trained raters (all $rs > .94$). Given that reliability was consistently high across expository texts and raters, each remaining protocol was scored by one of the trained raters.

**Results and Discussion**

Mean recall across all text sections during practice and final test for each group is reported in Figure 1. A 2 (Grain size: section versus whole text) X 2 (Test: practice versus final) mixed-factor analysis of variance (ANOVA) resulted in significant main effects of Grain size, $F (1, 51) = 6.93, MSE = 842.61, p = .011, \eta_p^2 = .12$, and Test, $F (1, 51) = 40.37, MSE = 76.71, p < .001, \eta_p^2 = .44$. Importantly, main effects were qualified by a significant interaction, $F (1, 51) = 37.15, MSE = 76.71, p < .001, \eta_p^2 = .42$. Practice
recall was significantly greater for the section group versus the whole text group, \( t(46) = 4.37, p < .001, d = 1.20 \). Surprisingly, the substantial advantage in recall success during practice did not persist across the delay. At final test, recall was not significantly greater for the section group versus the whole text group, \( t(51) = 0.76, p = .449, d = .21 \). As converging evidence, performance on the final cued recall test was not significantly greater for the section group (\( M = 14.3, SE = 2.4 \)) versus the whole text group (\( M = 9.4, SE = 1.7 \)), \( t(51) = 1.64, p = .107, d = .45 \).
Experiment 2

Experiment 1 partially disconfirmed the grain size hypothesis. Although section recall increased the amount of information recalled during practice as predicted, the advantage did not persist on the final test. One potential explanation is that elevated recall during practice was due to learners in the section group retrieving some information from working memory versus long-term memory. Research has shown that facilitative effects of testing occur when information is retrieved from long-term memory (Butler & Roediger, 2007; Carpenter & DeLosh, 2006) and that subsequent memory is enhanced to a greater degree when retrieval practice is more difficult versus less difficult (Pyc & Rawson, 2009). Therefore, the facilitative effects that arise from engaging in retrieval practice may have been undermined given that learners in Experiment 1 engaged in recall immediately after studying each section. To minimize retrieval from working memory, Experiment 2 included two new groups that completed math equations after studying each text section (hereafter referred to as section-math and whole text-math; see Table 1). If elevated recall during practice is due to retrieval from working memory, minimizing retrieval from working memory via the filler task will decrease initial recall in the section-math group.

Method

Participants and design. Undergraduates who participated for course credit were randomly assigned to one of four groups (section, whole text, section-math, or
whole text-math, ns = 29, 28, 30, and 28, respectively), with grain size (section versus whole text) and task (math versus no math) manipulated between subjects and test (practice versus final) manipulated within subjects.

**Materials and procedure.** Materials were the same as in Experiment 1. The procedure for the section and whole text groups was the same as in Experiment 1. The procedure for the section-math and whole text-math groups was the same except that participants solved math equations for two minutes after studying each text section (see Table 1).

**Results and Discussion**

As shown in Figure 2, results showed the same qualitative pattern as in Experiment 1, such that recall for the section group versus the whole text group was significantly greater during practice but not on the final test. Furthermore, the fragility of the effect apparently was not due to retrieval from working memory. These conclusions were supported by outcomes of a 2 (Grain size: section versus whole text) X 2 (Task: math versus no math) X 2 (Test: practice versus final) mixed factor ANOVA. The main effects of Grain size, $F (1, 111) = 10.21$, $MSE = 383.30$, $p = .002$, $\eta^2_p = .08$, and Test, $F (1, 111) = 243.84$, $MSE = 70.34$, $p < .001$, $\eta^2_p = .69$, and the two-way interaction, $F (1, 111) = 104.89$, $MSE = 70.34$, $p < .001$, $\eta^2_p = .49$ were significant. Completing math equations resulted in no significant main effects or interactions (all $Fs < 2.80$), indicating that elevated recall during practice was not due to the section group retrieving some information from working memory.
Collapsing across math groups, practice recall was significantly greater for the section group versus the whole text group, $t(113) = 6.23, p < .001, d = 1.16$. Replicating Experiment 1, the advantage of section recall during retrieval practice did not persist across the delay. At final test, recall was not significantly greater for the section group versus the whole text group, $t(113) = 1.27, p = .207, d = .24$. Furthermore, performance on the final cued recall test was not significantly greater for the section group ($M = 14.0, SE = 1.4$) versus the whole text group ($M = 13.4, SE = 1.2$), $t(113) = 0.33, p = .742, d = .06$. 
Experiment 3

Experiment 2 partially disconfirmed the grain size hypothesis and established that the significant interaction was not due to the section group retrieving some information from working memory during practice. The goal of Experiment 3 was to rule out another methodological explanation for why the grain size effect observed during practice did not extend to final test. Previous research has shown that the facilitative effects of retrieval practice are more likely to emerge on delayed tests versus immediate tests (e.g., Roediger & Karpicke, 2006; Whitten & Bjork, 1977). For example, Roediger and Karpicke (2006) showed that additional study trials versus additional test trials during practice increased performance on an immediate test but decreased performance on a delayed test. By comparison, the relatively short retention interval (20 minutes) between retrieval practice and final test in Experiments 1-2 may have been insufficient for the recall advantage of engaging section recall versus whole text recall to emerge. Specifically, the whole text group may show a greater memory deficit on a delayed final test. To examine this possibility, Experiment 3 included two new groups that took the final test after two days (hereafter referred to as section-delay and whole text-delay; see Table 1). If a longer retention interval is necessary for the recall advantage of section recall versus whole text recall to emerge, the section-delay group will outperform the whole text-delay group at final test.

In addition to a longer retention interval, Experiment 3 explored whether the same
recall pattern emerges when using a different expository text. Establishing that a
different text produces the same effect will extend the findings of Experiments 1-2 by
showing that results are not specific to one set of materials.

Method

Participants and Design. Undergraduates who participated for course credit were
randomly assigned to one of four groups (section, whole text, section-delay, whole text-
delay, \( ns = 30, 29, 28, 27 \), respectively), with grain size (section versus whole text) and
delay (15 minutes versus 2 days) manipulated between subjects and test (practice versus
final) manipulated within subjects.

Materials and procedure. Materials included an expository text on capturing
and storing greenhouse gases (1,333 words, 12.0 Flesch grade level) based on materials
developed by Rawson and Kintsch (2005). The text was divided into four sections, each
including a subtopic header (Capturing Greenhouse Gases, A New Approach in Norway,
Underground or Underwater, and Safe and Sound). In general, the text sections discussed
the issue of greenhouse gases and potential solutions. Participants were given 4-5
minutes (depending on section length) to study each section. During Session 1, the
procedure for the section and whole text groups was the same as in Experiment 1 with
two exceptions. First, participants completed an unrelated norming study for 15 minutes
as the filler task instead of the computer administered survey. Second, participants did
not complete a final cued recall test. During Session 1, the procedure for the section-
delay and whole text-delay groups was the same except that participants did not take the
final free recall test following the 15 minute filler task (see Table 1). Instead, participants
were informed that they had completed all the tasks for the first session and were reminded to return to the lab two days later. After two days had elapsed, participants in all four groups returned to the lab to complete the second session of the study. During Session 2, all participants were asked to recall as much information as they could remember from the four text sections. Performance on the delayed tests for the section and whole text groups ($M = 19.4$ and $M = 18.4$, respectively) was not of interest but was collected in order to equate participants’ expectations about returning for a delayed test session across all four groups. Therefore, although all participants completed the second session, only outcomes on the delayed tests for the section-delay and whole text-delay groups are discussed further.

**Results and Discussion**

As shown in Figure 3, results showed the same qualitative pattern seen in Experiments 1-2, such that recall for the section group versus the whole text group was significantly greater during practice but not on the final test. Furthermore, the recall pattern was not affected by the length of the retention interval between practice and final test. These conclusions were supported by a 2 (Grain size: section versus whole text) X 2 (Delay: 15 minutes versus 2 days) X 2 (Test: practice versus final) mixed factor ANOVA. The main effects of Grain size, $F(1, 110) = 13.19$, $MSE = 533.59$, $p < .001$, $\eta^2_p = .11$, and Test, $F(1, 110) = 201.88$, $MSE = 78.47$, $p < .001$, $\eta^2_p = .65$, and the two-way interaction, $F (1, 110) = 98.71$, $MSE = 78.47$, $p < .001$, $\eta^2_p = .47$ were significant. Increasing the retention interval between retrieval practice and final test resulted in no significant main effects or interactions (all other $Fs < 2.44$). In fact, the trend was in the
opposite direction than expected, such that a numerical disadvantage on the final test emerged for the section-delay group versus the whole text-delay group.

Collapsing across delay groups, practice recall was significantly greater for the section group versus the whole text group, $t(112) = 6.28, p < .001, d = 1.18$. As in Experiments 1-2, the recall advantage during practice did not persist across the delay. At final test, recall was not significantly greater for the section group versus the whole text group, $t(112) = 0.15, p = .885, d = .03$. 
Experiments 1-3: Relative Memorability Analysis

Experiments 1-3 indicated that although section recall enhanced retrieval success during practice, the section group lost access to a substantial amount of information across the delay. What might explain the recall advantage for the section group over the whole text group during practice, and what might explain why the section group shows such a precipitous drop in recall from practice to final test? One possibility concerns the kind of information that participants are recalling during practice and on the final test. To briefly summarize, theories of text comprehension assume that the memory representation for text material functionally includes three different levels: the surface level (i.e., a representation of exact words and/or phrases used in the text), the textbase (i.e., a propositional representation of the concepts and ideas expressed in the text and their associations), and the situation model (i.e., a multi-modal representation of the meaning of the text integrated with prior knowledge) (for reviews, see Kintsch 1998; Zwaan & Radvansky, 1998). Furthermore, the textbase is assumed to include both a macrolevel (consisting of the main ideas of a text) and a microlevel (consisting of the details of a text). Important for current purposes, prior research has shown that these different kinds of information have different forgetting functions, such that main ideas are more likely to be remembered over time whereas details are more likely to be quickly forgotten (Kintsch, Welsch, Schmalhofer, & Zimny, 1990; Rawson & Kintsch, 2005).
Additionally, prior research suggests that extraneous information (e.g., details) may distract the reader from processing main ideas (Lehman, Schraw, McCrudden, & Hartley, 2007; Mayer & Jackson, 2005; Reder & Anderson, 1982). Taken together, these theories of text comprehension and empirical findings provide a plausible explanation for the recall pattern observed in the section group during practice and on the final test. Specifically, the recall advantage for the section group versus the whole text group during practice may reflect increased retrieval of details. Additionally, the decrease in recall from practice to final test observed in the section group may reflect the loss of details. Accordingly, the *details-up details-down (DUDD) hypothesis* states that elevated recall during practice for the section group reflects retrieval of detail information, which is subsequently lost across the delay.

One approach to explore the kind of information that is being recalled during practice and on the final test is a *relative memorability analysis* (Stine & Wingfield, 1988; Stine & Wingfield, 1990; Verhaeghen & Marcoen, 1993). In a relative memorability analysis (hereafter referred to as RMA), the probability of recall for each idea unit across participants is the dependent measure and the kind of information is empirically defined by normative levels of recall. Idea units that are normatively recalled more frequently are assumed to reflect main ideas whereas ideas that are normatively recalled less frequently are assumed to reflect details. Importantly, RMA can be used to diagnose recall differences between two different groups or at two different time points by comparing the normative level of recall for more memorable ideas (i.e., main ideas) or for less memorable ideas (i.e., details). For example, Stine and Wingfield (1988)
conducted RMA to compare normative levels of recall for younger adults versus older adults. Recall was greater for younger versus older adults, and moreso for main ideas than for details.

For current purposes, we conducted two sets of RMA to evaluate the two basic claims of the DUDD hypothesis. To evaluate whether the recall of details was greater for the section group versus the whole text group during practice, we plotted the probability of recall for idea units during practice in the section group as a function of the probability of recall during practice for the whole text group. The DUDD hypothesis predicts that the recall advantage for the section group during practice primarily reflects an increase in the recall of details. The top row of Figure 4 illustrates three qualitative patterns that could emerge for the RMA plotting the normative level of recall for the section group versus the whole text group during practice. Panel A illustrates the qualitative pattern of recall predicted by the DUDD hypothesis. Specifically, a slope of less than 1.0 suggests increased recall of details for the section group versus the whole text group during practice. Panel B illustrates a pattern of recall opposite to that predicted by the DUDD hypothesis. A slope of greater than 1.0 suggests increased recall of main ideas for the section group versus the whole text group during practice. Panel C illustrates the other pattern of recall that may emerge, which is that the recall advantage for the section group over the whole text group reflects main ideas and details to a similar extent.

To evaluate the second claim of the DUDD hypothesis (that the decline in recall from practice to final test in the section group reflects a loss of details), we plotted the probability of recall for idea units for the section group during practice as a function of
the probability of recall for the section group on the final test. The bottom row of Figure 4 illustrates three qualitative patterns that could emerge for the RMA plotting the normative level of recall for the section group during practice and on the final test. Panel D illustrates the pattern of recall predicted by the DUDD hypothesis. Specifically, a slope of less than 1.0 suggests that the decrease in recall from practice to final test is disproportionately due to the loss of details. Panel E illustrates a pattern of recall opposite to that predicted by the DUDD hypothesis. A slope of greater than 1.0 would indicate that the decrease in recall from practice to final test is disproportionately due to the loss of main ideas. Panel F illustrates the other pattern of recall that may emerge, which is that memory for main ideas and details decreases equally from practice to final test. Taken together, the DUDD hypothesis predicts that both sets of RMA will show shallow slopes, suggesting that the section group recalls disproportionately more details during practice and then loses this information across the delay.

The top three rows of Table 2 report slopes from linear regressions and corresponding $R^2$ values for the RMA conducted on Experiments 1-3. Surprisingly, both sets of RMA revealed a recall pattern opposite to that predicted by the DUDD hypothesis. The RMA comparing the section group to the whole text group during practice for Experiments 1-3 showed slopes greater than 1.0, suggesting that the recall advantage for the section group was primarily due to enhanced recall of main ideas rather than details during practice. The RMA comparing recall during practice and on the final test for the section group for Experiments 1-3 also showed slopes greater than 1.0, suggesting that the section group was disproportionately losing main ideas versus details across the
delay. To foreshadow, we also conducted RMA for Experiments 4-6 for purposes of completeness (see the bottom three rows of Table 2). Outcomes for these RMA are somewhat mixed, and the model fits are not quite as good as for the first three experiments (as indicated by $R^2$ values). Nonetheless, the overall pattern of RMA outcomes provides initial evidence against the DUDD hypothesis. Given that RMA was conducted post-hoc for all experiments, the primary goal of Experiment 4 was to provide a stronger test of the DUDD hypothesis.
Experiment 4

Experiment 3 provided further evidence against the grain size hypothesis by establishing the same qualitative pattern with a new expository text and by ruling out the possibility that significant recall differences on the final test between the section and whole text groups would emerge after a longer delay. In addition to eliminating the explanation of a longer retention interval, RMA outcomes provided evidence against the DUDD hypothesis, a seemingly plausible hypothesis given prior theoretical and empirical work on text comprehension. Surprisingly, RMA outcomes suggested that the recall advantage in the section group during practice was primarily due to main ideas rather than details, and this information was subsequently lost across the delay. However, one potential limitation of the RMA technique in general is that main ideas and details are empirically defined post-hoc via normative levels of recall. As discussed earlier, idea units with lower levels of recall are assumed to reflect details, and idea units with higher levels of recall are assumed to reflect main ideas. However, main ideas are not the only kind of information in text material that may show elevated recall. For example, prior research has shown elevated recall of information that readers rate as highly interesting but that is unimportant to the main idea of a text, referred to as seductive details (e.g., Garner, Gillingham, & White, 1989; Garner, Alexander, Gillingham, & Brown, 1991).

To more directly evaluate the DUDD hypothesis, Experiment 4 used an expository text
including target sentences that were constructed to primarily contain main ideas, important details, or unimportant details (Rawson & Kintsch, 2005). In addition to replicating the basic pattern in overall free recall with another expository text, the pattern of recall for the target sentences during practice and on the final test were of particular interest. The DUDD hypothesis predicts that the section group will show increased recall of unimportant details during practice and decreased recall of unimportant details on the final test.

**Method**

**Participants and design.** Undergraduates who participated for course credit were randomly assigned to one of two groups (section or whole text, $n_s = 27$ and 27), with test (practice versus final) manipulated within subjects.

**Materials and procedure.** Materials included the expository text developed by Rawson and Kintsch (2005), which discussed inconsistencies between Hollywood’s depiction of history and factual history (1,319 words, 11.4 Flesch grade level). The text was divided into four sections, describing why filmmakers choose to modify historical facts and describing a relevant example of a film that does so, *The Charge of the Light Brigade*. Rawson and Kintsch (2005) revised a subset of sentences from the text so that the content of each sentence expressed a main idea, important detail, or unimportant detail. They also conducted a pilot study to collect normative ratings for the kind of information contained in each sentence, in which participants were asked to indicate whether the sentence expressed a main idea, an important detail, or an unimportant detail.
Sentences showing the highest agreement were selected as target sentences (5 main-idea sentences, 10 important-detail sentences, and 10 unimportant-detail sentences).

The procedure for the section and whole text groups was the same as in Experiment 3 except that after the final free recall test all participants completed a recognition test. During the recognition test, participants were presented with 60 phrases one at a time and asked to decide whether the idea was or was not presented in the text. Half of the phrases contained correct ideas from the text and half of the phrases contained incorrect ideas. For example, the following phrase is an actual sentence from the text: “Balaklava was a small but beautiful costal village nestled at the base of the mountains.” On the recognition test, participants saw either “Balaklava was at the foot of the mountains” (i.e., correct idea) or “Balaklava was high in the mountains” (i.e., incorrect idea). Thus, participants had to decide whether the idea had been stated in the text based on the meaning rather than on the exact wording.

Results and Discussion

Mean free recall during practice and final test for each group is reported in Figure 5. A 2 (Grain size: section versus whole text) X 2 (Test: practice versus final) mixed-factor ANOVA resulted in significant main effects of Grain size, $F (1, 52) = 18.25, MSE = 154.07, p < .001, \eta_p^2 = .26$, and Test, $F (1, 52) = 59.91, MSE = 26.62, p < .001, \eta_p^2 = .54$, which were qualified by a significant interaction, $F (1, 52) = 47.37, MSE = 26.62, p < .001, \eta_p^2 = .48$. Practice recall was significantly greater for the section group versus the whole text group, $t(39) = 5.91, p < .001, d = 1.61$. At final test, recall was not
significantly greater for the section group versus the whole text group, $t(52) = 1.50, p = .140, d = .41$

Of greater interest for evaluating the DUDD hypothesis, recall of target sentences for the section group and whole text group during practice and final test are plotted in Figure 6. The left panel reports recall of main ideas, the middle panel reports recall of important details, and the right panel reports recall of unimportant details. To revisit, the DUDD hypothesis predicted that the section group would show increased recall of unimportant details during practice and decreased recall of unimportant details on the final test. Contradictory to the DUDD hypothesis and converging with the overall pattern of RMA outcomes, recall of unimportant details during practice was not significantly greater in the section group versus the whole text group, $t(52) = 0.21, p = .833, d = .06$, whereas recall of main ideas and important details during practice was significantly greater in the section group versus the whole text group, $t(36) = 5.71, p < .001, d = 1.55$, and $t(45) = 4.50, p < .001, d = 1.22$, respectively. Furthermore, the section group showed a substantial decrease in the recall of main ideas and important details from practice recall to final test, $t(26) = 4.44, p < .001, d = 0.85$, and $t(26) = 7.03, p < .001, d = 1.35$, respectively. Thus, contrary to the DUDD hypothesis, elevated recall during practice for the section group consisted primarily of main ideas and important details rather than unimportant details, which learners subsequently lost across the delay.

Concerning performance on the recognition test, the proportion of hits was significantly greater for the whole text group ($M = .78, SE = .02$) versus the section group ($M = .69, SE = .03$), $t(42) = 2.40, p = .021, d = .65$. The proportion of false alarms was
not significantly greater for the whole text group ($M = .40, SE = .04$) versus the section group ($M = .35, SE = .03$), $t(52) = 0.98, p = .329, d = .27$. Corrected recognition (hits – false alarms) was not significantly greater for the whole text group ($M = .38, SE = .05$) versus the section group ($M = .34, SE = .05$), $t(52) = 0.54, p = .595, d = .15$. 
Experiment 5

Taken together, outcomes of the RMA and the pattern of recall for target sentences in Experiment 4 disconfirmed the DUDD hypothesis. Results showed that elevated recall in the section group during practice was due to increased retrieval of main ideas, which surprisingly were then lost across the delay. This finding is contrary to the DUDD hypothesis and is also inconsistent with prior text comprehension research showing greater durability of main ideas versus details. The loss of main ideas observed in the section group is likely not due to the particular expository text used here, given that the whole text group exhibited the conventional pattern of recall (i.e., maintenance of main ideas across the delay, with loss only for unimportant details) and that this text has shown similar patterns in prior research (Rawson & Kintsch, 2005; Rawson, 2012). Theories of text comprehension assume that main ideas (versus details) are maintained across time because they are highly interconnected with other ideas in the text representation (Kintsch, 1998). Empirical findings have also shown that memory for text material is affected by the extent to which ideas are integrated, showing enhanced recall for ideas that have a greater number of connections to other ideas in the text (Kintsch & Keenan, 1973; Miller & Keenan, 2009; Miller & Keenan, 2011). However, although main ideas afford a high degree of interconnectedness with other ideas in the text learners may not necessarily encode these connections during text processing. If learners do not
connect ideas throughout the text, the information is less likely to be retained across time. More generally, connecting ideas throughout the text produces a coherent representation that provides an effective retrieval structure for subsequent recall, such that ideas that are recalled can in turn serve as cues to retrieve other ideas that are connected in the text representation. Conversely, if ideas in the text are not integrated, access is diminished due to fewer retrieval routes. Thus, one explanation for why the section group is losing main ideas across the delay is that section recall may interfere with the integration of information across different text sections, reducing the coherence of the resulting text representation that is encoded.

Recent research on how interim retrieval affects subsequent memory for word lists (Jang & Huber, 2008; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011) is consistent with this explanation. Pastötter et al. (2011) proposed that taking interim tests in between the study of word lists produces an internal context change that may be used as a cue during subsequent retrieval. The context-specific cues facilitate recall by decreasing cross-list interference and increasing discrimination across lists. Although discrimination enhances the retrieval of subsequent word lists, it may have the opposite effect for text material. Specifically, engaging in section recall may segregate the text sections and disrupt the learner from connecting ideas across different sections. If ideas throughout the text are not connected then the overall level of integration may be impaired. Thus, the extent to which information across the text sections is not integrated provides a plausible explanation for why the section group loses main ideas across the
delay. Accordingly, the integration hypothesis states that section recall disrupts the integration of ideas across different text sections.

To test the integration hypothesis, we used a test format developed by Rawson (2012) to explore the role of integration in rereading lag effects. In that study, learners initially studied a text and then reread the text either immediately or after two days. To evaluate the degree of integration of information in the text, the final test included an intersentential cued recall test. Learners were presented with sentences extracted from the text and asked to recall what other information from the text the prompt brought to mind. Cued recall was greater for the long-lag group versus the short-lag group, indicating more extensive integration of information when rereading occurred after a long lag versus a short lag. Cued recall responses were further coded as retrieval of near information (information recalled from the same paragraph in the text that contained the prompt) and retrieval of distant information (information recalled from any paragraph in the text other than the one containing the prompt). Long-lag rereading protected against the loss near information and increased the recall of distant information (although to a lesser extent).

Although Rawson (2012) designed the intersentential cued recall measure to investigate rereading effects, the method can be adapted for present purposes. To revisit, the integration hypothesis predicts that section recall during practice decreases the extent to which ideas are integrated across different text sections. To evaluate the integration hypothesis, Experiment 5 used the intersentential cued recall test developed by Rawson (2012). If engaging in section recall decreases integration across different text sections,
recall of distant information will be lower in the section group than in the whole text group.

Method

Participants and design. Undergraduates who participated for course credit were randomly assigned to one of two groups (section or whole text, ns = 32 and 30), with test (practice versus final) manipulated within subjects.

Materials and procedure. Materials were the same as in Experiment 4. The procedure for the section group and whole text group was the same as in Experiment 4 except that participants did not take a final free recall test. Instead, all participants completed the intersentential cued recall test used in Rawson (2012). The cued recall test included the 8 prompts used in Rawson (2012) plus two additional prompts. On each trial, participants were shown a sentence from the text and asked to type in any other information from the text that the prompt brought to mind. Each prompt afforded a high degree of integration such that it was related to information within the same text section and information from one or more other text section. For example, the prompt “The first way in which filmmakers fictionalize history is by changing the historical story” is related to an idea from the same text section, “Some of the changes made are minor” and to an idea from a different text section, “[The movie] gets the place and date of the battle right, but absolutely nothing else.” Each idea contained in a cued recall response was coded as either near information or distant information. Near refers to retrieval of an idea from the same text section that contained the prompt sentence, whereas distant refers
to retrieval of an idea from a section of the text other than the one containing the prompt sentence.

**Results and Discussion**

Mean free recall during practice and overall performance on the final cued recall test for each group are reported in Figure 7. Replicating Experiments 1-4, a 2 (Grain size: section versus whole text) X 2 (Test: practice versus final) mixed-factor ANOVA resulted in significant main effects of Grain size, $F(1, 60) = 4.96, MSE = 144.33, p = .300, \eta^2_p = .08$, and Test, $F(1, 60) = 26.42, MSE = 50.17, p < .001, \eta^2_p = .31$, which were qualified by a significant interaction, $F(1, 60) = 20.11, MSE = 50.17, p < .001, \eta^2_p = .25$. Practice recall was significantly greater for the section group versus the whole text group, $t(60) = 4.01, p < .001, d = 1.02$. Overall final cued recall was not significantly different between the two groups, $t(60) = 0.38, p = .707, d = .10$.

To directly explore the integration hypothesis, cued recall for near information versus distant information is reported in Table 3. As shown in Table 3, the numerical trend was in the direction predicted by the integration hypothesis. Recall of distant information was lower for the section group versus the whole text group. Although the integration hypothesis makes no a priori prediction concerning retrieval of near information, recall was greater for the section group versus the whole text group. However, differences in the recall of near information and distant information between the two groups were not statistically significant. A 2 (Group: section versus whole text) X 2 (Distance: near information versus distant information) mixed-factor ANOVA
resulted in no significant main effects (all $F$s < 1.93) and a non-significant interaction, $F(1, 60) = 0.93, \text{MSE} = 19.02, p = .339, \eta_p^2 = .02$. 
Experiment 6

Although Experiment 5 showed a numerical trend that was consistent with the integration hypothesis, the recall differences were not statistically significant. Thus, the integration hypothesis was not supported conclusively. One goal of Experiment 6 was to replicate the numerical trend observed in Experiment 5. The second goal of Experiment 6 was to provide an additional test of the integration hypothesis by adding an instructional manipulation intended to enhance integration across the text sections. Experiment 6 included two new groups that were explicitly instructed to connect ideas across the four text sections (hereafter referred to as section-connection and whole text-connection; see Table 1). The integration hypothesis predicts that the instructional manipulation will have a greater effect on the extent to which information is integrated across different text sections in the section-connection group versus the whole text-connection group, to the extent that learners are less likely to spontaneously engage in integration in the section group than in the whole text group.

Method

Participants and design. Undergraduates who participated for course credit were randomly assigned to one of four groups (section, whole text, section-connection, or whole text-connection, ns = 31, 33, 31, and 32, respectively), with grain size (section
versus whole text) and instructions (with versus without instructions encouraging integration) manipulated between subjects and test (practice versus final) manipulated within subjects.

**Materials and procedure.** Materials were the same as in Experiment 5. The procedure for the section and whole text groups was the same as in Experiment 5. The procedure for the section-connection and whole text-connection groups was the same except that participants were explicitly instructed to connect ideas across the different text sections. Experimental instructions were as follows:

“The information contained in each of the text sections is related, so you should be connecting ideas across the different text sections. **Each time you read an idea that is related to an idea you studied in a previous text section, you should click the button at the bottom of your screen.** You will see this button starting with the second section you are asked to study. Again, although you will read different text sections, all of the information is related and each time you make a connection between the current section and a previous section, you should click the button on your screen.”

Starting with Section 2, a button labeled “I made a connection” appeared on the computer screen and participants were asked to click the button each time they read an idea that was related to an idea from a previously studied section. On average, participants in the section-connection and whole-text-connection groups clicked the button 3.6 and 3.7 times, respectively, providing evidence for compliance with instructions.

**Results and Discussion**

As shown in Figure 8, instructing participants to connect ideas across the sections did not affect recall during practice or on the final cued recall test. This conclusion was supported by outcomes of a 2 (Grain size: section versus whole text) X 2 (Instructions: with versus without instructions encouraging integration) X 2 (Test: practice versus final)
mixed-factor ANOVA. The main effects of Grain size, $F(1, 123) = 73.34$, $MSE = 100.47$, $p < .001$, $\eta_p^2 = .37$, and Test, $F(1, 123) = 282.81$, $MSE = 30.64$, $p < .001$, $\eta_p^2 = .70$, and the two-way interaction, $(1, 123) = 123.13$, $MSE = 30.64$, $p < .001$, $\eta_p^2 = .50$ were significant. Instructions encouraging the integration of ideas across the text sections resulted in no significant effects or interactions (all other $Fs < 1$). Collapsing across connection groups, practice recall was significantly greater for the section group versus the whole text group, $t(105) = 10.51$, $p < .001$, $d = 1.87$. Final cued recall was also significantly greater for the section group versus the whole text group, $t(125) = 2.98$, $p = .003$, $d = .53$.

As shown in Table 3, the numerical trend for recall of distant information was inconsistent with the integration hypothesis. Recall of distant information was greater in the section group versus the whole text group. Recall of near information was also greater in the section group versus the whole text group. A 2 (Grain size: section versus whole) X 2 (Distance: near information versus distant information) mixed-factor ANOVA resulted in significant main effects of Grain size, $F(1, 125) = 8.88$, $MSE = 16.83$, $p = .003$, $\eta_p^2 = .06$, and Distance, $F(1, 125) = 87.26$, $MSE = 5.81$, $p < .001$, $\eta_p^2 = .41$, and the interaction was marginal, $F(1, 125) = 2.76$, $MSE = 5.81$, $p = .10$, $\eta_p^2 = .02$. Recall of distant information was significantly greater in the section group versus the whole text group, $t(125) = 2.00$, $p = .048$, $d = 0.35$, and recall of near information was significantly greater in the section group versus the whole text group, $t(125) = 3.05$, $p = .003$, $d = 0.54$. 
Collapsing the data from Experiments 5 and 6 to increase statistical power, the numerical trend was still inconsistent with the integration hypothesis. A 2 (Grain size: section versus whole text) X 2 (Distance: near information versus distant information) mixed-factor ANOVA resulted in a significant main effect of Distance, $F(1, 187) = 46.17, MSE = 10.38, p < .001, \eta_p^2 = .20$. The main effect of grain size and the interaction term were not statistically significant (all other $Fs < 1$). Recall of distant information was not significantly different in the section group versus the whole text group, $p = .564$. Recall of near information was significantly greater in the section group versus the whole text group, $t(187) = 2.36, p = .019, d = 0.34$. 
General Discussion

Six experiments tested the grain size hypothesis, which states that engaging in section recall will yield greater retrieval during practice and in turn will enhance subsequent memory for lengthy text material. As expected, outcomes consistently showed that section recall versus whole text recall significantly increased recall during practice. However, contrary to the grain size hypothesis, the consistently sizeable recall advantages observed during practice ($d$s ranging from 1.02 to 1.87) did not persist on the final test ($d$s ranging from 0.03 to 0.65). Furthermore, the striking interaction was highly robust, manifesting with three different sets of materials, two different retention intervals, and four different final test formats.

Experiments 2-6 systematically explored methodological and theoretical explanations for the surprisingly fragile effect of grain size on text memory. Although a conclusive explanation for the fragility of the grain size effect remains elusive, our concerted effort to identify why the effect was not durable across a delay provides a significant contribution by ruling out several plausible accounts. Experiment 2 ruled out the explanation that elevated recall during practice was due to the section group recalling some information from working memory versus long-term memory. Experiment 3 ruled out the possibility that recall differences between the section group and the whole text group would emerge after a longer retention interval. Experiments 3 and 4 established that the effect was unlikely due to a particular text by showing that the same qualitative
pattern of recall emerged with other expository texts.

Concerning theoretical explanations, two sets of relative memorability analyses provided initial evidence against the DUDD hypothesis, which states that elevated recall during practice for the section group reflects retrieval of detail information that is subsequently lost across the delay. Surprisingly, RMA outcomes suggested that the section group was primarily gaining and then losing main ideas across the delay. Experiment 4 provided a more direct test of the DUDD hypothesis by using an expository text in which target sentences were specifically constructed to include main ideas, important details, and unimportant details. Outcomes disconfirmed the DUDD hypothesis, showing more definitively that the recall advantage for the section group during practice primarily reflected main ideas that were subsequently lost across the delay. Experiments 5 and 6 explored the integration hypothesis, which states that engaging in section recall during practice disrupts the integration of ideas across different text sections. Although results of Experiment 5 showed a numerical trend consistent with the integration hypothesis, such that recall of distant information was greater in the whole text group versus the section group, this trend did not replicate in Experiment 6. Furthermore, including instructions that encouraged cross-section integration resulted in no significant effects or interactions. In sum, although we were unable to resolve why the grain size effect was not durable across a delay we have considerably narrowed the field for future research by eliminating several plausible explanations.

Although the importance of falsification tends to be undervalued in the field of psychology more generally (Ferguson & Heene, 2012; Kerr, 1998; Ledgerwood &
Sherman, 2012; Pashler & Harris, 2012; Schmidt, 2009; Simmons, Nelson, & Simonsohn, 2011), it is a critical first step in establishing and exploring an effect. By way of notable example, in the seminal work on generation effects, Slamecka and Graf (1978, p. 592) unabashedly declared:

This is an empirically oriented article whose purpose is to report a set of simple experiments that establish the existence of a robust and interesting phenomenon. This phenomenon, called the generation effect, is robust in that it manifests itself across a variety of testing procedures, encoding rules, and other situational changes. It is interesting in that it does not seem to be easily or satisfactorily accommodated by any of the currently familiar explanatory notions. We expect that once the phenomenon is described in its initial form, it will be the subject of wider experimental analysis and will eventually become better understood… Findings constitute a reasonable start toward a delineation of the phenomena and provide the necessary empirical base from which more analytic experiments can be launched in the future.

Although they provided no theoretical explanation for the generation effect, their prediction that the seminal work would provide an important foundation for future investigation was borne out, given the extensive amount and impact of research that has since explored the phenomenon.

We certainly are not claiming that the grain size effect will have the same scope of impact as the seminal work on generation effects, but the grain size effect does merit further exploration given the theoretical and practical implications. Specifically, the current research has implications for current theories of testing effects. For example, the bifurcated distribution model (Kornell, Bjork, & Garcia, 2011) was originally forwarded to explain why advantages of retrieval practice versus restudy often are not apparent on an immediate test but emerge after a delay. The model assumes that the memory strength of a given item lies on a continuum and that the item is later recalled if its memory strength is above the recall threshold. Engaging in retrieval practice bifurcates the
distribution of items, such that memory strength increases substantially for items that are recalled versus not at all for items that are not recalled. Engaging in restudy produces a normal distribution of items, such that the memory strengths of all restudied items increase to the same extent (although to a lesser extent compared to items that are successfully recalled). Recall on an immediate test is assumed to be greater for restudy items versus retrieval practice items because the recall threshold is low enough for many of the restudied items to be recalled. However, the recall threshold is assumed to be higher on a delayed test. Given that successfully retrieved items versus restudied items were initially strengthened to a greater extent, recall on a delayed test is greater for retrieval practice items versus restudy items.

Although the bifurcated distribution model was originally forwarded to explain the relative advantages of retrieval practice versus restudy, its assumptions concerning the consequences of successful versus unsuccessful retrieval practice would arguably be expected to hold here as well. As stated, the model would assume that memory strength was increased for a greater number of idea units in the section group versus the whole text group given the significant differences between the two groups during practice recall. However, recall on the final test is inconsistent with the bifurcated distribution model. Specifically, the increase in memory strength for a greater number of idea units in the section group versus the whole text group did not result in enhanced recall across the delay.

More generally, other theories of testing effects are also not well equipped to account for the effects of retrieval practice with text material. For example, the
elaborative retrieval hypothesis (ERH) states that engaging in retrieval practice (versus restudy) activates information related to the cue en route to retrieval of the target, which is then encoded along with the cue and the successfully retrieved target in an elaborated trace. These connections are assumed to create additional retrieval routes that can be used later to access the target from memory (Carpenter, 2011). As currently instantiated, ERH does not fully account for the fragility of grain size effect. Other than assuming that activation of related information is more likely during retrieval practice than during restudy, ERH is silent on factors that would affect the extent of elaboration. As such, ERH would assume a similar degree of activation of related information during section versus whole text recall. Furthermore, because a greater number of idea units are successfully recalled during section recall, more text ideas should be elaboratively encoded during section versus whole text recall. Thus, ERH would reasonably make the same prediction as the grain size hypothesis—namely, that success should breed success. However, ERH could reasonably be extended to include assumptions about factors that moderate the degree of elaboration during retrieval practice. Of interest here, one possible moderator is the relative ease of the memory search during practice recall, with less activation of elaborative information during easier or less extensive memory searches. If so, retrieval was likely easier during section recall versus whole text recall, and thus less elaborative information may have been activated during section recall versus whole text recall. This idea is consistent with theories of text comprehension, which assume that integrating information from the text with general world knowledge results in greater memory for text material. Subsequent memory for the text may have
been impaired if section recall resulted in less integration with general world knowledge that would have been activated during more elaborative retrieval. Of course, this account is speculative, but it does provide an illustrative example of the implications of the current work for evaluating and further refining current theories of testing effects.

Concerning practical implications, the current research suggests a potential boundary condition on the robust effects of retrieval practice observed in the testing effect literature (Roediger & Butler, 2011; Roediger & Karpicke, 2006). Although engaging in section recall versus whole text recall substantially increased retrieval during practice, the grain size effect did not persist across a delay. One possibility is that the fluency and the amount of recall with a smaller grain size during practice may deceive students, such that they incorrectly believe that the information is well learned and will be retained across time. This possibility is consistent with research in the metacognitive literature showing that students give higher confidence judgments when information is accessed quickly and when more information is recalled (referred to as momentary accessibility, Baker & Dunlosky, 2006; Morris, 1990).

In sum, results conclusively establish an initially sizeable but subsequently fragile effect of grain size, for which an explanation remains elusive. The practical and theoretical implications of the current research emphasize the importance of further exploring the grain size effect on the memorial consequences of retrieval practice for lengthy text.
References


Footnotes

1. Given that the cued recall measures showed significant effects of other manipulations in Rawson and Kintsch (2002, 2004) the null results are likely not due to insensitivity of the measure.

2. We thank Larry Jacoby for this suggestion.
### TABLE 1. Methodological diagram of Experiment 1-6.

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Section</th>
<th>S1</th>
<th>T1</th>
<th>S2</th>
<th>T2</th>
<th>S3</th>
<th>T3</th>
<th>S4</th>
<th>T4</th>
<th>20m</th>
<th>FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole text</td>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>T1-4</td>
<td>20m</td>
<td>FT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment 2</th>
<th>Section</th>
<th>S1</th>
<th>T1</th>
<th>S2</th>
<th>T2</th>
<th>S3</th>
<th>T3</th>
<th>S4</th>
<th>T4</th>
<th>20m</th>
<th>FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole text</td>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>T1-4</td>
<td>20m</td>
<td>FT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section-math</td>
<td></td>
<td>S1 M</td>
<td>T1</td>
<td>S2 M</td>
<td>T2</td>
<td>S3 M</td>
<td>T3</td>
<td>S4 M</td>
<td>T4</td>
<td>20m</td>
<td>FT</td>
</tr>
<tr>
<td>Whole text-math</td>
<td></td>
<td>S1 M</td>
<td>S2 M</td>
<td>S3 M</td>
<td>S4 M</td>
<td>T1-4</td>
<td>20m</td>
<td>FT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment 3</th>
<th>Section</th>
<th>S1</th>
<th>T1</th>
<th>S2</th>
<th>T2</th>
<th>S3</th>
<th>T3</th>
<th>S4</th>
<th>T4</th>
<th>15m</th>
<th>FT</th>
<th>2d</th>
<th>FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole text</td>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>T1-4</td>
<td>15m</td>
<td>FT</td>
<td>2d</td>
<td>FT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section-delay</td>
<td></td>
<td>S1</td>
<td>T1</td>
<td>S2</td>
<td>T2</td>
<td>S3</td>
<td>T3</td>
<td>S4</td>
<td>T4</td>
<td>15m</td>
<td>2d</td>
<td>FT</td>
<td></td>
</tr>
<tr>
<td>Whole text-delay</td>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>T1-4</td>
<td>15m</td>
<td>2d</td>
<td>FT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment 4 &amp; 5</th>
<th>Section</th>
<th>S1</th>
<th>T1</th>
<th>S2</th>
<th>T2</th>
<th>S3</th>
<th>T3</th>
<th>S4</th>
<th>T4</th>
<th>15m</th>
<th>FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole text</td>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>T1-4</td>
<td>15m</td>
<td>FT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment 6</th>
<th>Section</th>
<th>S1</th>
<th>T1</th>
<th>S2</th>
<th>T2</th>
<th>S3</th>
<th>T3</th>
<th>S4</th>
<th>T4</th>
<th>15m</th>
<th>FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole text</td>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>T1-4</td>
<td>15m</td>
<td>FT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section-connection</td>
<td></td>
<td>S1</td>
<td>T1</td>
<td>S2</td>
<td>T2</td>
<td>S3</td>
<td>T3</td>
<td>S4</td>
<td>T4</td>
<td>15m</td>
<td>FT</td>
</tr>
<tr>
<td>Whole text-connection</td>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>T1-4</td>
<td>15m</td>
<td>FT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.  
S = study, T = free recall test, FT = final test, M = 2 minutes of math problems.  
Numbers refer to text section that was studied or recalled, and italics denote length of retention interval (m = minute, d = day).
TABLE 2. Slopes from linear regressions and \((R^2)\) values for relative memorability analyses for Experiments 1-6.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Practice Recall for Section vs. Whole Text</th>
<th>Recall for Section Group in Practice vs. Final Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>1.29 (.55)</td>
<td>1.27 (.67)</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>1.17 (.72)</td>
<td>1.57 (.76)</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>1.03 (.48)</td>
<td>1.32 (.48)</td>
</tr>
<tr>
<td>Experiment 4</td>
<td>1.04 (.22)</td>
<td>1.18 (.48)</td>
</tr>
<tr>
<td>Experiment 5</td>
<td>0.62 (.35)</td>
<td>N/A</td>
</tr>
<tr>
<td>Experiment 6</td>
<td>0.98 (.35)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Note. Experiment 2 values are collapsed across math groups. Experiment 3 values are collapsed across delay groups. Experiment 6 values are collapsed across connection groups.
Note For Table 3 (see next page). Mean values reflect number of idea units recalled.

SE = standard error of the mean. Near refers to recall of information from the same section of the text that contained the prompt sentence. Distant refers to recall of information from a section other than the one containing the prompt sentence.

Experiment 6 values are collapsed across connection groups.
TABLE 3. Cued recall as a function of distance in the text between prompt sentence and recalled information.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Near information recalled:</th>
<th></th>
<th>Distant information recalled:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Section</td>
<td>M</td>
<td>Whole text</td>
<td>M</td>
</tr>
<tr>
<td>Experiment 5</td>
<td></td>
<td>9.91</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Whole text</td>
<td>9.60</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>Experiment 6</td>
<td></td>
<td>7.37</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Whole text</td>
<td>5.33</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Experiment 5-6 collapsed</td>
<td></td>
<td>8.21</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Whole text</td>
<td>6.68</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Section</td>
<td>4.04</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Whole text</td>
<td>3.01</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Section</td>
<td>5.40</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Whole text</td>
<td>4.99</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>
FIGURE 1. Mean number of idea units correctly recalled during practice and final test for each group, Experiment 1. Error bars represent standard errors.
FIGURE 2. Mean number of idea units correctly recalled during practice and final test for each group, Experiment 2. Error bars represent standard errors.
FIGURE 3. Mean number of idea units correctly recalled during practice and final test for each group, Experiment 3. Error bars represent standard errors.
FIGURE 4. Hypothetical outcomes for relative memorability analyses evaluating the claims of the DUDD hypothesis. DUDD hypothesis predicts the pattern shown in Panel A (in which the recall advantage for the section group over whole text group during practice is primarily due to elevated recall of details) and in Panel D (in which the decline in recall in the section group from practice to final test primarily reflects the loss of details). Panels B-C and E-F illustrate other logical possibilities. See text for further details about these predictions.
FIGURE 5. Mean number of idea units correctly recalled during practice and final test for each group, Experiment 4. Error bars represent standard errors.
FIGURE 6. Mean percent of idea units recalled from target sentences during practice and final test for each group, Experiment 4. The left panel represents recall of main ideas, the middle panel represents recall of important details, and the right panel represents recall of unimportant details. Error bars represent standard errors.
FIGURE 7. Mean number of idea units correctly recalled during practice and final cued recall test for each group, Experiment 5. Error bars represent standard errors.
FIGURE 8. Mean number of idea units correctly recalled during practice and final cued recall test for each group, Experiment 6. Error bars represent standard errors.