REPLICATION

CrossMark

How Much Mightier Is the Pen than the Keyboard for Note-Taking? A Replication and Extension of Mueller and Oppenheimer (2014)

Kayla Morehead[1] (ID) · John Dunlosky[1] · Katherine A. Rawson[1]

Abstract

Many students use laptops to take notes in classes, but does using them impact later test performance? In a high-profile investigation comparing note-taking writing on paper versus typing on a laptop keyboard, Mueller and Oppenheimer (*Psychological Science, 25*, 1159–1168, 2014) concluded that taking notes by longhand is superior. We conducted a direct replication of Mueller and Oppenheimer (2014) and extended their work by including groups who took notes using eWriters and who did not take notes. Some trends suggested longhand superiority; however, performance did not consistently differ between any groups (experiments 1 and 2), including a group who did not take notes (experiment 2). Group differences were further decreased after students studied their notes (experiment 2). A meta-analysis (combining direct replications) of test performance revealed small (nonsignificant) effects favoring longhand. Based on the present outcomes and other available evidence, concluding which method is superior for improving the functions of note-taking seems premature.

Keywords Note-taking · Laptop · Longhand · eWriter · Encoding · Storage

Introduction

College students take a lot of notes, with over 95% reporting that they take notes while attending a lecture (Morehead et al. 2019; Palmatier and Bennett 1974; Peverly and Wolf 2019). Moreover, reviewing notes is one of the most popular techniques for preparing for classroom examinations (Blasiman et al. 2017; Gurung 2005; Karpicke et al. 2009; Nandagopal and Ericsson 2012). These two facts—students take notes and report relying on them while studying—highlight the importance of note-taking for student learning. That is,

---

✉ Kayla Morehead
  kmorehea@kent.edu

[1]   Department of Psychological Sciences, Kent State University, Kent, OH 44240, USA

students' success will be related to how well taking notes promotes learning and the degree to which studying notes allows students to better learn target material.

Given the importance of quality note-taking, researchers have investigated the degree to which various approaches to note-taking (1) improve learning while taking notes and (2) later support effective exam preparation. The former refers to whether taking notes improves students' retention of target material (without subsequent study) and has been called the *encoding function* of note-taking. The latter refers to whether notes taken during a lecture helps students learn that material while studying and has been called the *storage function* (Di Vesta and Gray 1972). In the present research, our major goal was to provide evidence concerning how different note-taking methods (e.g., longhand vs. typing) affect both functions of note-taking, with a particular focus on replicating and extending research reported by Mueller and Oppenheimer (2014). We consider the storage function further in experiment 2 where it is evaluated, and for the remainder of the Introduction, we largely focus on the encoding function, which is evaluated in both experiments.

To illustrate the encoding function of note-taking (for a general overview, see Peverly and Wolf 2019), consider outcomes reported by Mueller and Oppenheimer (2014), who had college students watch one of five video lectures and take notes either by longhand on paper or by typing on a laptop keyboard (henceforth referred to as longhand and laptop note-taking, for brevity). Students then completed distracter tasks for about 30 min. Afterwards, a test was administered that included factual and conceptual questions about the lectures. Although performance on the factual questions did not differ between the groups, performance on the conceptual questions was significantly greater for those who took notes longhand than by laptop; that is, a *longhand-superiority effect* occurred for the encoding function of note-taking.

Why might taking notes by longhand be superior to laptop note-taking? For one, because laptops offer access to other programs (e.g., games) and applications (e.g., e-mail), such distractions can decrease students' performance (Glass and Kang 2018), especially for the students who already struggle to learn the material (Carter et al. 2017; Patterson and Patterson 2017; Sana et al. 2013). Such a problem would not occur for those taking notes by longhand. For example, Ragan et al. (2014) observed students using laptops in classrooms and reported that students were only on task 37% of the time. The rest of the time, students were on social media or doing something else on the computer. Research on laptop use in the classroom has also reported higher final test performance for classes where laptops were not allowed in the classroom than for classes where they were (Carter et al. 2017; Patterson and Patterson 2017). Because students presumably study their notes in preparation for high-stakes exams, these effects may arise from either the storage or encoding function of note-taking.

Although this research suggests that laptops may not be as effective as longhand for note-taking because of their potential for distraction, it does not address whether laptops can support note-taking when distractions are removed. It is possible that laptops can be as effective for note-taking as longhand (or even more effective, e.g., a *laptop-superiority effect*) when there are no distractions. However, some evidence suggests longhand may still be superior. In particular, as compared to when children type, handwriting connects more visual and motor networks in the brain (James 2017; Vinci-Booher et al. 2016). Children also perceive letters differently after handwriting practice than typing practice (James 2017; James and Engelhardt 2012; Vinci-Booher et al. 2016). And, handwriting and letter perception appear to share neural substrates (James and Gauthier 2006). These results suggest that neural differences arise after handwriting versus typing practice, but most of the research involved children, and hence, the results may not entirely generalize to adult note-taking.

Another possible explanation for the longhand-superiority effect provided by Mueller and Oppenheimer (2014) is that students who use laptops merely transcribe the lecture without deeply processing the content. Consistent with this hypothesis, in their experiments, the number of words in the notes (word count) and the verbatim overlap between the lecture content and the notes were higher for those who took notes by laptop than longhand. These outcomes reported by Mueller and Oppenheimer (2014) inspired headlines from news media that illustrate the impact of their research: "A Learning Secret: Don't Take Notes with a Laptop" (from *Mind*), "Attention, Students: Put Your Laptops Away" (from *National Public Radio: Education*), and "To Remember a Lecture Better, Take Notes by Hand" (from *The Atlantic*). These headlines and resulting recommendations are provocative, but we believe conclusions about which method (if any) is superior for improving the functions of note-taking are premature for two reasons.

First, some laboratory studies have not demonstrated a longhand-superiority effect on the encoding function of note-taking. In particular, Luo et al. (2018) had students take notes on a lecture about educational measurement and then take a test afterwards during the same session. Test performance was higher for those who took notes using laptops, although this laptop-superiority effect was not significant. Bui et al. (2013, experiment 1) reported that test performance after taking notes while listening to a lecture was superior for those taking notes using a laptop, and this effect was statistically significant. Moreover, several studies have reported that word count, verbatim overlap, and number of idea units are higher for laptop groups than longhand groups (Bui et al. 2013; Fiorella and Mayer 2017; Luo et al. 2018; Mueller and Oppenheimer 2014) but differed in their conclusions about whether these results support the longhand-superiority effect (Mueller and Oppenheimer 2014) or the laptop-superiority effect (Bui et al. 2013).

Second, direct and independent replications of the empirical outcomes reported by Mueller and Oppenheimer (2014) have not yet been published. For instance, the methods used by Luo et al. (2018) and Bui et al. (2013) differed from those used by Mueller and Oppenheimer (2014) in a variety of ways (e.g., topic of lectures, kinds of test, et cetera). As argued by Simmons et al. (2011), direct replications are essential because conceptual replications do not restrict researchers to the initial set of design decisions and hence may produce noncomparable outcomes (see also, Simons 2014). For instance, the methods used in the studies by Mueller and Oppenheimer (2014) versus Luo et al. (2018) and Bui et al. (2013) may have included key differences that moderated whether a longhand- or laptop-superiority effect would occur.

Major Aims of the Present Research

Accordingly, our major aim was to conduct a direct replication of Mueller and Oppenheimer's (2014, experiments 1 and 2) comparison of longhand and laptop note-taking for the encoding function of note-taking. We conducted as close to a direct replication as possible, which included using the same lectures and test questions. Two minor differences included adding a few test questions (to equate for the number of possible points on the test for each video) and using just a subset of the videos (experiment 2 only). For an overview of all replications of Mueller and Oppenheimer (2014) within the present experiments, see the top half of Table 1. Most important, as in their study, we compared the longhand and laptop groups on performance for an *immediate* (same day) test both for factual and conceptual questions. Mueller and Oppenheimer (2014) also reported that word count and verbatim overlap were greater for

Table 1 Summary of effect size estimates separated by replications and extensions

	Dependent variable	Note-taking groups	Mueller and Oppenheimer		Present experiments	
			1	2	1	2
Replications	Factual immediate[1]	Longhand and laptop	0.07	0.15	0.45*	0.27
	Conceptual immediate	Longhand and laptop	0.34	0.40*	0.13	− 0.35
	Word count	Longhand and laptop	− 1.43**	− 1.11**	− 0.44*	− 0.58**
	Verbatim overlap	Longhand and laptop	− 0.95**	− 1.13**	− 0.44*	− 0.20
Extensions	Factual immediate	eWriter comparisons[2]	–	–	0.04	< 0.01
	Conceptual immediate	eWriter comparisons	–	–	0.01	0.01
	Word count	eWriter comparisons	–	–	0.07**	0.06**
	Verbatim overlap	eWriter comparisons	–	–	0.04*	0.01
	Factual delayed	Longhand, laptop, and eWriter	–	–	0.02	< 0.01
	Conceptual delayed	Longhand, laptop, and eWriter	–	–	< 0.01	< 0.01
	Factual relevance	Longhand, laptop, and eWriter	–	–	0.02	0.01
	Conceptual relevance	Longhand, laptop, and eWriter	–	–	< 0.01	0.01
	Idea units	Longhand, laptop, and eWriter	–	–	0.07**	0.03
	Efficiency rating	Longhand, laptop, and eWriter	–	–	0.01	0.01

For replications, effect size estimates are Cohen's ds, with positive values indicating longhand superiority. For extensions, effect size estimates are eta squared based on ANOVA outcomes. For significant ANOVA outcomes, see text for relevant follow-up tests

* significant at $p < 0.05$; ** significant at $p < 0.01$. For replications, p values are one-tailed

[1] Factual and conceptual immediate and delayed refer to performance on corresponding tests

[2] All comparisons involving eWriter groups represent extensions

participants taking notes by laptop than by longhand, so outcomes on these measures in the present research represent replications.

Taking Notes with eWriters As important, we extended their research in multiple ways, which are presented in the bottom half of Table 1. Two extensions—eWriters and individual-difference analyses—are common to experiments 1 and 2, so we discuss them here (we address other extensions in the introduction to each experiment). In both experiments, we evaluated the efficacy of another note-taking device—*eWriters*. Besides notebooks and laptops, newer technologies can be used for note-taking in classes. eWriters have become a viable alternative to taking longhand notes on paper. In the present research, we used Boogie Board® Sync eWriters (manufactured by Kent Displays, Inc.), which reproduce the experience of writing on paper but are electronic and hence can decrease paper waste. They are about the size of paper and include a stylus. As compared to laptops, eWriters are lighter in weight and easier to carry, and their only function is writing. They do not support other applications (e.g., e-mail, Web browser), so students cannot be distracted by other internet-based activities (e.g., checking e-mail) while taking notes in a classroom. Importantly for the present purposes, given that eWriters allow people to write as they would by hand, we expected students using

an eWriter to perform as well as those taking notes by longhand on paper. Even so, in a large-scale survey, fewer than 2% of college students indicated using eWriters of any kind to take notes (Morehead et al. 2019), so using this relatively novel device may have an unforeseeable impact on the quality or quantity of notes taken.

Individual Differences in Note-Taking We also conducted secondary analyses aimed at estimating the degree to which individual differences in various aspects of the notes were related to final test performance. These analyses were exploratory and could provide preliminary evidence for why a given note-taking method (if any) is superior. For instance, differences in verbatim overlap wherein laptop note-takers are more likely to take notes in a rote and verbatim manner have been implicated in producing the longhand-superiority effect (Luo et al. 2018; Mueller and Oppenheimer 2014). Our exploratory analyses also follow from the research by Peverly and colleagues (Peverly et al. 2007, 2013, 2014; Peverly and Sumowski 2012; Reddington et al. 2015). In a typical study, students watched a lecture and took notes, and the notes were coded for the main content from the lecture. Later, students were tested by writing an organized summary of the lecture (without their notes available), which was also scored for the main content. Across several different large-scale studies (Peverly et al. 2007, 2013, 2014; Peverly and Sumowski 2012; Reddington et al. 2015), a consistent outcome has been that individual differences in the quality of notes is predictive of subsequent performance on the test. To contribute to this growing body of research on individual differences in note-taking, we estimated the relationships between test performance and a variety of aspects of the notes: overall word count, verbatim overlap, and number of idea units.

As important, we computed a measure that corresponds to whether the notes taken align with content tapped by the test questions (which we refer to as test relevance). This measure of notes is most closely related to Peverly and colleagues' measure of note quality. In their case, the measure of note quality (i.e., overlap between notes and the main content from the lecture) was similar to the measure of quality test performance (i.e., how well the summaries captured the main content from the lecture). In many cases, such as in the present studies, all the content within a lecture is not subsequently tested, so the present measure of test relevance focuses just on the main content of lectures that was subsequently tested. Because note quality has predicted test performance in prior research, we expected that this measure of test relevance would positively correlate with test performance.

Finally, Peverly and colleagues have reported that transcription—or writing—speed predicts individual differences in the quality of notes. If this were the case for typing speed, then students who struggle to fluently type may take fewer (and perhaps poorer quality) notes and would be at a disadvantage if assigned to take notes with a laptop. To empirically evaluate this possibility, we measured student's typing speed and estimated the degree to which it predicted key outcome measures (e.g., quantity of notes, test performance) for those taking notes using a laptop.

Experiment 1

In experiment 1, along with the aims mentioned above, we also evaluated the effects of note-taking after a longer retention interval by including a delayed test of lecture content. Mueller and Oppenheimer (2014) used an immediate test that was administered during the same

session that notes were taken (and we included the same kind of immediate test for the replications) but not a delayed test.[1] We included a delayed test because sometimes the impact of a variable on test performance can differ between immediate and delayed tests. For example, the benefits of retrieval practice over restudy are often weaker (or reversed) when a final test is administered during the same session as practice versus delayed one or more days after practice (e.g., Johnson and Mayer 2009; Kornell et al. 2011; Toppino and Cohen 2009). Thus, in the present case, perhaps the note-taking method that initially supports the best performance will result in poorer performance on the delayed test. To evaluate this possibility in experiment 1, about half the participants completed the test during the first session (the replication of Mueller and Oppenheimer 2014). Then all participants returned 2 days later to complete the delayed test. Including a group who completed only the delayed test was necessary to obtain a better estimate of how note-taking methods influence delayed performance, because completing the first test during the initial session could have reactive effects on delayed performance. This reactive effect would be consistent with test-enhanced learning effects (Roediger and Karpicke 2006).

Method

Participants The experiment was approved by the university IRB. Participants were undergraduate students at a large Midwestern university who completed the experiment for course credit. Informed consent was obtained from all participants. To determine the sample size, we conducted a power analysis to obtain a power of 0.90 for detecting the focal effect (Cohen's $d = 0.76$) between longhand and laptop note-taking for the conceptual questions (from Mueller and Oppenheimer 2014, experiment 1). One hundred ninety-three people participated in the experiment. Five participants did not complete the experiment due to programming errors, and their data were removed. Seven participants did not return for the second session of the experiment, but their data from the first session were included in the analyses. One hundred forty-seven participants were female, and ages ranged from 18 to 44 (84% of participants were under the age of 22).

Design The experiment had a 3 (note-taking method: longhand, laptop, eWriter) × 5 (video) × 2 (test group: immediate and delayed, delayed only) between-participants design. Participants were randomly assigned to one of the groups. Sixty-three participants were in the longhand group (ns = 32 and 31, for the immediate-and-delayed test group and for the delayed test only group, respectively), 64 were in the laptop group (ns = 31 and 33), and 61 were in the eWriter group (ns = 31 and 30).

Materials We used the five TED talks (approximately 17 m each) from Mueller and Oppenheimer (2014). The content of these talks was meant to be interesting, but not common knowledge. For example, one talk was about how scientists are deciphering the Indus script, and another was about how the collaboration of ideas leads to progress. The mean speaking rate of all talks was 164.80 words per minute. Participants were randomly assigned to watch one of the five videos (as per Mueller and Oppenheimer 2014).

---

[1] Although Mueller and Oppenheimer (2014) used a delayed test in experiment 3, the delayed test results from the present experiment do not constitute a direct replication of their experiment 3 because they used different materials.

Participants were randomly assigned to a note-taking group. Some participants took notes longhand and were given a pen (or pencil) and two sheets of paper. Some took notes on a laptop in Microsoft Word. This group was allowed to use all functions available in Microsoft Word. Some took notes longhand on a Boogie Board® Sync eWriter. Participants wrote on the eWriter with a stylus. When they filled the screen, they pushed the *save* button and then the *erase* button to start a new blank screen. All files were saved on the device and accessed later on a computer by the experimenters. Even files that were erased were saved, so if a participant forgot to save a screen before erasing, their notes could still be accessed.

The tests included the original set of questions from Mueller and Oppenheimer (2014) and consisted of factual-recall (e.g., "The talk mentions three hypotheses about what the Indus script could represent. What are they?") and conceptual-application (e.g., "What is significant about the examples of the Indus script found in Mesopotamia?") short-answer questions. In the original set of questions, the number of points possible varied between videos; to equate them, we added additional test questions for some of the videos, so that the total number of points was 20 for each video. The experiment instructions, video transcripts, and test questions including answers are available on the Open Science Framework at the following link: https://osf.io/dyga5/?view_only=843c2187b4894 aefbfc6218b2d6eaed4.

Procedure Participants in the eWriter group were given brief instructions on how to use the eWriter and asked if they had questions. No participants reported problems using the eWriter. Participants completed the experiment in a cubicle with a computer, headphones, and their note-taking device. Participants were instructed to watch the videos and take notes as they normally would in class. They were then told that their notes would be taken away after watching the video, they would answer general questions, and then they would be tested on the video content. They then watched one of the TED talks on the computer with headphones on while taking notes using their assigned note-taking method (longhand, laptop, or eWriter). After watching the video, the experimenter took away their notes. Students were no longer allowed access to their notes, and the remainder of the experiment was completed on a computer.

Participants then made judgments about the quality of their notes; these judgments were not used by Mueller and Oppenheimer (2014). Although these judgments may have had a minor effect on outcomes, they were made by all groups and hence would not contribute to any group differences, so we do not discuss them further. Participants then completed distractor tasks for 30 min, including a measure of typing speed. For the typing speed measure, participants were shown a passage and given 1 min to type as much of the passage as they could, and the number of words typed in 1 min was calculated. Next, about half the participants completed the short answer test. Two days later, all participants returned to the lab to complete the delayed test, which consisted of the same questions used on the immediate test.

Results

Longhand and eWriter notes were transcribed into word documents for scoring (e.g., to measure word count, verbatim overlap, et cetera). Responses on the short answer tests were scored by two research assistants. Interrater reliability on a subset (14) of tests scored by both

assistants was high (Cronbach's alpha = 0.94). All discrepancies were discussed and resolved, then each assistant scored approximately half the remaining tests. Given that the immediate test was a replication and the delayed test was an extension, we conducted analyses separately for performance on the immediate and delayed tests. We also conducted separate analyses for performance on the factual and conceptual questions (as per Mueller and Oppenheimer 2014). For all analyses, alpha was set at 0.05.

We began the analyses of test performance by conducting omnibus analysis of variance (ANOVAs) that included all note-taking groups, video groups, and test groups (and were conducted separately for performance on factual and conceptual questions). After each omnibus ANOVA, we then conducted planned comparisons to address the replication. For all planned comparisons, we used one-tailed tests.

We then conducted secondary analyses using one-way ANOVAs comparing note-taking groups on word count, verbatim overlap, test relevance, idea units, and efficiency rating. For word count and verbatim overlap, we also conducted planned comparisons for the longhand and laptop groups, because these comparisons comprise a direct replication of Mueller and Oppenheimer (2014). For all other secondary analyses, we followed up significant main effects using Tukey's honest significant difference (HSD) test. For a summary of main outcomes separated by replications and extensions, see Table 1. After presenting outcomes for test performance and secondary measures, we then present correlational analyses relevant to the degree to which these secondary measures may account for individual differences in test performance.

Finally, for analyses of most measures in both experiments, we included video as a factor in the ANOVA. Because video was a nuisance variable and in all cases but one did not interact with other variables, we collapsed across video group in all the figures (as per Mueller and Oppenheimer 2014) to highlight the critical outcomes pertaining to note-taking method. The one significant three-way interaction was small, and the differences were inconsistent and not obviously interpretable, so we do not consider it further. Interested readers can contact the first author for outcomes conditionalized by video.
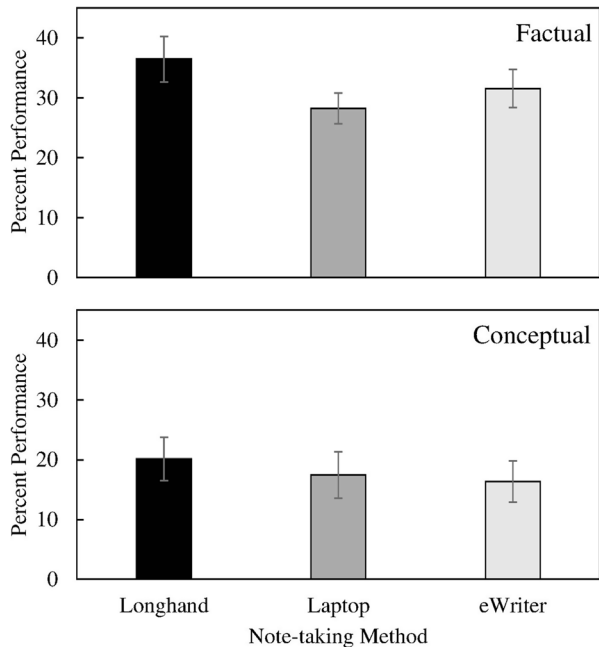
Immediate Test Performance

Factual Questions Performance on the immediate test for the factual questions is presented in the top panel of Fig. 1. We conducted a 3 (note-taking method) × 5 (video) ANOVA. The main effect of note-taking method was not significant, $F(2, 79) = 2.43$, MSE = 0.03, $p = 0.10$, $\eta^2 = 0.01$; the main effect of video was significant, $F(4, 79) = 6.10$, MSE = 0.03, $p < 0.001$, $\eta^2 = 0.05$ (performance was greater for the video about the Indus script than for the others). The interaction between method and video was not significant, $F(8, 79) = 1.12$, MSE = 0.03, $p = 0.36$, $\eta^2 = 0.02$.

Mueller and Oppenheimer (2014) reported no significant difference in performance between the two groups on factual questions, although performance was numerically higher for the longhand group. In the present experiment, a planned comparison revealed that performance was significantly greater for the longhand ($M = 0.36$, SE = 0.04) than laptop ($M = 0.28$, SE = 0.03) group, $t(61) = 1.79$, $p = 0.04$, Cohen's $d = 0.45$.[2]

---

[2] As stated in the Method section, we added questions to those used by Mueller and Oppenheimer (2014). To ensure that any different outcomes were not due to the new questions, we also conducted the planned comparisons based on performance for only test questions that were used in the original report. Conclusions were the same whether analyses were conducted on the entire question set (reported in the text) or the original questions (analyses available from the first author).

Fig. 1 Immediate factual and conceptual test performance for the three note-taking groups in experiment 1. Error bars are standard errors of the mean
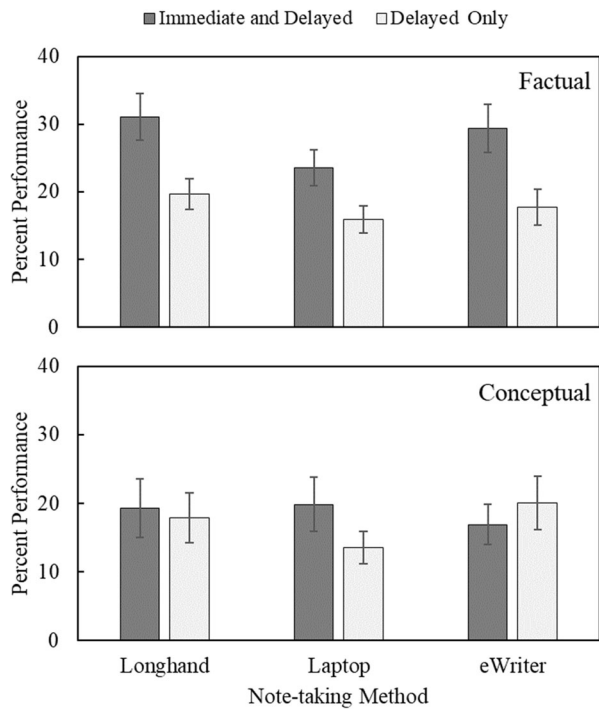

Conceptual Questions Performance on the immediate test for the conceptual questions is presented in the bottom panel of Fig. 1. The main effects of note-taking method, $F(2, 79) = 0.32$, MSE = 0.04, $p = 0.73$, $\eta^2 < 0.01$, and the interaction, $F(8, 79) = 1.10$, MSE = 0.04, $p = 0.37$, $\eta^2 = 0.05$, were not significant. The main effect of video was significant, $F(4, 79) = 3.85$, MSE = 0.04, $p < 0.01$, $\eta^2 = 0.08$ (performance was higher for the video about the collaboration of ideas than the other videos). In contrast to Mueller and Oppenheimer (2014), who reported significantly higher performance for the longhand group, the planned comparison revealed that performance did not significantly differ between the longhand ($M = 0.20$, SE = 0.04) and laptop ($M = 0.17$, SE = 0.04) groups, $t(61) = 0.50$, $p = 0.31$, $d = 0.13$.

Delayed Test Performance

Factual Questions Performance on the delayed test for the factual questions is presented in the top panel of Fig. 2 separated by test group. We conducted a 3 (note-taking method) × 5 (video) × 2 (test group) ANOVA. The main effect of note-taking method was not significant, $F(2, 151) = 2.55$, MSE = 0.02, $p = 0.08$, $\eta^2 = 0.01$. The main effects of video, $F(4, 151) = 11.68$, MSE = 0.02, $p < 0.001$, $\eta^2 = 0.06$, and test group, $F(1, 151) = 22.89$, MSE = 0.02, $p < 0.001$, $\eta^2 = 0.03$, were significant. Performance was higher for the video about the Indus script than the others. Performance was higher for those who took the immediate and delayed test ($M = 0.28$, SE = 0.02) than for those who only took the delayed test ($M = 0.18$, SE = 0.01). This effect did not interact with note-taking method, $F(2, 151) = 0.29$, MSE = 0.02, $p = 0.75$, $\eta^2 < 0.01$. The other interaction terms were also not significant: method by video, $F(8, 151) = 0.49$, MSE = 0.02, $p = 0.86$, $\eta^2 = 0.01$; video by test group, $F(4, 151) = 0.74$, MSE = 0.02, $p = 0.57$, $\eta^2 < 0.01$; and the three-way interaction, $F(8, 151) = 1.67$, MSE = 0.02, $p = 0.11$, $\eta^2 = 0.02$.

Fig. 2 Delayed factual and conceptual test performance for the three note-taking groups in experiment 1 separated by test group. Error bars are standard errors of the mean

Conceptual Questions Performance on the delayed test for the conceptual questions is presented in the bottom panel of Fig. 2 separated by test group. The main effects of note-taking method, $F(2, 151) = 0.19$, MSE = 0.03, $p = 0.83$, $\eta^2 < 0.01$, and test group, $F(1, 151) = 0.55$, MSE = 0.03, $p = 0.46$, $\eta^2 < 0.01$, were not significant. The main effect of video was significant, $F(4, 151) = 8.52$, MSE = 0.03, $p < 0.001$, $\eta^2 = 0.09$. Performance was higher for the video on the collaboration of ideas than the others. None of the two-way interactions were significant, $Fs < 1.0$, $\eta^2 < 0.01$, but the three-way interaction was significant, $F(8, 151) = 2.02$, MSE = 0.03, $p < 0.05$, $\eta^2 = 0.04$.

Secondary Measures and Individual-Differences Analyses

To further investigate differences between the note-taking groups, we conducted analyses of secondary measures, which are presented in the top half of Table 2. To explore potential predictors of performance outcomes, we then present the results of correlational analyses.

Word Count A one-way ANOVA across note-taking method revealed a significant effect, $F(2, 182) = 7.18$, MSE = 6073.87, $p < 0.01$, $\eta^2 = 0.07$. Consistent with the outcomes from Mueller and Oppenheimer (2014), a planned comparison revealed that more words were produced by the laptop than the longhand group, $t(122) = 2.43$, $p < 0.01$, $d = 0.44$. A post hoc test revealed that the laptop group produced significantly more words than did the eWriter group ($p < 0.01$) and the longhand group ($p < 0.05$), whereas the longhand and eWriter groups did not significantly differ ($p = 0.54$).

Table 2 Means for secondary outcome measures within note-taking method

		Word Count	Verbatim overlap	Test relevance		Idea units	Efficiency rating
				Factual	Conceptual		
Experiment 1	Longhand	133.90 (8.23)	4.20 (0.43)	39.08 (2.72)	44.13 (4.60)	10.34 (0.77)	15.41 (1.61)
	Laptop	170.73 (12.75)	6.50 (0.86)	33.12 (2.49)	45.93 (4.29)	14.21 (1.24)	13.77 (0.71)
	eWriter	119.07 (8.01)	4.50 (0.49)	36.26 (2.85)	41.18 (3.75)	9.30 (0.90)	16.21 (1.28)
Experiment 2	Longhand	118.88 (7.59)	5.41 (0.55)	42.49 (3.17)	44.49 (4.06)	12.73 (0.88)	10.05 (0.48)
	Laptop	156.05 (8.82)	6.20 (0.45)	38.78 (3.43)	52.08 (4.38)	15.25 (0.89)	10.91 (0.54)
	eWriter	126.78 (7.49)	5.27 (0.50)	35.71 (3.11)	45.00 (4.07)	12.76 (0.79)	11.06 (0.82)

Standard error of the mean is presented in parentheses. Word count is the number of words present in notes. Verbatim overlap is the percentage of content overlap between notes and lecture content. Test relevance is the percentage of test content that was present in the notes. Idea units are the number of idea units that are correct based on the video content. Efficiency rating was calculated by dividing word count by the number of idea units (see text for details)

Verbatim Overlap To compute the amount of verbatim overlap between notes and video content, we used an n-gram program that counts every overlapping three-word string between the notes and the video content as a match. It then reports the percentage of three-word strings in the notes that overlap with the video content. A one-way ANOVA across note-taking method revealed a significant effect, $F(2, 182) = 4.11$, MSE $= 0.002$, $p < 0.05$, $\eta^2 = 0.04$. Consistent with a key outcome from Mueller and Oppenheimer (2014), verbatim overlap was significantly greater for the laptop than the longhand group, $t(122) = 2.44$, $p < 0.01$, $d = 0.44$. A post hoc test revealed that the eWriter group did not significantly differ from either group (longhand: $p = 0.92$; laptop: $p = 0.07$).

Test Relevance We also analyzed notes by coding the amount of information targeted by test questions that were captured in participants' notes. For each test answer that was included in a participant's notes, one point was awarded. We then computed the percentage of points earned out of the number of possible points, which differed by video. Participants received two scores for test relevance: one for factual questions and one for conceptual questions. We conducted separate one-way ANOVAs across note-taking method for test relevance on factual and conceptual information. Neither effect was significant [factual test relevance, $F(2, 184) = 1.52$, MSE $= 0.04$, $p = 0.22$, $\eta^2 = 0.02$; conceptual test relevance, $F(2, 182) = 0.43$, MSE $= 0.11$, $p = 0.65$, $\eta^2 < 0.01$].

Idea Unit Analysis We scored notes by coding for the number of idea units. Each idea unit included in the notes that was correct based on information in the video was given one point. Five research assistants were trained on scoring and then scored notes from one of the five videos. The first author scored 20 notes (four per video) for reliability (Cronbach's alpha $= 0.94$). We then calculated an efficiency rating for each participant by calculating word count divided by idea units (Carter and Van Matre 1975; Kiewra 1985), which represents how many words a given participant used to produce an idea unit from a lecture. We conducted a separate one-way ANOVA across note-taking method for number of correct idea units and for efficiency rating. There was a significant difference in the number of correct idea units between note-taking method, $F(2, 182) = 6.82$, MSE $= 60.16$, $p < 0.01$, $\eta^2 = 0.07$. A post hoc analysis revealed that the number of idea units was significantly higher for the laptop group than the other two groups (longhand:

$p < 0.05$; eWriter: $p < 0.01$). Efficiency rating did not differ between note-taking methods, $F(2, 175) = 0.96$, MSE $= 95.18$, $p = 0.38$, $\eta^2 = 0.01$.

Correlational Analyses Toward exploring individual differences that may account for differences in test performance, we correlated test performance with the other relevant measures. Based on Mueller and Oppenheimer's (2014) outcomes, we expected a negative correlation between word count and test performance and between verbatim overlap and test performance. Table 3 presents correlational outcomes collapsed across groups (correlations computed separately for each note-taking method are presented in Appendix 1).

Consider some notable outcomes. Word count positively correlated with conceptual performance both on the immediate and delayed tests, whereas verbatim overlap did not correlate with performance. Factual test relevance correlated with factual test performance but did not correlate with conceptual test performance. Likewise, conceptual test relevance correlated with conceptual test performance but not with factual test performance. The number of idea units correlated with immediate test performance and delayed conceptual test performance and also strongly correlated with word count and correlated with both measures of test relevance.

To evaluate effects of typing speed on performance, we correlated typing speed with the other outcome measures only for the laptop group. The results are presented at the bottom of Table 8 in Appendix 1. Typing speed only correlated with the number of idea units and efficiency rating.

Discussion

Given the rich data set produced by the present experiment, we summarize here some of the more critical replications and extensions (see Table 1).

Replications Results for immediate test performance replicated the direction of the effect reported by Mueller and Oppenheimer (2014), in that performance for factual and conceptual

Table 3 Correlations between test performance and secondary outcome measures for experiment 1

Measure	1	2	3	4	5	6	7	8	9	10
1 Factual immediate	–									
2 Conceptual immediate	0.37**	–								
3 Factual delayed	0.84**	0.45**	–							
4 Conceptual delayed	0.14	0.53**	0.19**	–						
5 Word count	0.20	0.27**	0.01	0.23**	–					
6 Verbatim overlap	− 0.02	0.06	− 0.02	0.00	0.14	–				
7 Factual relevance	0.35**	0.04	0.21**	− 0.03	0.38**	− 0.09	–			
8 Conceptual relevance	0.16	0.26*	0.01	0.28**	0.52**	0.07	0.32**	–		
9 Idea units	0.27**	0.27**	0.04	0.19*	0.83**	0.09	0.39**	0.55**	–	
10 Efficiency rating	− 0.09	0.02	− 0.04	0.00	− 0.08	− 0.06	− 0.16*	− 0.18*	− 0.44*	–

* significant at $p < 0.05$; ** significant at $p < 0.01$

questions was numerically higher for the longhand than the laptop group. However, in the present experiment, the effect was significant for factual questions, but not conceptual, whereas Mueller and Oppenheimer (2014) reported the opposite. Regarding secondary outcomes, results of word count and verbatim overlap analyses were consistent with Mueller and Oppenheimer (2014) in that both were significantly higher for the laptop than the longhand group. However, given the apparent inconsistencies for some measures, we later present outcomes from a meta-analysis that combines across replications for each measure, so as to provide updated estimates of the impact of longhand versus laptop note-taking on these outcome measures.

Extensions Test performance for factual and conceptual questions (either immediate or delayed) did not significantly differ between the eWriter group and the other note-taking methods. Regarding secondary outcomes, word count was significantly higher for the laptop than the eWriter group, whereas the longhand and eWriter groups did not differ. Considering analyses of individual differences, the quantity (word count) and quality (number of idea units) of notes positively predicted test performance, consistent with prior research (Peverly and Sumowski 2012; Peverly et al. 2007). Moreover, factual test relevance positively correlated with factual test performance, and conceptual test relevance positively correlated with conceptual test performance. In experiment 2, we attempted to replicate these potentially important relationships between the content of notes and test performance.

Experiment 2

Results of experiment 1 numerically replicated Mueller and Oppenheimer (2014), but two critical aspects of note-taking in general were not evaluated. First, although we estimated the relative encoding benefits of longhand versus laptop note-taking, to evaluate whether note-taking itself has an encoding benefit, performance for the note-taking groups needs to be compared to performance for participants who do not take notes and just watch the video. To our knowledge, no one has compared laptop note-taking to a no-notes group. Moreover, in prior research, performance outcomes have not been consistently higher for longhand note-takers than for those who just listened to a lecture (Kiewra 1989; Kobayashi 2005), so outcomes from a no-notes group may have implications for interpreting any note-taking effects. For instance, if a longhand-superiority effect is found yet taking notes by longhand is no better than not taking notes, it would suggest that taking notes by laptop is detrimental to performance (and not that longhand note-taking improves it). Hence, in experiment 2, we included a group who listened to the video without taking notes.[3]

Second, the three note-taking methods may lead to differential storage benefits, and a subsequent study session is needed to evaluate this possibility. Other researchers have investigated storage benefits of longhand and laptop notes, but they have reported inconsistent results. Luo et al. (2018) and Mueller and Oppenheimer (2014, experiment 3) reported longhand-superiority effects when participants studied notes before the test, but

---

[3] Given that our initial aim was to replicate Mueller and Oppenheimer (2014), who did not include a no-notes group, we also did not include this group in experiment 1. We included it in experiment 2 because it could potentially offer insight into the overall encoding benefits of note-taking.

Fiorella and Mayer (2017, experiment 2) reported a laptop-superiority effect after study. To evaluate storage benefits in the present context, before the delayed test in experiment 2, all participants who took notes during the first session studied their notes before completing the test. If notes taken by longhand afford better storage benefits than notes taken by laptop, then performance on this second test will be superior for the longhand group (on paper and on eWriter) as compared to the laptop group.

Mueller and Oppenheimer (2014, experiment 3) investigated the storage function of note-taking and found a longhand-superiority effect, but they used different materials in their third experiment where they investigated these benefits. Hence, the present experiment does not provide a direct replication of this aspect of their research. Instead, we used the same methods as in the present experiment 1 to provide another direct replication of performance outcomes comparing longhand and laptop groups on the immediate test (and on word count and verbatim overlap) from Mueller and Oppenheimer (2014) and experiment 1 (see top half of Table 1).

Method

Participants Two hundred twenty-two undergraduate students at a large Midwestern university completed the experiment for course credit. Informed consent was obtained from all participants. Five participants did not complete the study due to a programming error so their data were removed. One student did not complete session 1 and another student did not complete session 2. Their data for the session they did complete were analyzed. One hundred eighty-seven participants were female, and ages ranged from 18 to 38 with 90% of participants under the age of 22.

Design The design was similar to experiment 1 except (a) we included a group who did not take notes and only completed the first part of the study (because they did not have notes to study before the delayed test, so it did not seem justifiable to compare their delayed test performance to the other note-taking groups), (b) only two of the videos were used, and (c) a study period was included before the delayed test. The experiment had a 4 (note-taking method: longhand, laptop, eWriter, no notes) × 2 (test group: immediate and delay, delay only) × 2 (video) between-participants design. Participants were randomly assigned to one of the groups. Sixty-one participants were in the longhand group ($ns = 31$ and 30, for the immediate-and-delayed test group and for the delayed test only group, respectively), 61 were in the laptop group ($ns = 31$ and 30), 63 were in the eWriter group ($ns = 32$ and 31), and 32 were in the no-notes group.

Materials Materials were the same as in experiment 1, but only two of the videos were used. Because overall test performance in experiment 1 was near floor (see Figs. 1 and 2), we selected the two videos with the highest mean test performance in experiment 1 ("When Ideas have Sex" and "Computing a Rosetta Stone for the Indus Script").

Procedure The procedure was similar to experiment 1 except for two differences. First, the no-notes groups did not take notes while watching the video, and hence, this group only completed the immediate test. Second, participants studied their notes for 7 min[4] before

---

[4] Mueller and Oppenheimer (2014, experiment 3) allowed students 10 min to study their notes. Most participants in the present study took only one or two pages of notes; thus, we expected that 7 min would be plenty of time for study, and no participants reported needing more time.

completing the delayed test and were not allowed to write on their notes during the study session. When participants came to the lab for session 2 of the experiment, they were given back their notes (which they had not had access to after the initial session) and told they would be allowed to study them for 7 min before being tested. Notes for the laptop and eWriter groups were printed out for study. Immediately after study, participants' notes were taken away, and they completed the delayed test.

Results

Results were analyzed as per the analytic plan used in experiment 1.

Immediate Test Performance: Encoding Benefits

With the exception of the no-notes groups, results pertaining to immediate test performance represent a direct replication of Mueller and Oppenheimer (2014) and experiment 1 (see top half of Table 1).

Factual Questions Performance on the immediate test for the factual questions is presented in the top panel of Fig. 3. We conducted a 4 (note-taking method) × 2 (video) ANOVA. The main effect of note-taking method was not significant, $F(3, 117) = 0.47$, $MSE = 0.04$, $p = 0.70$, $\eta^2 < 0.01$, whereas the main effect of video was significant, $F(1, 117) = 48.50$, $MSE = 0.04$, $p < 0.001$, $\eta^2 = 0.09$ (performance was higher for the video on the Indus script). The interaction was not significant, $F(3, 117) = 0.65$, $MSE = 0.04$, $p = 0.59$, $\eta^2 < 0.01$. Concerning the replication, a planned comparison indicated that performance did not significantly differ between the longhand ($M = 0.39$, $SE = 0.04$) and laptop ($M = 0.32$, $SE = 0.04$) groups, $t(59) = 1.07$, $p = 0.15$, $d = 0.28$, although the trend is consistent with the significant outcome from experiment 1.

Fig. 3 Immediate factual and conceptual test performance for the four note-taking groups in experiment 2. Error bars are standard errors of the mean
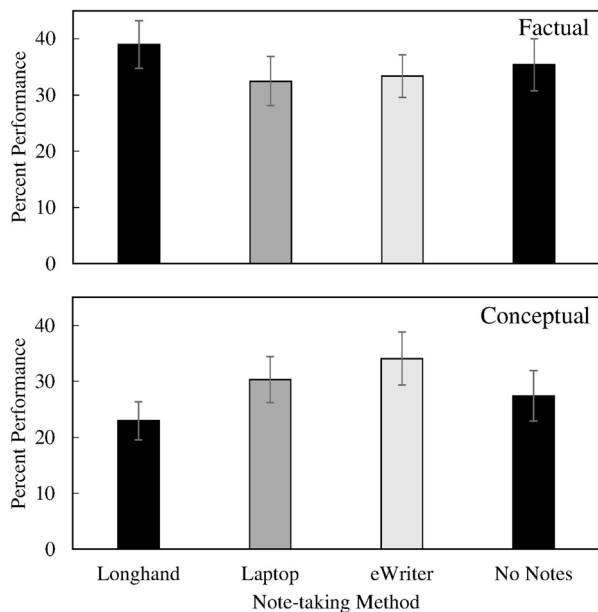
Conceptual Questions Performance on the immediate test for the conceptual questions is presented in the bottom panel of Fig. 3. The main effect of note-taking method was not significant, $F(3, 117) = 1.27$, $MSE = 0.05$, $p = 0.29$, $\eta^2 = 0.01$. The main effect of video was significant, $F(1, 117) = 6.25$, $MSE = 0.05$, $p < 0.05$, $\eta^2 = 0.02$ (performance was higher for the video on the collaboration of ideas). The interaction was not significant, $F(3, 117) = 1.31$, $MSE = 0.05$, $p = 0.28$, $\eta^2 = 0.01$. A planned comparison indicated that performance was not significantly different between the longhand ($M = 0.23$, $SE = 0.03$) and laptop ($M = 0.30$, $SE = 0.04$) groups, $t(59) = 1.37$, $p = 0.09$, $d = 0.35$. The trend is toward laptop superiority and hence does not replicate the significant longhand-superiority effect reported by Mueller and Oppenheimer (2014).

Delayed Test Performance: Storage Benefits

Unlike experiment 1, participants studied their notes before answering the test questions during session 2. Thus, performance on the delayed test represents the potential storage benefits of the three note-taking methods. The no-notes groups did not take the delayed test so they were not included in this analysis.

Factual Questions Performance on the delayed test for the factual questions collapsed across all participants (i.e., those who had taken the immediate test and those who had not) are presented in the top panel of Fig. 4 (given that all note-takers studied their notes prior to the delayed test, we collapsed outcomes across the two test groups in Fig. 4). We conducted a 3 (note-taking method) × 2 (video) × 2 (test group) ANOVA. The main effects

Fig. 4 Delayed factual and conceptual test performance for the three note-taking groups in experiment 2. Error bars are standard errors of the mean
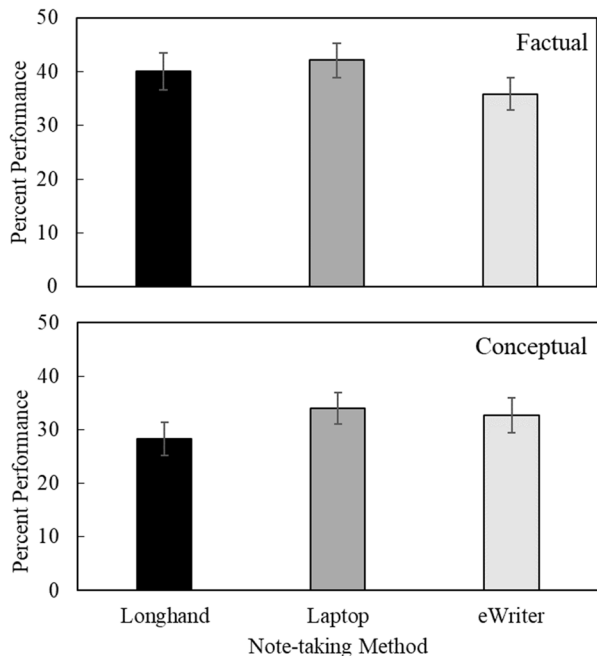
of note-taking method, $F(2, 168) = 1.62$, MSE = 0.04, $p = 0.20$, $\eta^2 < 0.01$, and test group, $F(1, 168) = 2.25$, MSE = 0.04, $p = 0.14$, $\eta^2 < 0.01$, were not significant. The main effect of video was significant, $F(1, 168) = 101.72$, MSE = 0.04, $p < 0.001$, $\eta^2 = 0.10$ (performance was higher for the video on the Indus script), but none of the interactions were significant: method by video, $F(2, 168) = 2.31$, MSE = 0.04, $p = 0.10$, $\eta^2 < 0.01$; video by test group, $F(1, 168) = 3.27$, MSE = 0.04, $p = 0.07$, $\eta^2 < 0.01$; method by test group, $F(2, 168) = 1.67$, MSE = 0.04, $p = 0.19$, $\eta^2 < 0.01$; and the three-way interaction, $F(2, 168) = 0.61$, MSE = 0.04, $p = 0.55$, $\eta^2 < 0.01$.

Conceptual Questions Performance on the delayed test for the conceptual questions is presented in the bottom panel of Fig. 4. The main effects of note-taking method, $F(2, 168) = 1.17$, MSE = 0.05, $p = 0.31$, $\eta^2 < 0.01$, and test group, $F(1, 168) = 1.28$, MSE = 0.05, $p = 0.26$, $\eta^2 < 0.01$, were not significant. The main effect of video was significant, $F(1, 168) = 36.17$, MSE = 0.05, $p < 0.001$, $\eta^2 = 0.06$ (performance was higher for the video on the collaboration of ideas), but none of the interactions were significant: note-taking method by test group, $F(2, 168) = 2.02$, MSE = 0.05, $p = 0.14$, $\eta^2 < 0.01$; method by video, $F(2, 168) = 0.89$, MSE = 0.05, $p = 0.41$, $\eta^2 < 0.01$; video by test group, $F(1, 168) = 1.39$, MSE = 0.05, $p = 0.24$, $\eta^2 < 0.01$; and the three-way interaction, $F(2, 168) = 0.48$, MSE = 0.05, $p = 0.62$, $\eta^2 < 0.01$. The outcomes indicate that the three methods did not produce differences in the storage benefits for either factual or conceptual questions.

Secondary Measures and Individual-Differences Analyses

For analyses involving secondary measures, the no-notes group was not included because they did not take notes. Outcomes for secondary measures separated by note-taking method are presented in the bottom half of Table 2.

Word Count A one-way ANOVA across note-taking method revealed a significant effect, $F(2, 175) = 6.04$, MSE = 3808.45, $p < 0.01$, $\eta^2 = 0.06$. Consistent with Mueller and Oppenheimer (2014) and experiment 1, more words were produced by the laptop than the longhand group, $t(117) = 3.18$, $p < 0.01$, $d = 0.58$. A post hoc test revealed that the laptop group produced significantly more words than the eWriter ($p < 0.05$) and the longhand ($p < 0.01$) groups, and the longhand and eWriter groups did not differ on word count ($p = 0.77$).

Verbatim Overlap A one-way ANOVA across note-taking method did not reveal a significant effect, $F(2, 175) = 1.01$, MSE = 0.001, $p = 0.37$, $\eta^2 = 0.01$. Unlike the results from Mueller and Oppenheimer (2014) and experiment 1, verbatim overlap did not significantly differ between the laptop and the longhand groups, $t(117) = 1.11$, $p = 0.14$, $d = 0.20$, but the numerical difference between groups was in the same direction.

Test Relevance We conducted separate one-way ANOVAs across note-taking method for test relevance on factual and conceptual information. As in experiment 1, neither effect was significant [factual test relevance, $F(2, 176) = 1.09$, MSE = 0.06, $p = 0.34$, $\eta^2 = 0.01$; conceptual test relevance, $F(2, 176) = 1.04$, MSE = 0.10, $p = 0.36$, $\eta^2 = 0.01$].

Idea Unit Analysis We scored notes for the number of idea units and computed an efficiency rating as in experiment 1. We conducted a separate one-way ANOVA across note-taking method for the number of correct idea units and for efficiency rating. Neither the number of correct idea units, $F(2, 176) = 2.88$, MSE $= 43.69$, $p = 0.06$, $\eta^2 = 0.03$, nor the efficiency rating, $F(2, 172) = 0.74$, MSE $= 23.24$, $p = 0.48$, $\eta^2 = 0.01$, significantly differed by note-taking method.

Correlational Analyses Table 4 presents correlational analyses collapsed across all groups (analyses separated by note-taking method are presented in Appendix 2). Unlike experiment 1, word count was positively correlated only with immediate conceptual performance, but not delayed test performance. Verbatim overlap was negatively correlated with every performance outcome except delayed conceptual performance. As in experiment 1, factual test relevance correlated positively with factual test performance, but did not positively correlate with conceptual test performance, whereas conceptual test relevance correlated positively with immediate conceptual test performance.

Finally, we correlated typing speed with the other outcome measures for only the laptop group. Typing speed did not significantly correlate with any outcome measure (see bottom of Table 11 in Appendix 2).

Discussion

Replications For a summary of the results of the replications, see the top half of Table 1. Immediate test performance for factual questions did not significantly differ between the longhand and laptop groups, and immediate test performance for conceptual questions did not significantly differ between the longhand and laptop group, which does not replicate the

Table 4 Correlations between test performance and secondary outcome measures for experiment 2

Measure	1	2	3	4	5	6	7	8	9	10
1 Factual immediate	–									
2 Conceptual immediate	0.23**	–								
3 Factual delayed	0.86**	0.00	–							
4 Conceptual delayed	0.01	0.81**	−0.01	–						
5 Word count	0.06	0.22*	0.11	0.07	–					
6 Verbatim overlap	−0.23*	−0.29**	−0.23**	−0.11	−0.25**	–				
7 Factual relevance	0.52**	−0.07	0.51**	−0.20**	0.38**	−0.21**	–			
8 Conceptual relevance	0.01	0.22*	0.05	0.11	0.55**	−0.19*	0.27**	–		
9 Idea units	0.12	0.18	0.15*	0.12	0.74**	−0.21**	0.28**	0.48**	–	
10 Efficiency rating	−0.01	0.08	0.00	−0.01	0.21**	−0.12	0.07	0.01	−0.40**	–

* significant at $p < 0.05$; ** significant at $p < 0.01$

significant effect reported by Mueller and Oppenheimer (2014). Regarding secondary outcomes, word count and verbatim overlap were higher for the laptop than the longhand group (although the trend for overlap was not significant), which replicates Mueller and Oppenheimer (2014).

Extensions Immediate test performance on factual and conceptual questions did not significantly differ between the eWriter group and the other note-taking groups. As important, the no-notes group performed as well (and sometimes numerically better) than the note-taking groups, suggesting no benefit of encoding for any note-taking method. Regarding the extension relevant to the storage function, delayed test performance for factual and conceptual questions did not significantly differ between the three note-taking methods.

Regarding secondary outcomes, word count was significantly greater for the laptop group than for the eWriter group and did not differ between the longhand and eWriter groups. Verbatim overlap did not significantly differ between the three note-taking methods, although it was higher for the laptop group than the other two groups. These results generally replicate outcomes from experiment 1. Considering the predictors of test performance, factual test relevance positively correlated with factual test performance, and conceptual test relevance positively correlated with only immediate conceptual test performance. The correlational outcomes were generally consistent with those reported in experiment 1 and highlight the potentially important role of test relevance in predicting criterion performance.

Combining Effect Estimates of Replications Across Experiments 1 and 2 and Mueller and Oppenheimer (2014)

Given that the major aim of the present experiments was the replication of the outcomes of Mueller and Oppenheimer (2014), we combined the available data with the aim of providing a better estimate of target effects. Toward this aim, we conducted a continuously cumulating meta-analysis (CCMA), which has been used to combine results of direct replications (Braver et al. 2014). In the CCMA, we combined direct replication outcomes from Mueller and Oppenheimer's (2014) experiments 1 and 2 with the corresponding outcomes from the present experiments. For each of the four experiments, we conducted an independent samples *t* test between the longhand and laptop groups on immediate test performance for factual questions, immediate test performance for conceptual questions, word count, and verbatim overlap.

To conduct the analysis, we used Excel software from Sibley (2008). The software calculates a *t* value, *p* value, Cohen's *d*, and *z*-score for each experiment and an overall *p* value, Cohen's *d*, *z*-score, *Q*-statistic, and 95% confidence interval (CI) across all experiments. Results of the CCMA for each of these outcomes are presented in Tables 5 and 6 (results for performance are presented in Table 5; results for word count and verbatim overlap are presented in Table 6). For test performance, there occurred a longhand-superiority effect, but effect size estimates were small and not significant for both the factual (95% CI = − 0.01 to 0.46) and conceptual (95% CI = − 0.07 to 0.40) questions. Word count and verbatim overlap were both substantially higher for the laptop

Table 5 Continuously cumulating meta-analysis (CCMA) outcomes for performance

Outcome measure	Experiment	Mean diff[1]	S_{pooled}	t	p (two-tailed)	Cohen's d	Z	Q
Factual performance	M & O experiment 1	0.02	0.24	0.29	0.77	0.07	0.29	–
	M & O experiment 2	0.04	0.23	0.77	0.44	0.15	0.76	–
	Present experiment 1	0.08	0.18	1.79	0.08	0.45	1.76	–
	Present experiment 2	0.07	0.24	1.07	0.29	0.27	1.06	–
	CCMA results				0.06	0.22	1.94	1.32
Conceptual performance	M & O experiment 1	0.08	0.23	1.39	0.17	0.34	1.37	–
	M & O experiment 2	0.10	0.26	1.97	0.05	0.40	1.95	–
	Present experiment 1	0.03	0.21	0.50	0.62	0.13	0.50	–
	Present experiment 2	−0.07	0.21	−1.37	0.18	−0.35	1.35	–
	CCMA results				0.16	0.17	2.58	5.83

[1] Positive mean differences indicate higher performance for the longhand group

than the longhand groups (word count: 95% CI = − 0.98 to − 0.57; verbatim overlap: 95% CI = − 0.79 to − 0.39).

General Discussion

Is the Pen Mightier than the Keyboard for Note-Taking?

Mueller and Oppenheimer (2014) answered this question with a resounding "yes." Indeed, the conclusion was part of the title of the article. With this conclusion in mind, consider the relevant outcomes from the present experiment, which involved independent and direct replications involving longhand and laptop groups. First, as per Mueller and Oppenheimer (2014), notes were longer and had more verbatim overlap for the laptop than the longhand group. Second, trends indicated the superiority of longhand over laptops on immediate test performance relevant to the encoding benefits of note-taking, but these trends were more apparent for the factual than conceptual questions (and the effect was even reversed for conceptual test performance in experiment 2, see bottom panel of Fig. 3). This pattern differs

Table 6 Continuously cumulating meta-analysis (CCMA) outcomes for word count and verbatim overlap

Outcome measure	Experiment	Mean diff[1]	S_{pooled}	t	p (two-tailed)	Cohen's d	Z	Q
Word count	M & O experiment 1	−136.23	95.30	−5.76	<0.001	−1.43	5.14	–
	M & O experiment 2	−104.94	94.66	−5.52	<0.001	−1.11	5.13	–
	Present experiment 1	−36.83	84.48	−2.43	0.02	−0.44	2.39	–
	Present experiment 2	−37.17	63.70	−3.18	<0.01	−0.58	3.11	–
	CCMA results				<0.001	−0.77	7.89	12.32
Verbatim overlap	M & O experiment 1	−0.06	0.06	−3.77	<0.001	−0.94	3.57	–
	M & O experiment 2	−0.05	0.05	−5.58	<0.001	−1.12	5.18	–
	Present experiment 1	−0.02	0.05	−2.40	0.02	−0.43	2.36	–
	Present experiment 2	−0.01	0.04	−1.11	0.27	−0.20	1.10	–
	CCMA results				<0.001	−0.59	6.13	12.97

[1] Negative mean differences indicate higher word count or verbatim overlap for the laptop group

from the one reported by Mueller and Oppenheimer (2014), who found a longhand-superiority effect for the conceptual questions (and not the factual ones) on an immediate test.

We do not view these minor discrepancies as necessarily problematic, because they represent independent attempts to estimate effects. Moreover, direct replications would not be expected to produce identical (and always significant) outcomes for every new experiment, especially when small effect sizes are involved (for detailed discussion of why effects will not always replicate even if they are real, see Francis 2012). Thus, perhaps most important, the combined outcomes from the CCMA support the following conclusions. First, although direct replications (using methods from Mueller and Oppenheimer 2014) yielded longhand-superiority effects, the effects relevant to the encoding function of note-taking were small and did not reach conventional levels of significance. Such outcomes do not support strong recommendations about whether students should take notes longhand or by laptop in class. Second, based on their evidence, Mueller and Oppenheimer (2014) hypothesized that the higher word count and verbatim overlap for laptop groups were in part responsible for diminished performance for laptop users. However, when combining results across the direct replications, differences in word count and verbatim overlap were large, whereas differences in performance were small (and not statistically significant). Thus, differences in word count and verbatim overlap do not appear sufficient to produce performance differences between longhand and laptop groups.

These outcomes and estimated effect sizes bring us back to a key applied question: Which method—longhand (on paper or eWriter) or laptop—should students use to take notes? At this point, we would argue that the available evidence does not provide a definitive answer to this question. Consider the mixed results from prior research. First, as discussed in the Introduction, the laptop-superiority effects reported by Bui et al. (2013) and Fiorella and Mayer (2017) should make one pause before recommending that laptops should not be used to take notes. In another study, longhand superiority was demonstrated when the criterion assessment was hand written, whereas the trend favored laptop note-taking when the assessment occurred on a computer (Barrett et al. 2014). Moreover, Luo et al. (2018) reported a laptop-superiority effect on the encoding function of note-taking and a longhand-superiority effect on the storage function (as did Mueller and Oppenheimer 2014, experiment 3). In sum, such mixed results run counter to making any general claims about which note-taking method is superior—factors that may moderate when a given method will be most effective need to be discovered.

Another reason we hesitate to offer general conclusions about which method is superior concerns results from the no-notes group (Fig. 3). In particular, performance on the immediate test (experiment 2) did not differ between those who took notes and those who did not. That is, all note-taking methods (laptop, longhand, and eWriter) examined in the present research had a minimal impact on the encoding function of note-taking. These outcomes reflect the note-taking literature, which has generally reported that note-taking has its largest influence on performance through its storage function (Kiewra 1989; Kobayashi 2005, 2006; Peverly and Wolf 2019). Also, as shown in Tables 3 and 4, one of the best predictors of test performance (either factual or conceptual) in the present experiments was whether students had included the lecture content relevant to the test questions (factual or conceptual, respectively) in their notes. These correlational outcomes are consistent with those from Peverly and colleagues (for a review, see Peverly and Wolf 2019) who have consistently shown that the amount of main content in notes is predictive of how much main content will later be remembered

on the criterion test. Taken together, these observations suggest that a key focus for future research should be to understand the degree to which a particular note-taking method increases the likelihood that students include the most important and to-be-tested content in their notes.

Which method supports the most effective note-taking (in terms of supporting the storage function) may also be moderated by several factors. For example, Fiorella and Mayer (2017) and Luo et al. (2018) reported that people included more images in their notes when they took longhand than laptop notes. Thus, if the to-be-learned material includes images that need illustrating, longhand may allow students to take more complete notes and subsequently gain more from studying them. Such moderators require systematic exploration as the field works toward prescriptions about when each note-taking method will be most effective.

Do eWriters Support Effective Note-Taking?

As more technology is introduced in the classroom, it is important to understand how these technologies affect learning outcomes. In the present research, we evaluated the use of eWriters because of some of their practical advantages for note-taking. Students who used eWriters tended to perform as well as those who took notes using other methods. Perhaps most relevant, consider the quantity and quality of the notes generated by students (Table 2), as described by word count, verbatim overlap, test relevance, and number of idea units. Outcomes from those taking notes on an eWriter most closely paralleled outcomes from those taking notes by longhand on paper, indicating that using eWriters to take notes mimics taking notes on paper. Although further research is needed to more fully evaluate the use of eWriters as effective note-taking devices, the preliminary evidence presented here indicates that they offer a viable alternative to longhand note-taking on paper.

Constraints on Generality

Simons et al. (2017) proposed that all primary research articles should include a statement specifying the constraints on generality. Given the relevance of note-taking for educational outcomes and that relatively few investigations have compared laptop to longhand note-taking, we agree that acknowledging these constraints is relevant for informing students and instructors and for guiding future research. We consider three constraints here.

Lecture Content We thank Mueller and Oppenheimer (2014) for sharing their materials (TED talks), which we needed to attempt a direct replication of their research. However, these materials differ from typical educational lectures in ways that may affect the relationship between note-taking and test performance. For example, many teachers construct their lectures with the aim of helping students learn critical content, and in doing so, teachers may emphasize main points. By contrast, although TED talks may be informative, they are often presented to entertain an audience, and in some cases, it is unclear which information would be most important for a subsequent test. Of course, it will also be valuable to understand note-taking efficacy using poorly constructed lectures because some teachers ramble and do not always reveal their main objectives.

In summary, one possible constraint is that conclusions from note-taking research involving entertainment-focused talks may not generalize to well-constructed lectures aimed at helping students learn assigned content. With this conclusion in mind, however, Luo et al. (2018) conducted a laboratory experiment in which they used materials that arguably better represent a standard lecture (in their case, one on educational measurement), and test performance relevant to the encoding function of note-taking did not differ significantly for the longhand and laptop groups. Thus, consistent with the small effect sizes from the present CCMAs, a general conclusion may be that in laboratory studies, the method of note-taking has a minor impact on the encoding function.

Distractions We aimed to compare laptops and longhand solely as note-taking devices, so we did not allow participants to access anything on the laptops besides the document they used to take notes (as in Mueller and Oppenheimer 2014). As stated in the Introduction, such access to other programs and applications on laptops may distract students and decrease their performance (Patterson and Patterson 2017; Sana et al. 2013). Thus, even if laptops themselves are effective note-taking devices, the distractions they encourage may decrease performance outcomes. Most important, the small longhand-superiority effects (reported here and in Mueller and Oppenheimer 2014) may not generalize to classroom settings, where such superiority may be larger due to dual-tasking that can occur when students take notes on laptops in classrooms.

Classroom Versus Laboratory Research Besides distractions, other differences between classroom and laboratory research may influence the generality of conclusions. Consider two here. First, in classrooms, students often decide how they want to take notes, which may affect the benefits of different note-taking methods. For example, students may choose to use a laptop because they write slowly or their handwriting is difficult to read. Alternatively, they may choose to handwrite notes because the lecture materials include symbols, figures, or illustrations. Second, students do not just take notes to learn content in classes but also rely on other activities to support learning, such as completing homework, studying flashcards, and completing study guides. These other activities may decrease the effects of note-taking by compensating for any initial differences in the encoding function.

To date, only a few studies have investigated the impact of laptop note-taking on performance in the classroom, and those reported higher final test performance when laptop use was not allowed in the classroom than when it was (Carter et al. 2017; Patterson and Patterson 2017). Even so, Carter et al. (2017) stated that the economics course in which their research was conducted consisted of material that was graphical or consisted of equations, which may provide an advantage to longhand note-taking because taking longhand notes can afford the use of spatial strategies (Fiorella and Mayer 2017; Luo et al. 2018). Further exploring plausible factors that could moderate the relative benefits of laptop versus longhand note-taking comprises an important avenue for future research.

Appendix 1. Correlation Tables Within the Note-Taking Method Group for Experiment 1

Table 7 Longhand group correlations between test performance and secondary outcome measures for experiment 1

Measure		1	2	3	4	5	6	7	8	9	10	11
1	Factual immediate	–										
2	Conceptual immediate	0.25	–									
3	Factual delayed	0.83**	0.36*	–								
4	Conceptual delayed	0.03	0.41*	0.15	–							
5	Word count	0.10	− 0.02	− 0.04	0.07	–						
6	Verbatim overlap	0.22	0.32	0.13	0.23	− 0.27*	–					
7	Factual relevance	0.28	− 0.14	0.21	− 0.17	0.32*	− 0.13	–				
8	Conceptual relevance	0.02	0.24	− 0.02	0.25*	0.47**	0.04	0.29*	–			
9	Idea units	0.37*	0.18	0.20	0.11	0.64**	− 0.06	0.41**	0.57**	–		
10	Efficiency rating	− 0.25	− 0.13	− 0.21	0.07	0.22	− 0.15	− 0.08	− 0.07	− 0.37**	–	
11	Typing speed	0.40*	0.21	0.17	0.19	0.12	0.04	0.08	0.05	0.13	0.18	–

* significant at $p < 0.05$; ** significant at $p < 0.01$

Table 8 Laptop group correlations between test performance and secondary outcome measures for experiment 1

Measure		1	2	3	4	5	6	7	8	9	10	11
1	Factual immediate	–										
2	Conceptual immediate	0.37*	–									
3	Factual delayed	0.77*	0.48**	–								
4	Conceptual delayed	0.43*	0.81**	0.44**	–							
5	Word count	0.44*	0.48**	0.07	0.36**	–						
6	Verbatim overlap	− 0.19	− 0.23	− 0.02	− 0.08	0.37**	–					
7	Factual relevance	0.45*	0.21	0.11	0.18	0.49**	0.08	–				
8	Conceptual relevance	0.40*	0.41*	0.07	0.32*	0.62**	0.22	0.34**	–			
9	Idea units	0.42*	0.36	0.08	0.33*	0.90**	0.23	0.51**	0.57**	–		
10	Efficiency rating	− 0.13	− 0.14	− 0.12	− 0.20	− 0.30*	− 0.03	− 0.21	− 0.15	− 0.60**	–	
11	Typing speed	0.31	0.18	0.13	0.11	0.19	− 0.14	0.10	0.03	0.33**	− 0.29*	–

* significant at $p < 0.05$; ** significant at $p < 0.01$

Table 9 eWriter group correlations between test performance and secondary outcome measures for experiment 1

Measure	1	2	3	4	5	6	7	8	9	10	11
1 Factual immediate	–										
2 Conceptual immediate	0.55**	–									
3 Factual delayed	0.90**	0.53**	–								
4 Conceptual delayed	0.06	0.31	0.05	–							
5 Word count	0.24	0.35	0.10	0.36**	–						
6 Verbatim overlap	−0.13	0.10	−0.02	−0.06	−0.34**	–					
7 Factual relevance	0.36*	0.08	0.24	−0.05	0.48**	−0.25	–				
8 Conceptual relevance	0.15	0.11	0.03	0.29*	0.49**	−0.25	0.36**	–			
9 Idea units	0.23	0.28	0.01	0.21	0.84**	−0.26*	0.40**	0.57*	–		
10 Efficiency rating	0.13	0.34	0.16	−0.01	−0.25	0.07	−0.28*	−0.41**	−0.55**	–	
11 Typing speed	−0.02	−0.11	0.16	0.00	0.10	−0.12	0.31*	0.00	0.01	−0.08	–

* significant at $p < 0.05$; ** significant at $p < 0.01$

Appendix 2. Correlation Tables Within Note-Taking Method Group for Experiment 2

Table 10 Longhand group correlations between test performance and secondary outcome measures for experiment 2

Measure	1	2	3	4	5	6	7	8	9	10	11
1 Factual immediate	–										
2 Conceptual immediate	0.03	–									
3 Factual delayed	0.87**	−0.16	–								
4 Conceptual delayed	−0.18	0.75**	−0.22	–							
5 Word count	0.34	0.38*	0.13	0.22	–						
6 Verbatim overlap	−0.19	−0.50**	−0.22	0.00	−0.26*	–					
7 Factual relevance	0.61**	−0.08	0.60**	−0.23	0.40**	−0.13	–				
8 Conceptual relevance	0.13	0.35	0.05	0.15	0.54**	−0.14	0.27*	–			
9 Idea units	0.32	0.18	0.13	0.20	0.78**	−0.24	0.31*	0.37**	–		
10 Efficiency rating	−0.07	0.26	0.04	−0.14	0.16	−0.03	0.23	0.29*	−0.44**	–	
11 Typing speed	0.31	−0.08	0.10	−0.10	0.20	−0.02	0.14	0.10	0.18	−0.13	–

* significant at $p < 0.05$; ** significant at $p < 0.01$

Table 11 Laptop group correlations between test performance and secondary outcome measures for experiment 2

Measure		1	2	3	4	5	6	7	8	9	10	11
1	Factual immediate	–										
2	Conceptual immediate	0.19	–									
3	Factual delayed	0.90**	0.09	–								
4	Conceptual delayed	0.07	0.81**	−0.06	–							
5	Word count	0.09	0.21	0.14	0.03	–						
6	Verbatim overlap	−0.16	−0.14	−0.29*	−0.12	−0.28*	–					
7	Factual relevance	0.42**	0.02	0.51**	−0.23	0.51**	−0.13	–				
8	Conceptual relevance	0.00	0.23	0.03	0.16	0.56**	−0.20	0.31*	–			
9	Idea units	0.29	0.20	0.19	0.08	0.70**	−0.17	0.36**	0.50**	–		
10	Efficiency rating	−0.05	0.04	0.02	0.12	0.31*	−0.28*	0.17	0.07	−0.36**	–	
11	Typing speed	−0.15	0.12	−0.12	0.07	−0.06	0.15	0.02	0.03	−0.03	−0.07	–

* significant at $p < 0.05$; ** significant at $p < 0.01$

Table 12 eWriter correlations between test performance and secondary outcome measures for experiment 2

		1	2	3	4	5	6	7	8	9	10	11
1	Factual immediate	–										
2	Conceptual immediate	0.22	–									
3	Factual delayed	0.79**	0.11	–								
4	Conceptual delayed	0.14	0.83**	0.23	–							
5	Word count	−0.21	−0.01	0.03	−0.05	–						
6	Verbatim overlap	−0.30	−0.19	−0.22	−0.24	−0.30*	–					
7	Factual relevance	0.55**	−0.13	0.41**	−0.13	0.27*	−0.39**	–				
8	Conceptual relevance	−0.04	0.11	0.06	0.00	0.55**	−0.25	0.25	–			
9	Idea units	−0.23	0.12	0.10	0.05	0.74**	−0.28*	0.15	0.53**	–		
10	Efficiency rating	0.09	−0.02	−0.04	−0.04	0.17	−0.09	−0.06	−0.20	−0.47**	–	
11	Typing speed	0.14	0.05	0.27*	0.02	0.07	−0.14	0.10	−0.06	0.13	0.05	–

* significant at $p < 0.05$; ** significant at $p < 0.01$

References

Barrett, M. E., Swan, A. B., Mamikonian, A., Ghajoyan, I., Kramarova, O., & Youmans, R. J. (2014). Technology in note taking and assessment: the effects of congruence on student performance. *International Journal of Instruction, 7*, 49–58.

Blasiman, R., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: comparing intended versus actual study behavior. *Memory, 25*, 784–792. https://doi.org/10.1080/09658211.2016.1221974.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*, 333–342. https://doi.org/10.1177/1745691614529796.

Bui, D. C., Myerson, J., & Hale, S. (2013). Note-taking with computers: Exploring alternative strategies for improved recall. *Journal of Educational Psychology, 105*, 299–309. https://doi.org/10.1037/a0030367.

Carter, J. F., & Van Matre, N. H. (1975). Note taking versus note having. *Journal of Educational Psychology, 67*, 900–904. https://doi.org/10.1037/0022-0663.67.6.900.

Carter, S. P., Greenberg, K., & Walker, M. S. (2017). The impact of computer usage on academic performance: evidence from a randomized trial at the United States Military Academy. *Economics of Education Review, 56*, 118–132. https://doi.org/10.1016/j.econedurev.2016.12.005.

Di Vesta, F. J., & Gray, G. S. (1972). Listening and note taking. *Journal of Educational Psychology, 63*, 8–14. https://doi.org/10.1037/h0032243.

Fiorella, L., & Mayer, R. E. (2017). Spontaneous spatial strategy use in learning from scientific text. *Contemporary Educational Psychology, 49*, 66–79. https://doi.org/10.1016/j.cedpsych.2017.01.002.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19*, 975–991. https://doi.org/10.3758/s13423-012-0322-y.

Glass, A. L., & Kang, M. (2018). Dividing attention in the classroom reduces exam performance. *Educational Psychology.* 1–14. On-line first publication. https://doi.org/10.1080/01443410.2018.1489046.

Gurung, R. A. (2005). How do students really study (and does it matter)? *Education, 39*, 323–340.

James, K. H. (2017). The importance of handwriting experience on the development of the literate brain. *Current Directions in Psychological Science, 26*, 502–508. https://doi.org/10.1177/0963721417709821.

James, K. H., & Engelhardt, L. (2012). The effects of handwriting experience on functional brain development in pre-literate children. *Trends in Neuroscience and Education, 1*, 32–42. https://doi.org/10.1016/j.tine.2012.08.001.

James, K. H., & Gauthier, I. (2006). Letter processing automatically recruits a sensory-motor brain network. *Neuropsychologia, 44*, 2937–2949. https://doi.org/10.1016/j.neuropsychologia.2006.06.026.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*, 621–629. https://doi.org/10.1037/a0015183.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: do students practice retrieval when they study on their own? *Memory, 17*, 471–479. https://doi.org/10.1080/09658210802647009.

Kiewra, K. A. (1985). Students' note-taking behaviors and the efficacy of providing the instructor's notes for review. *Contemporary Educational Psychology, 10*, 378–386. https://doi.org/10.1016/0361-476X(85)90034-7.

Kiewra, K. A. (1989). A review of note-taking: the encoding-storage paradigm and beyond. *Educational Psychology Review, 1*, 147–172. https://doi.org/10.1007/BF01326640.

Kobayashi, K. (2005). What limits the encoding effect of note-taking? A meta-analytic examination. *Contemporary Educational Psychology, 30*, 242–262. https://doi.org/10.1016/j.cedpsych.2004.10.001.

Kobayashi, K. (2006). Combined effects of note-taking/reviewing on learning and the enhancement through interventions: a meta-analytic review. *Educational Psychology, 26*, 459–477. https://doi.org/10.1080/01443410500342070.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: a distribution-based bifurcation model. *Journal of Memory and Language, 65*, 85–97. https://doi.org/10.1016/j.jml.2011.04.002.

Luo, L., Kiewra, K. A., Flanigan, A. E., & Peteranetz, M. S. (2018). Laptop versus longhand note taking: effects on lecture notes and achievement. *Instructional Science, 46*, 1–25. https://doi.org/10.1007/s11251-018-9458-0.

Morehead, K., Dunlosky, J., Rawson, K. A., Blasiman, R., & Hollis, R. B. (2019). Note-taking habits of 21st century college students: implications for student learning, memory, and achievement. *Memory.* https://doi.org/10.1080/09658211.2019.156969.

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: advantages of longhand over laptop note taking. *Psychological Science, 25*, 1159–1168. https://doi.org/10.1177/0956797614524581.

Nandagopal, K., & Ericsson, K. A. (2012). An expert performance approach to the study of individual differences in self-regulated learning activities in upper-level college students. *Learning and Individual Differences, 22*, 597–609. https://doi.org/10.1016/j.lindif.2011.11.018.

Palmatier, R. A., & Bennett, J. M. (1974). Notetaking habits of college students. *Journal of Reading, 18*, 215–218. http://www.jstor.org/stable/40009958. Accessed 12 April 2016.

Patterson, R. W., & Patterson, R. M. (2017). Computers and productivity: evidence from laptop use in the college classroom. *Economics of Education Review, 57*, 66–79. https://doi.org/10.1016/j.econedurev.2017.02.004.

Peverly, S. T., & Sumowski, J. F. (2012). What variables predict quality of text notes and are text notes related to performance on different types of tests? *Applied Cognitive Psychology, 26*, 104–117. https://doi.org/10.1002/acp.1802.

Peverly, S. T., & Wolf, A. D. (2019). Note-taking. To appear in J. Dunlosky & K. A. Rawson (Eds.), *Cambridge handbook of cognition and education* (pp. 320–355). New York, NY: Cambridge University Press.

Peverly, S. T., Sumowski, J. F., & Garner, J. (2007). Skill in lecture note-taking: what predicts? *Journal of Educational Psychology, 99*, 167–180. https://doi.org/10.1037/0022-0663.99.1.167.

Peverly, S. T., Vekaria, P. C., Reddington, L. A., Sumowski, J. F., Johnson, K. R., & Ramsay, C. M. (2013). The relationship of handwriting speed, working memory, language comprehension and outlines to lecture note-taking and test-taking among college students. *Applied Cognitive Psychology, 27*, 115–126. https://doi.org/10.1002/acp.2881.

Peverly, S. T., Garner, J. K., & Vekaria, P. C. (2014). Both handwriting speed and selective attention are important to lecture note-taking. *Reading and Writing: An Interdisciplinary Journal, 27*, 1–30. https://doi.org/10.1007/s11145-013-9431-x.

Ragan, E. D., Jennings, S. R., Massey, J. D., & Doolittle, P. E. (2014). Unregulated use of laptops over time in large lecture classes. *Computers & Education, 78*, 78–86. https://doi.org/10.1016/j.compedu.2014.05.002.

Reddington, L. A., Peverly, S. T., & Block, C. J. (2015). An examination of some of the cognitive and motivation variables related to gender differences in lecture note-taking. *Reading and Writing: An Interdisciplinary Journal, 28*, 1155–1185. https://doi.org/10.1007/s11145-015-9566-z.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x.

Sana, F., Weston, T., & Cepeda, N. J. (2013). Laptop multitasking hinders classroom learning for both users and nearby peers. *Computers & Education, 62*, 24–31. https://doi.org/10.1016/j.compedu.2012.10.003.

Sibley, C. G. (2008). *Utilities for examining simple meta-analytic avergages [computer software]*. Auckland: University of Auckland.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. https://doi.org/10.1177/0956797611417632.

Simons, D. J. (2014). The value of direct replications. *Perspectives on Psychological Sciences, 9*, 76–80. https://doi.org/10.1177/1745691613514755.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): a proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*, 1123–1128. https://doi.org/10.1177/1745691617708630.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval. *Experimental Psychology, 56*, 252–257. https://doi.org/10.1027/1618-3169.56.4.252.

Vinci-Booher, S., James, T. W., & James, K. H. (2016). Visual-motor functional connectivity in preschool children emerges after handwriting experience. *Trends in Neuroscience and Education, 5*, 107–120. https://doi.org/10.1016/j.tine.2016.07.006.