



Don't Ditch the Laptop Just Yet: A Direct Replication of Mueller and Oppenheimer's (2014) Study 1 Plus Mini Meta-Analyses Across Similar Studies



Psychological Science
1–14
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797620965541
www.psychologicalscience.org/PS


Heather L. Urry¹, Chelsea S. Crittle¹, Victoria A. Floerke¹, Michael Z. Leonard¹, Clinton S. Perry, III¹, Naz Akdilek¹, Erica R. Albert¹, Avram J. Block¹, Caroline Ackerley Bollinger¹, Emily M. Bowers¹, Renee S. Brody¹, Kelly C. Burk¹, Ally Burnstein¹, Allissa K. Chan¹, Petrina C. Chan¹, Lena J. Chang¹, Emily Chen¹, Chakrapand Paul Chiarawongse¹, Gregory Chin¹, Kathy Chin¹, Ben G. Cooper¹, Katherine Adele Corneilson¹, Amanda M. Danielson¹, Elizabeth S. Davis¹, Ycar Devis¹, Melissa Dong¹, Elizabeth K. Dossett¹, Nick Dulchin¹, Vincent N. Duong¹, Ben Ewing¹, Julia Mansfield Fuller¹, Thomas E. Gartman¹, Chad R. Goldberg¹, Jesse Greenfield¹, Selena Groh¹, Ross A. Hamilton¹, Will Hodge¹, Dylan Van Hong¹, Joshua E. Insler^{1,2}, Aava B. Jahan¹, Jessica Paola Jimbo¹, Emma M. Kahn¹, Daniel Knight¹, Grace E. Konstantin¹, Caitlin Kornick¹, Zachary J. Kramer¹, Meghan S. Lauzé¹, Misha S. Linnehan¹, Tommaso Lombardi¹, Hayley Long¹, Alec J. Lotstein¹, Myrna-Nahisha A. Lyncee¹, Monica Gabriella Lyons¹, Eli Maayan¹, Nicole Marie May¹, Elizabeth C. McCall¹, Rhea Ann Charlotte Montgomery-Walsh¹, Michael C. Morscher¹, Amelia D. Moser^{1,3}, Alexandra S. Mueller¹, Christin A. Mujica¹, Elim Na^{1,4}, Isabelle R. Newman¹, Meghan K. O'Brien¹, Katherine Alexandra Ochoa Castillo¹, Zaenab Ayotola Onipede¹, Danielle A. Pace¹, Jasper H. Park¹, Angeliki Perdikari¹, Catherine E. Perloff¹, Rachel C. Perry¹, Akash A. Pillai¹, Avni Rajpal¹, Emma Ranalli¹, Jillian E. Schreier¹, Justin R. Shangguan¹, Micaela Jen Silver¹, Avery Glennon Spratt¹, Rachel E. Stein¹, Grant J. Steinhauer¹, Devon K. Valera¹, Samantha M. Vervoordt¹, Lena Walton¹, Noah W. Weinflash¹, Karen Weinstock¹, Jiaqi Yuan¹, Dominique T. Zarrella¹, and Jonah E. Zarrow¹

¹Department of Psychology, Tufts University; ²Rush Medical College, Rush University; ³Department of Psychology and Neuroscience, University of Colorado Boulder; and ⁴School of Medicine, Boston University

Corresponding Author:

Heather L. Urry, Tufts University, Department of Psychology
E-mail: heather.urry@tufts.edu

Abstract

In this direct replication of Mueller and Oppenheimer's (2014) Study 1, participants watched a lecture while taking notes with a laptop ($n = 74$) or longhand ($n = 68$). After a brief distraction and without the opportunity to study, they took a quiz. As in the original study, laptop participants took notes containing more words spoken verbatim by the lecturer and more words overall than did longhand participants. However, laptop participants did not perform better than longhand participants on the quiz. Exploratory meta-analyses of eight similar studies echoed this pattern. In addition, in both the original study and our replication, higher word count was associated with better quiz performance, and higher verbatim overlap was associated with worse quiz performance, but the latter finding was not robust in our replication. Overall, results do not support the idea that longhand note taking improves immediate learning via better encoding of information.

Keywords

note taking, laptop, longhand, open data, open materials, preregistered

Received 3/29/19; Revision accepted 8/4/20

Ditch the laptop and pick up a pen, class. Researchers say it's better for note taking.

—Elahe Izadi (*The Washington Post*, August 26, 2014)

In educational settings, students and professors alike are keen to facilitate student learning. One common strategy that students adopt is to take notes during class using pen and paper or a laptop. Which note-taking medium promotes better learning? Mueller and Oppenheimer (2014) conducted a set of three experiments to find out.

In each experiment, participants watched a prerecorded lecture. Prior to watching the lecture, participants received either a laptop or pen and paper so they could take notes. They subsequently took a quiz about the lecture material. In two of three experiments, in which participants had no opportunity to study their notes, longhand note taking resulted in better performance than laptop note taking on items putatively tapping conceptual understanding. In a third experiment, the difference was found only among participants who studied their notes prior to taking the quiz a week later. In all three studies, participants took more notes in the laptop condition than the longhand condition, and their notes included more of the words used by the lecturer in the laptop condition than the longhand condition.

This work has been influential. For one thing, it may be guiding teaching decisions; it is frequently featured as a point of discussion among educators about the decision to allow or ban laptops in the classroom (e.g., see Holstead, 2015). Moreover, the work captured public imagination, with pieces published by *The Washington Post*, *NPR*, *Scientific American*, and other outlets. It has inspired headlines suggesting that students should ditch their laptops and take notes by hand, as highlighted in the epigraph; otherwise, they perform worse.

It has also captured the academic imagination; as of January 15, 2021, Mueller and Oppenheimer (2014) have been cited more than 1,100 times (Google Scholar), and the article's Attention score places it in "the top 5% of all research outputs ever tracked by Altmetric" (more than 16 million; <https://sage.altmetric.com/details/2300218#score>). It has also inspired several close replications (Kirkland, 2016; Luo et al., 2018; Mitchell & Zheng, 2017; Morehead et al., 2019).

In the current study ($N = 142$), we conducted a preregistered direct replication of Study 1 by Mueller and Oppenheimer (2014; $N = 65$). Participants took notes with a laptop or a pen while they watched one of five TED Talks. After a distractor-filled delay, they took a quiz that assessed their grasp of the material. We measured quiz performance, the number of words in their notes, and verbatim overlap between their notes and words used by the lecturer.

In confirmatory analyses, we tested the hypothesis that longhand note taking would lead to better performance on conceptual quiz items than laptop note taking. Such a result would indicate that the note-taking medium impacts transfer of new information to long-term memory, an extension of Di Vesta and Gray's (1972) encoding hypothesis. We also tested the hypothesis that laptop note taking would lead to more words in the notes (and, specifically, more words spoken verbatim by the lecturer) than longhand note taking. Finally, we conducted exploratory mini meta-analyses across similar studies to generate cumulative estimates of the size of note-taking effects on immediate quiz performance and notes contents to date.

Method

We report all data exclusions, manipulations, and measures in the study as well as how we determined our

sample size. This study was approved by the Social, Behavioral, and Educational Research Institutional Review Board at Tufts University. All participants provided written informed consent prior to participating.

We preregistered this study on March 7, 2017 (see <https://osf.io/qe3wb/wiki/home/>). Our materials, data, and analysis scripts are available on OSF at <https://osf.io/tr868/>. When reporting results for the studies published by Mueller and Oppenheimer (2014), we relied on updated data files posted at <https://osf.io/t43ua/> as part of their 2018 Corrigendum for the original article.

Participants

We recruited participants through posts on social media, e-mails to acquaintances and outreach lists, and flyers in heavily trafficked locations on campus. Undergraduates interested in participating were directed to complete an online screening survey that confirmed that they were college students and at least 18 years old; eligible individuals were then redirected to a scheduling website. Our recruitment materials are available on pages 58 and 59 of the pdf at <https://osf.io/y3ty8>.

A total of 145 undergraduate students from Tufts University participated in the experiment individually, typically with two experimenters. Two participants provided no responses to quiz items, and condition assignment was unclear for a third; thus, we excluded these three observations, leaving us with 142 participants for analysis. Notes were unavailable for two participants; thus, analyses involving variables derived from notes (word count, verbatim overlap) have two fewer observations. We present our sample-size rationale in the Supplemental Material available online.

Each participant was randomly assigned to view one of five lectures in either the laptop or longhand note-taking condition. Overall, there were 68 participants in the laptop condition (12–14 per lecture) and 74 in the longhand condition (13–18 per lecture). We thus had 80% power to detect a standardized effect (Cohen's d) of note-taking condition of ± 0.47 or larger for quiz-performance variables and ± 0.48 or larger for notes variables (two-tailed $\alpha = .05$). We also had 80% power to detect equivalence of the note-taking effect to zero within equivalence bounds (Cohen's d) of -0.49 to $+0.49$ for quiz-performance and notes variables (two-tailed $\alpha = .05$). Equivalence tests examined whether one can reject the presence of an effect as extreme or more so than one's equivalence bounds, ideally the smallest effect size of interest (Lakens et al., 2018).

Participants from all four graduation years were represented; the majority were sophomores (first year: 12%, second year: 49%, third year: 23%, and fourth year: 17%). With regard to gender, 62% identified as female

Statement of Relevance

In educational settings, students and professors alike are keen to facilitate learning. One common strategy that students adopt to help them learn is to take notes during class either longhand or with a laptop. Which note-taking medium promotes better learning? A 2014 article by Mueller and Oppenheimer provided evidence that longhand note taking was better for learning. Their work has been highly influential in and outside of academic circles and has arguably impacted teaching decisions. Given its impact, we tested whether the effect could be replicated. As a collaborative group of 88 researchers, we conducted our own original research and also considered eight independent studies on the same topic. Overall, our results did not support the idea that longhand note taking improves learning at least over short delays between learning and testing. This research sets the stage for potentially fruitful future research on this important topic.

and 37% as male; one person declined to report gender. With regard to race/ethnicity, 5% were African American or Black, 24% were Asian, 58% were White, 5% were Hispanic or Latinx, and 7% were multiracial; two people declined to report their race/ethnicity. Participants received \$15 in compensation.

Materials

Lectures. The lectures for this study were the same TED Talks used in the original study. They lasted approximately 15 min each. Links to their location at www.ted.com, from which the videos were streamed and transcripts obtained, are available in the Supplemental Material.

Quiz performance. Participants responded to open-ended quiz items from the original study for each of the five lectures. Per Mueller and Oppenheimer, we divided items into two types, factual recall and conceptual application. The extent to which the quiz items reflect a valid distinction between factual and conceptual understanding is unclear. However, we used the factual versus conceptual labels from the original study to facilitate comparison.

There were five to seven items for factual-recall performance, for example, "According to the speaker, what kinds of stressful tasks most reliably raise the level of cortisol (a stress-related hormone)?" and three to five items for conceptual-application performance, for example, "Why are the negative outcomes the speaker discusses (social problems, life expectancy, etc.)

correlated with economic status within countries, but not across countries?” depending on which lecture the participant viewed.

A total of 12 to 15 raters scored participants’ open-ended responses on the basis of a standard scoring key. For details regarding scoring and interrater reliability, see the Supplemental Material.

For both factual-recall and conceptual-application item types, interrater reliability was excellent for all lectures. The minimum and maximum intraclass correlations (ICCs) across lectures, respectively, were .98 and .99 for factual recall and .90 and .98 for conceptual application. We calculated a total index score for each participant as the mean across raters separately for factual-recall and conceptual-application scales (maximum = 10). We then standardized these scores across lectures as Mueller and Oppenheimer did; we also computed the proportion of correct responses.

Content of notes: word count and verbatim overlap. For each participant, we determined the number of words in their lecture notes and the degree to which three-word chunks of text (trigrams) from a transcript of the lecture were present in those notes using the *tidytext* package (Version 0.1.9; Silge & Robinson, 2016) in the R programming environment (R Core Team, 2020). We expressed verbatim overlap as a percentage: $100 \times L/T$, where L is the number of lecture trigrams in participant notes, and T is the total number of trigrams in participant notes.

Distractor tasks. As a distraction after the video lectures, participants completed, in order, a typing test, a questionnaire, and a reading span task. On the basis of experimenter reports of typing-test start times and reading-span-task end times, we found that participants were distracted for 24.02 min on average across distractor tasks (95% confidence interval [CI] = [23.36, 24.69]). Distraction duration was similar for participants in the laptop condition ($M = 23.73$, 95% CI = [22.8, 24.66]) and longhand condition ($M = 24.29$, 95% CI = [23.33, 25.26]), $\Delta M = 0.56$, 95% CI = [-0.77, 1.89], $t(136.00) = 0.84$, $p = .403$. Thus, the distraction period was sufficient to build in an approximately 30-min delay between lecture and quiz as in the original study, and it did not vary by note-taking condition. For detailed information about the distractor tasks, see the Supplemental Material.

Design and procedure

We collected data in person in various locations on Tufts University’s Medford campus. Before participants arrived, experimenters assembled the relevant forms and opened the Qualtrics survey that would be used to administer all study procedures, including random

assignment of participants to conditions. Our Qualtrics survey template is available at <https://osf.io/s5gfd>.

We used a 2 (condition: laptop, longhand) \times 5 (lecture) between-subjects factorial design for this experiment. After obtaining written informed consent, experimenters provided each participant with either a pen and paper or an experimenter-owned laptop on which to take notes. If the participant brought headphones to the session, the experimenter ensured that participants were wearing them and that they were plugged into the jack on a second computer, typically another experimenter-owned laptop, that would display the lecture. They then said, “You will now watch a lecture on this monitor. Please use your normal classroom note-taking strategy. We’re interested in how information is actually recorded during class lectures.” The experimenter ensured that the display screen was visible and then moved to an area of the room outside of the participant’s line of sight.

When the video ended, the experimenter retrieved the note-taking laptop or pen and paper and said, “Now, we’d like for you to complete several tasks here on this computer. This part of the study will take about 30 minutes in total. Please let me know after you’ve finished each task.” Participants then completed the distractor tasks, with the experimenter moving out of the participant’s line of sight for each task. After completing the distractor tasks, participants completed the quiz for the lecture they had viewed earlier.

Finally, participants responded to a number of self-report questions about their note-taking-medium preferences and beliefs, described in the Supplemental Material. They also indicated with which gender and racial or ethnic group (or groups) they most identify. At this point, experimenters debriefed participants, collected information required to compensate them via PayPal, thanked them, and excused them.

Deviations from the original method

Our replication of Study 1 differed from Mueller and Oppenheimer’s in the following ways. First, we did not collect grade point averages or SAT scores from participants because this sensitive information was not critical to replicating the key findings of the original study. Because these variables were collected after the manipulations and measures of interest in the original study, their omission could not have affected our replication results.

Second, we administered all manipulations and measures via a Qualtrics survey. The survey linked to other websites for the reading-span-task and typing-test distractors. Doing so facilitated our ability to collect the data in a standardized way for every participant and minimized the risk of data loss given the number of

experimenters who collected the data. It is possible that this change could have affected the key results.

Third, we added an open-ended question asking participants to tell us what they thought the study was about. We administered this item after the manipulations and measures of interest; its inclusion could not have affected the key results.

Fourth, we recruited college undergraduates at Tufts University in 2017 rather than Princeton University circa 2013. These are both selective private institutions, and thus, the populations of interest are similar; nevertheless, it is possible that drawing from a different population at a different time could have affected the key results. For example, there could have been a difference in the frequency with which students typically took notes with a laptop versus longhand in the two studies.

Fifth, in Mueller and Oppenheimer's original Study 1, participants completed the study in a classroom, generally in groups of two, and the video lecture was presented via a projector on a screen at the front of the room. We could not ensure that a classroom setting would always be accessible to our experimenters and could not provide a standardized set of laptops to experimenters. Thus, participants viewed the lecture on a monitor, typically a laptop owned by an experimenter. When available, participants wore headphones or earbuds to minimize distraction. In addition, participants took notes on a laptop that was owned by an experimenter, when applicable. We do not think that using a laptop-headphone setup is likely to have affected the key results of interest for our replication of Study 1 because Mueller and Oppenheimer adopted the same procedure in their Study 2. Variation in settings and types of laptops used for note taking and lecture watching could, however, have introduced variability that affected key results.

Sixth, Mueller kindly provided the two 5-min distractor tasks used in the original Study 1, but these materials were not amenable to administration via Qualtrics. Thus, alongside the reading span task—one of the three original distractors in Study 1—we administered a 5-min typing test and the Need for Cognition Scale (Cacioppo et al., 1984) instead. These are the same distractors used by Mueller and Oppenheimer in their Study 2. Thus, we do not think this change is likely to have affected the key results of interest for our replication of Study 1.

Confirmatory data analysis and inference criteria

We analyzed our data using R (Version 4.0.2; R Core Team, 2020) and wrote the manuscript for this article in R Markdown via RStudio 1.3.1056 (RStudio Team, 2020). The *papaja* (Version 0.1.0.9997; Aust & Barth, 2018) and

knitr (Version 1.29; Xie, 2015) packages were instrumental to producing the formatted document. We used an alpha of .05 for null-hypothesis significance testing.

In accordance with Mueller and Oppenheimer's original study and our preregistered analysis plan, we conducted independent-samples *t* tests to determine whether assignment to the laptop condition, compared with the longhand condition, influenced word count and degree of verbatim overlap in the notes that participants took. We also conducted mixed-effects analyses of variance (ANOVAs) in quiz performance with note-taking condition as a fixed effect, lecture as a random effect, and a random slope for note-taking condition. The original authors used the UNIANOVA command in *SPSS* for these analyses; we conducted them using the *afex* package in R (Version 0.27-2; Singmann et al., 2019).

In addition, we examined whether the effect sizes in the present study were significantly different from those reported in the original study by conducting several pairs of one-sided tests using the *TOSTER* package (Version 0.3.4; Lakens, 2017). For the factual- and conceptual-performance variables, we set the upper bound to the size of the original effect ($d = 0.01$ and 0.34 , respectively) and the lower bound to -999 . For the word-count and verbatim-overlap variables, we set the lower equivalence bound to the size of the original effect ($d = -1.43$ and -0.94 , respectively) and the upper bound to $+999$. These analyses amount to inferiority tests (Lakens et al., 2018).

We also examined whether the effect sizes (d s) in the present study were equivalent to 0 using *TOSTER*. In this case, we set equivalence bounds from -0.49 to 0.49 for the two quiz-performance variables and the two notes variables. We selected these equivalence bounds because they yield 80% power to detect equivalence given the final sample size, an approach recommended by Lakens (2017) as one way of defining the smallest effect size of interest.

A successful replication of results should yield a statistically significant effect of note-taking condition on number of words (laptop > longhand), verbatim overlap (laptop > longhand), and conceptual-application performance (longhand > laptop). In addition, replication effect sizes should be neither significantly different from the original effect sizes nor equivalent to 0. For a list of deviations from our preregistered analysis plan, see the Supplemental Material.

Results

Preliminary analyses

Summary statistics. For descriptive statistics for and correlations between measured variables in our replication, see Table A1 in the Supplemental Material.

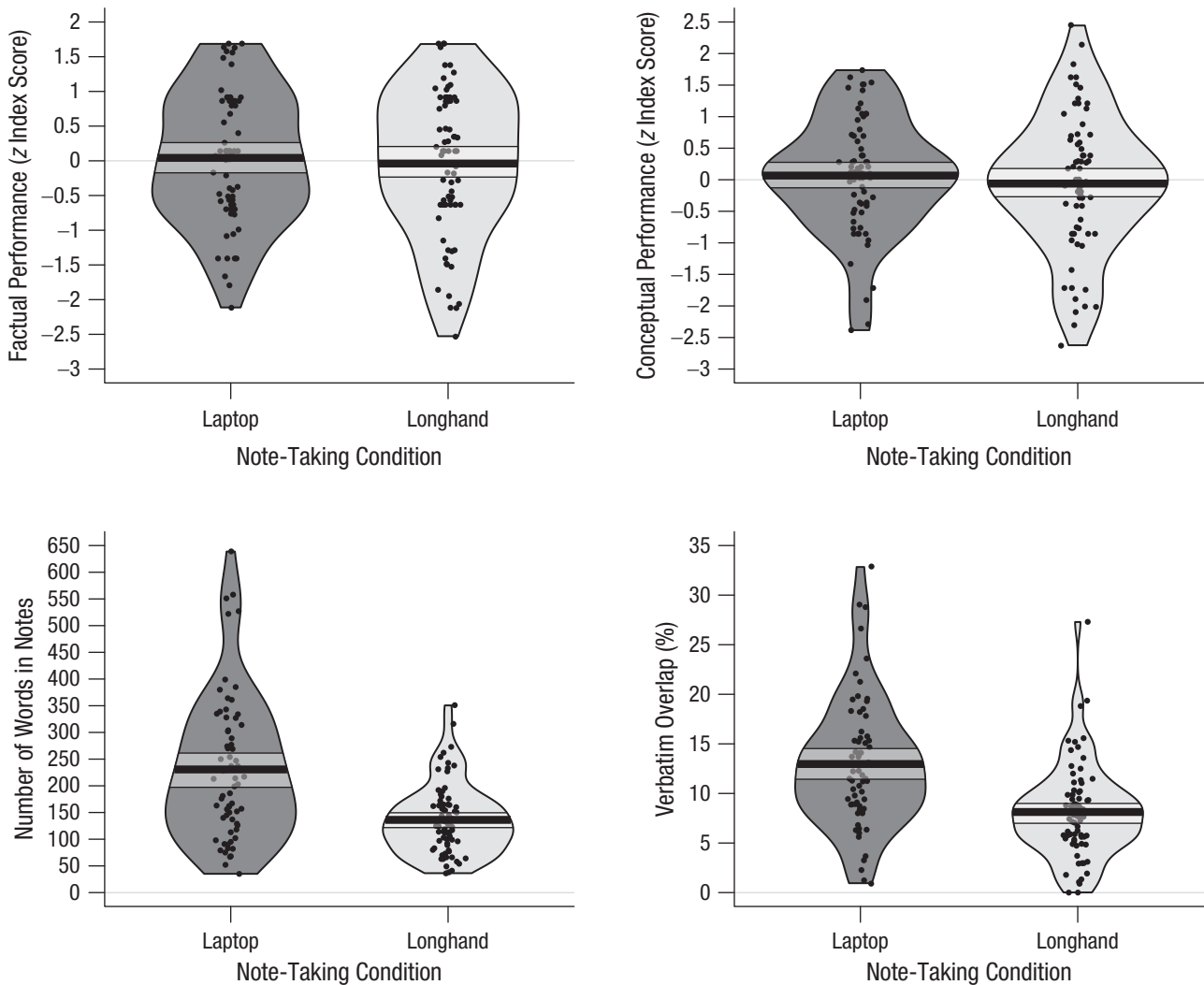


Fig. 1. Violin plots depicting results for the four primary dependent variables as a function of note-taking condition in the present replication study. For quiz performance (top row), we depict standardized scores. In all plots, the mean is represented by the thick black line in each condition. Error bars, shown in lighter shading around the mean, represent 95% confidence intervals. The width of each shaded area indicates the density of the data, and dots are individual data points. Plots were generated using the *yarr* package (Version 0.1.5; Phillips, 2017).

Influential observations. We identified potentially influential observations, that is, observations that may have biased the effect of note-taking condition on our four criterion variables, as described in the Supplemental Material. We repeated our confirmatory analyses after excluding these influential observations, as specified in our preregistration (see <https://osf.io/qe3wb/wiki/home/>); their exclusion did not alter conclusions about experimental effects.

Confirmatory analyses

Figure 1 shows results for all four primary dependent variables as a function of note-taking condition. The

top row plots standardized quiz performance; factual-recall items are on the left, and conceptual-application items are on the right. The bottom row plots notes content; word count is on the left, and verbatim overlap is on the right.

Effect of note-taking condition on quiz performance.

Table 1 shows the fixed effect of note-taking condition on quiz performance in the original study by Mueller and Oppenheimer (2014; Study 1) and the present replication study. We show results for standardized performance scores, as presented in the original study, as well as for the proportion of correct responses, a more intuitive measure

Table 1. Analysis-of-Variance Results (Type III Sums of Squares): Effect of Note-Taking Condition on Quiz Performance (Standardized Scores and Proportion of Correct Responses) in Mueller and Oppenheimer's (2014) Study 1 and in the Present Replication Study

Measure and study	<i>MSE</i>	<i>F</i> (1, 4)	η_G^2	<i>p</i>
Factual recall (<i>z</i> index score)				
Mueller and Oppenheimer Study 1	0.22	0.05	.00	.84
Replication study—all observations	0.09	0.13	.00	.74
Replication study—excluding influential observations	0.07	0.00	.00	.98
Conceptual application (<i>z</i> index score)				
Mueller and Oppenheimer Study 1	0.04	8.08	.20	.05
Replication study—all observations	0.08	0.59	.02	.48
Replication study—excluding influential observations	0.03	2.68	.05	.18
Factual recall (proportion correct)				
Mueller and Oppenheimer Study 1	0.01	0.04	.00	.86
Replication study—all observations	0.00	0.07	.00	.81
Replication study—excluding influential observations	0.00	0.02	.00	.91
Conceptual application (proportion correct)				
Mueller and Oppenheimer Study 1	0.00	9.40	.34	.04
Replication study—all observations	0.01	0.35	.02	.59
Replication study—excluding influential observations	0.00	2.34	.04	.20

Note: The fixed effects of condition on factual-recall and conceptual-application performance in Mueller and Oppenheimer's Study 1 were $F(1, 4.01) = 0.046, p = .841$, and $F(1, 4.09) = 8.05, p = .046$, respectively, based on the UNIANOVA command in SPSS (see the files for Mueller and Oppenheimer's 2018 Corrigendum at <https://osf.io/t43ua/>). We report results from the corresponding analysis in R using the `afex::aov_4` command. The values differ slightly because of differences in how SPSS and *afex* handle random effects, but substantive conclusions remain the same.

of performance. As shown, removing influential observations from the replication data had little effect on conclusions; thus, results below include all observations.

Consistent with the findings of Mueller and Oppenheimer, the difference in factual-recall performance between the laptop and longhand conditions was not significant (see Fig. 1, top left). Their study yielded a near-zero effect slightly favoring better factual-recall performance in the longhand than the laptop condition (Hedges's $g = 0.01$, 95% CI = $[-0.48, 0.50]$) on the basis of standardized scores. The effect in the present replication study was negligible in the opposite direction (Hedges's $g = -0.08$, 95% CI = $[-0.41, 0.25]$); not significantly different from the original effect, $t(139.96) = -0.55, p = .291$; and equivalent to -0.49 to 0.49 , $t(139.96) = 2.45, p = .008$.

Contrary to the findings of the Mueller and Oppenheimer, the difference in conceptual-application performance was not significant (see Fig. 1, top right). Their study yielded a small to medium-sized effect, suggesting better performance in the longhand than the laptop condition (Hedges's $g = 0.34$, 95% CI = $[-0.16, 0.83]$), on the basis of standardized units. The effect in the replication study was negligible in the opposite direction (Hedges's $g = -0.13$, 95% CI = $[-0.45, 0.20]$); significantly different from the original effect, $t(139.03) = -2.78, p = .003$; and equivalent to -0.49 to 0.49 , $t(139.03) = 2.17, p = .016$.

In units of proportion correct, mean factual-recall performance in the laptop condition was .63 ($SD = .20$, 95% CI = $[.58, .68]$); mean factual-recall performance in the longhand condition was .62 ($SD = .23$, 95% CI = $[.57, .68]$). Mean conceptual-application performance in the laptop condition was .74 ($SD = .19$, 95% CI = $[.69, .78]$); mean conceptual-application performance in the longhand condition was .70 ($SD = .23$, 95% CI = $[.65, .75]$).

Averaged across note-taking conditions, results showed that participants in Mueller and Oppenheimer's study scored lower by a proportion of .05 on factual-recall items ($M = .58, SD = .21$, 95% CI = $[.53, .63]$) than did participants in this replication study ($M = .63, SD = .21$, 95% CI = $[.59, .66]$). Similarly, participants in the original study scored lower by a proportion of .08 on conceptual-application items ($M = .63, SD = .23$, 95% CI = $[.58, .69]$) than participants in this replication study ($M = .72, SD = .21$, 95% CI = $[.68, .75]$).

Effect of note-taking condition on content of notes.

Consistent with the original study, results showed that taking notes using a laptop led to a higher word count ($M = 230.69, SD = 133.87$, 95% CI = $[198.03, 263.34]$) than taking notes longhand ($M = 136.16, SD = 66.26$, 95% CI = $[120.71, 151.62]$), $t(94.63) = -5.22, p < .001$ (see Fig. 1, bottom left). Removing influential observations had little effect on the statistical results, $t(99.41) = -4.91, p < .001$.

Both the original and replication studies yielded large effects, suggesting a higher word count in the laptop than the longhand condition (Mueller and Oppenheimer: Hedges's $g = -1.41$, 95% CI = $[-1.96, -0.86]$; present replication: Hedges's $g = -0.90$, 95% CI = $[-1.25, -0.56]$). The effect in the replication study was significantly different from the original effect, $t(94.63) = 3.02$, $p = .002$, but it was not equivalent to -0.49 to 0.49 , $t(94.63) = -2.34$, $p = .989$.

Again consistent with Mueller and Oppenheimer's study, results showed that taking notes using a laptop ($M = 12.97\%$, $SD = 6.53$, 95% CI = $[11.37, 14.56]$) led to more verbatim overlap with the lecture than writing notes longhand ($M = 8.13\%$, $SD = 4.73$, 95% CI = $[7.03, 9.23]$), $t(119.46) = -4.98$, $p < .001$ (see Fig. 1, bottom right). Removing influential observations had little effect on the statistical results, $t(119.30) = -5.58$, $p < .001$. Both studies yielded large effects, suggesting a higher verbatim overlap in the laptop than the longhand condition (Mueller and Oppenheimer: Hedges's $g = -0.93$, 95% CI = $[-1.44, -0.41]$; present replication: Hedges's $g = -0.85$, 95% CI = $[-1.19, -0.51]$). The effect in the replication study was not significantly different from the original effect, $t(119.46) = 0.45$, $p = .326$, nor was it equivalent to -0.49 to 0.49 , $t(119.46) = -2.08$, $p = .980$.

Averaged across note-taking conditions, results showed that participants in the original study typed 56.95 more words ($M = 238.35$, $SD = 116.79$, 95% CI = $[209.41, 267.29]$) than participants in this replication study ($M = 181.40$, $SD = 114.14$, 95% CI = $[162.33, 200.47]$). Levels of verbatim overlap were similar; participants in the original study exhibited just 1.09% greater verbatim overlap ($M = 11.53\%$, $SD = 6.69$, 95% CI = $[9.87, 13.19]$) than participants in this replication ($M = 10.44\%$, $SD = 6.14$, 95% CI = $[9.42, 11.47]$).

Exploratory mini meta-analyses

Our replication's experimental results are consistent with those of Mueller and Oppenheimer in that both demonstrate a laptop-superiority effect when it comes to the number of words in notes and the extent of verbatim overlap with the lecture, and they demonstrate no effect of note-taking condition on factual-recall performance. However, our replication results are inconsistent with those of the original study in that they do not demonstrate a longhand-superiority effect when it comes to conceptual-recall performance. It may be that our particular instantiation resulted in false-negative effects for conceptual items because of methodological differences (e.g., use of Qualtrics to collect the data; different population of undergraduates; variation in data collection settings, experimenters, and computer equipment). Thus, next we conducted exploratory mini

meta-analyses to integrate evidence across multiple similar studies as a more robust test of the hypothesis.

To estimate the effect of note-taking condition on quiz performance, word count, and verbatim overlap, we located a total of eight very similar studies that met the following criteria: (a) experimentally manipulated laptop versus longhand note taking; (b) assessed immediate quiz performance on the same day as exposure to the lecture; (c) used video lecture material; (d) measured and reported results for quiz performance, word count, and verbatim overlap; and (e) studied undergraduates. See the Supplemental Material for information about our search strategy. Although eight studies is insufficient to make definitive conclusions, it does afford an interim aggregation of cumulative knowledge that can yield testable predictions for future work.

The set of eight studies comprised Studies 1 and 2 by Mueller and Oppenheimer (not Study 3, which assessed quiz performance 1 week later), two studies reported by Morehead et al. (2019; immediate condition only), the current study, and three more single-study replications (Kirkland, 2016; Luo et al., 2018; Mitchell & Zheng, 2017). We excluded participants in the laptop-intervention condition in Mueller and Oppenheimer's Study 2 given that the goal of the intervention was to eliminate or reduce the difference between the laptop and longhand conditions. Also, we could meta-analyze only seven studies for factual-recall and conceptual-application performance because the authors reported performance across item types in one study (Luo et al., 2018).

One source of variation across studies, despite otherwise similar methods, is the lecture video material. Morehead et al. (2019), Mitchell and Zheng (2017), and the current replication used at least one of the original TED Talk lectures; Kirkland (2016) and Luo et al. (2018) used other video material that lasted a bit longer than the original 15-min videos (28 min and 23 min, respectively). We excluded the recent study of 7th- to 9th-grade students (Frantz et al., 2018) in part because participants were not university undergraduates, which could introduce age-related heterogeneity, and in part because they did not report results for word count and verbatim overlap.

We ran five random-effects meta-analyses to estimate the effect of note-taking condition on quiz performance (total, factual, conceptual) and the content of notes (word count, verbatim overlap) using the *metafor* package (Version 2.4-0; Viechtbauer, 2010). We computed effect sizes as longhand minus laptop using the *effsize* (Version 0.8.0; Torchiano, 2019) and *compute.es* (Version 0.2-5; Re, 2013) packages.

The forest plot in Figure 2 summarizes quiz-performance findings. Across studies, taking notes longhand as opposed to with a laptop boosted total quiz performance across factual and conceptual item types to a negligible

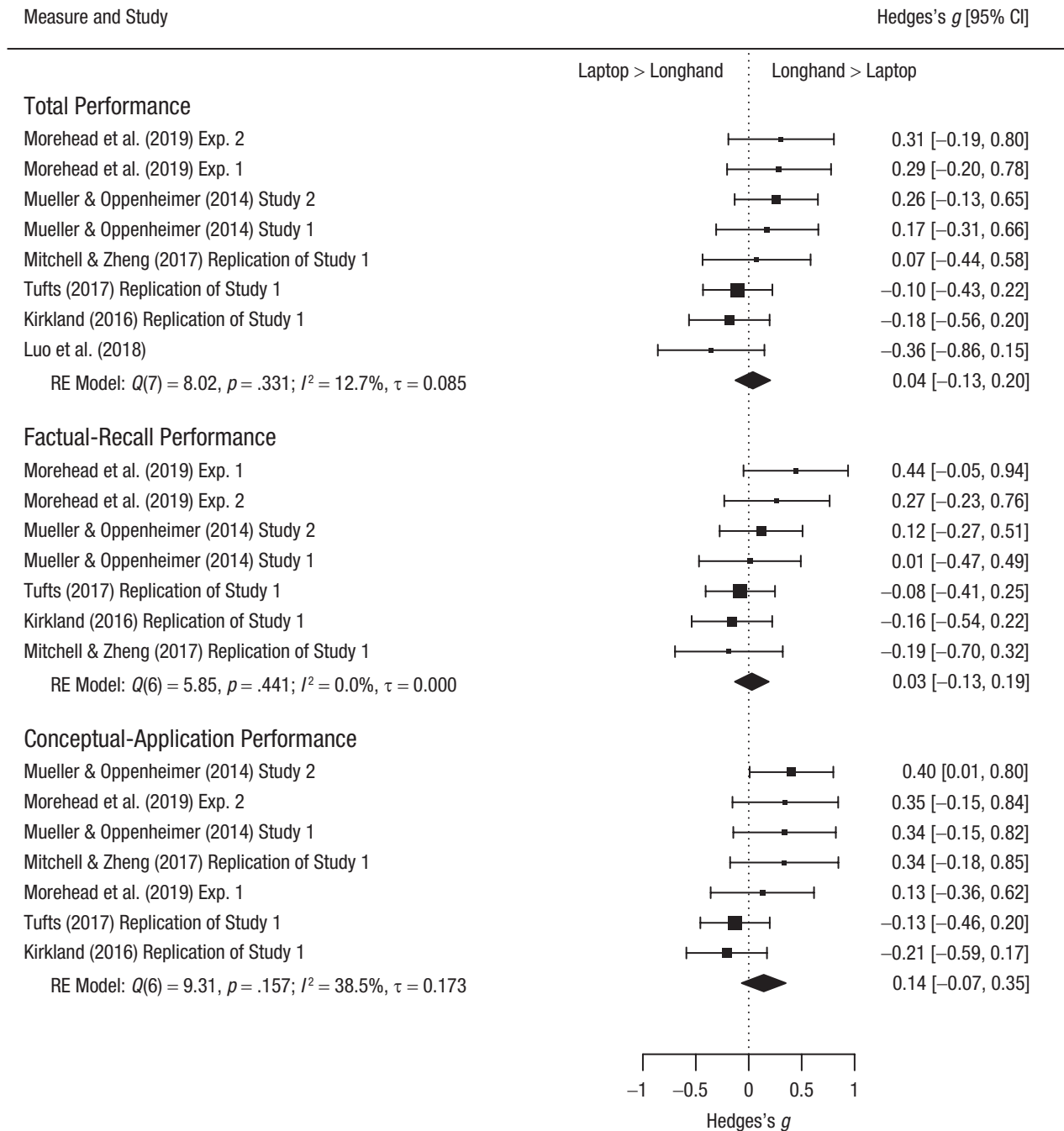


Fig. 2. Standardized effect sizes for the quiz-performance measures in Mueller and Oppenheimer’s (2014) study and all replications. For each measure, we present Hedges’s *g* point estimates in descending order. Error bars represent 95% confidence intervals (CIs). The size of the symbols is inversely proportional to the variance of the estimate; larger symbols indicate more precise estimation. We generated overall estimates using a random-effects (RE) model. Overall estimates are depicted with black diamonds.

degree (Hedges’s $g = 0.04$, 95% CI = [-0.13, 0.20]). This effect was not statistically significant ($z = 0.46, p = .645$), and it also was equivalent to -0.38 to 0.38 ($z = -4.10, p \leq .001$). The equivalence bound (d) of 0.38 reflects the effect size that Mueller and Oppenheimer’s study

could detect with 33% power. Equivalence, thus, suggests that the meta-analytic effect size was too small to have been detected in the original study.

The effects for the two item types separately, factual-recall performance (Hedges’s $g = 0.03$, 95% CI = [-0.13,

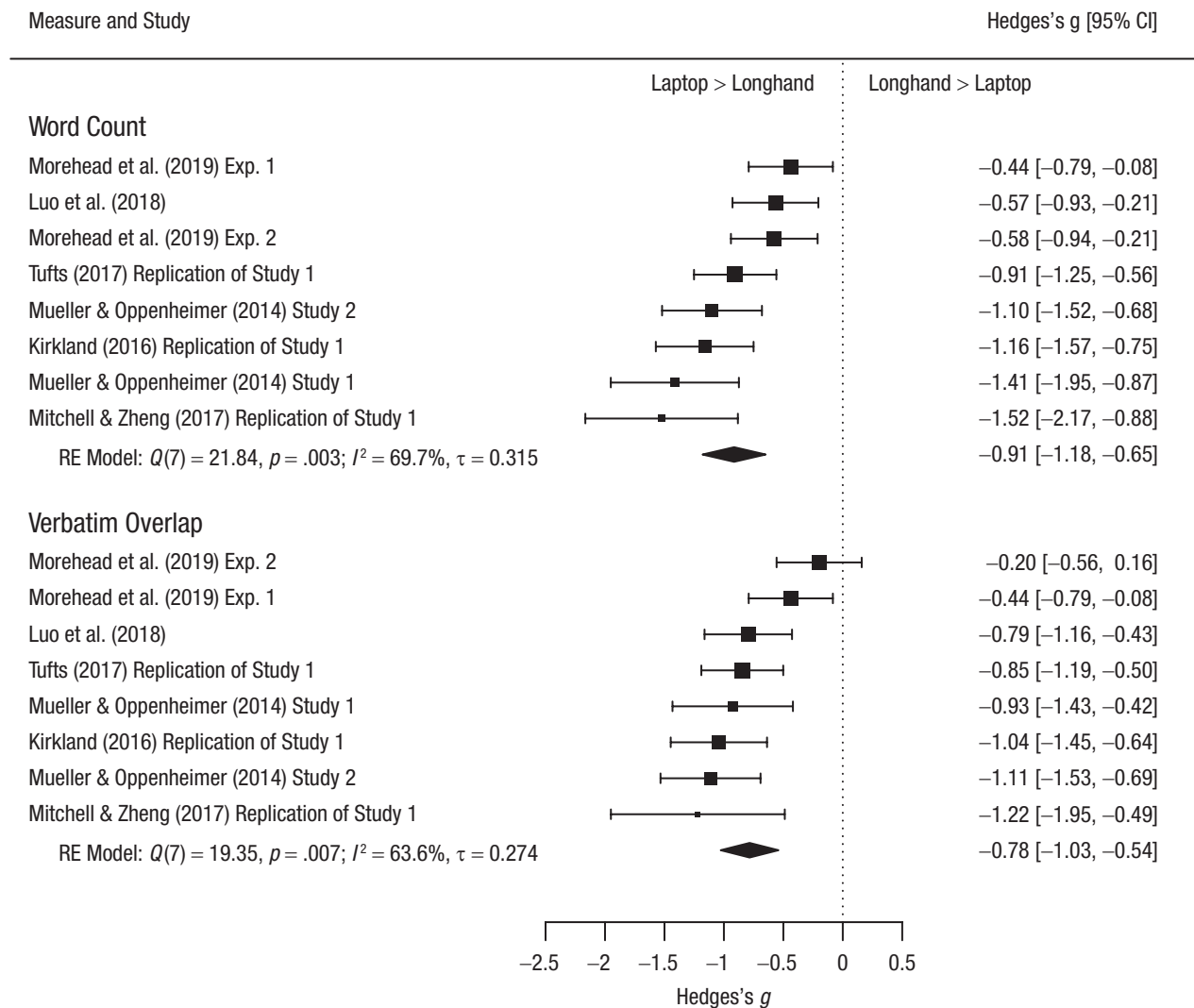


Fig. 3. Standardized effect sizes for the notes-content measures in Mueller and Oppenheimer's (2014) study and all replications. For each measure, we present Hedges's *g* point estimates in descending order. Error bars represent 95% confidence intervals (CIs). The size of the symbols is inversely proportional to the variance of the estimate; larger symbols indicate more precise estimation. We generated overall estimates using a random-effects (RE) model. Overall estimates are depicted with black diamonds.

0.19]) and conceptual-application performance (Hedges's $g = 0.14$, 95% CI = [-0.07, 0.35]), were negligible to very small. These effects were not statistically significant ($z = 0.36$, $p = .719$, and $z = 1.34$, $p = .182$, respectively), and both were equivalent to -0.38 to 0.38 ($z = -4.33$, $p \leq .001$, and $z = -2.27$, $p = .011$, respectively).

The forest plot in Figure 3 summarizes notes-content findings. Consistent with the original study, results showed that taking notes with a laptop boosted both word count (Hedges's $g = -0.91$, 95% CI = [-1.18, -0.65]) and degree of verbatim overlap (Hedges's $g = -0.78$, 95% CI = [-1.03, -0.54]) to a large degree. These effects were statistically significant ($z = -6.75$, $p \leq .001$, and

$z = -6.34$, $p \leq .001$, respectively), and neither of them was equivalent to -0.38 to 0.38 ($z = -3.92$, $p > .999$, and $z = -3.24$, $p = .999$, respectively).

A modest percentage of the total variability across studies was due to heterogeneity of true effects for total quiz performance ($I^2 = 12.73\%$). This appeared to be driven more so by conceptual-application performance ($I^2 = 38.54\%$) than factual-recall performance ($I^2 = .001\%$). In terms of notes content, a large percentage of the total variability across studies was due to heterogeneity of true effects for word count ($I^2 = 69.69\%$) and verbatim overlap ($I^2 = 63.58\%$). In the Supplemental Material, we address whether effects of note-taking

condition on notes-content variables are correlated with effects of note-taking condition on quiz-performance variables at the study level.

Additional exploratory analyses

We present a number of additional exploratory analyses in the Supplemental Material, which we summarize briefly here for the sake of completeness. In one set of exploratory analyses, we took a Bayesian approach to examine relative evidence for the replication and null hypotheses. Consistent with results presented above, results generally favored the replication hypothesis for notes variables and the null hypothesis for quiz-performance variables.

In a second set of exploratory analyses, we conducted mixed-effects ANOVAs on quiz performance with item type treated as a factor (instead of examining factual and conceptual performance in separate analyses). There was no significant effect of note-taking condition either on its own or in interaction with item type; this was true in our replication and in Mueller and Oppenheimer's original study.

In a third set of analyses, we examined continuous predictors of quiz performance in linear mixed-effects regressions; such analyses could reveal hypothesized effects of note-taking condition by accounting for variance in quiz performance otherwise attributed to error in confirmatory analyses. However, there were no significant effects of note-taking condition when analyses accounted for these extraneous variables. As in the original study, higher word count was associated with better quiz performance; higher verbatim overlap was associated with worse quiz performance, but inconsistently so, depending on analysis.

In a fourth set of analyses, we examined laptop versus longhand note-taking preferences. Our replication participants were more likely to say that they tended to take notes longhand; participants in Mueller and Oppenheimer's study were more likely to say that they tended to take notes using a laptop. Our replication participants also believed, on average, that taking notes longhand is better for learning; participants in Mueller and Oppenheimer's study believed, on average, that there was not much of a difference.

Discussion

Summary and evaluation of replication results

Mueller and Oppenheimer (2014) found in their first study that participants who took lecture notes on a laptop demonstrated poorer performance on putatively conceptual quiz items than their counterparts who took

lecture notes by hand. Laptop notes contained more words and greater verbatim overlap with lecture content than longhand notes. Moreover, people whose notes had more words but less verbatim overlap performed better. Laptop versus longhand note taking had no effect on factual-recall quiz performance.

In our replication study, laptop notes contained more words and greater verbatim overlap with lecture content than longhand notes. However, unlike Mueller and Oppenheimer, we found only small, statistically nonsignificant differences in quiz performance as a function of note-taking medium. This conclusion was borne out in mixed-effects ANOVAs, equivalence tests, Bayesian analyses, and linear mixed-effects regressions. Thus, we replicated the experimental effect of note-taking condition assignment on notes but not quiz performance.

We also replicated correlational results reported by Mueller and Oppenheimer. Consistent with their study, ours showed that higher word count was associated with better quiz performance. We also found that higher verbatim overlap was associated with worse quiz performance, albeit less robustly. It would be tempting to conclude that taking more notes causes better quiz performance or that taking verbatim notes causes worse performance. However, we did not manipulate word count or the extent to which the notes exhibited verbatim overlap with the lecture; thus, alternative explanations are plausible. Higher word count or lower verbatim overlap may be third-variable proxies for motivation, conscientiousness, or interest, any of which might prompt students to take more notes in their own words and do better on the test.

Mini meta-analyses of very similar studies

There have been several parallel efforts by other researchers to replicate the experimental effect of note-taking condition on both quiz performance and notes content in undergraduates watching lecture videos. This is not surprising in light of the theoretical and practical importance of the findings.

Our mini meta-analyses of studies that reported the same dependent measures in undergraduates—two in the original report plus six by other researchers—suggested that the experimental effect on quiz performance was near zero irrespective of item type. CIs around the point estimates indicated that negligible to small effects favoring laptop or longhand superiority were both compatible with the data. There was modest heterogeneity in the extent to which this was true across studies.

By contrast, across the board, these studies found that laptop note taking boosted both word count and verbatim overlap with the lecture relative to longhand note taking. CIs around the point estimates indicated

that medium to large effects favoring laptop superiority were compatible with the data. However, there was considerable heterogeneity in the extent to which this was true across studies.

Our mini meta-analyses, thus, replicated the experimental effect of note-taking condition on notes but not quiz performance. This reduces concern about the limitations of our single replication. However, because these meta-analyses included only eight studies, likely did not include all unpublished attempts to replicate the original study, and did not take into account publication bias, our meta-analytic estimates should be considered preliminary.

Limitations and future directions

Our direct replication of Mueller and Oppenheimer's Study 1 was limited by some deviations from the original study; we comment further on one deviation, namely, the nature of our data-collection sessions. Specifically, approximately 80 students partnered to run data-collection sessions on campus at various times of day outside of class. Many noted that sessions were subject to distractions and errors; sessions also varied in formality and equipment (i.e., laptops and headphones). Thus, situation noise was likely a considerable source of random error that could have reduced sensitivity to detect note-taking effects on quiz performance. In future studies examining effects of laptop versus longhand note-taking, the context should be controlled to minimize these sources of random error.

There are several important directions for future research. First, we considered the effect of laptop versus longhand note taking only on immediate testing with no opportunity to study. Some studies suggest that effects of note-taking condition occur only when participants have the opportunity to study their notes (Luo et al., 2018; Mueller & Oppenheimer, 2014, Study 3); in future studies, experimental efforts should be focused in this direction.

Second, the studies in our meta-analyses mostly used TED Talks as lectures. These are interesting and unfamiliar to students but also brief and unlike actual classroom lectures. Disallowing pauses to catch up on note taking or ask questions takes the experimental context further afield of reality. Future studies should use approaches that better represent real-world settings and new note-taking technologies (e.g., the eWriter examined by Morehead et al., 2019) and account for note-taking preferences. For example, our replication participants were more apt than Mueller and Oppenheimer's participants to say that they generally took class notes by hand. Maybe longhand note taking has bigger effects on performance in people who typically take laptop notes. Although one study failed to observe a moderating

effect of note-taking preference (Kirkland, 2016), higher powered research is needed.

Third, future studies should, ideally, include a no-notes control condition to see the effect of taking notes regardless of medium (Jansen et al., 2017). Focusing on the laptop–longhand comparison without a no-notes control encourages dichotomous thinking when the story likely is more complicated. Jansen and colleagues (2017) suggest, for example, that a note-taking benefit “depends on the way lectures are presented, how notes are taken, and individual differences in cognitive abilities” (p. 231).

Finally, the studies considered herein examined whether the note-taking medium influences information encoding; they did not address other important issues that bear on the utility of laptops in classrooms. For example, laptops (and other Web-enabled devices) can support active learning and are necessary for learning for some disabled students; they may also be a source of distraction. Future studies must address these other issues.

Psychological science in the classroom

Psychologists have spearheaded several large-scale replication efforts such as Reproducibility Project: Psychology (Open Science Collaboration, 2015) and Many Labs (e.g., Klein et al., 2014). Large-scale efforts alone, however, are insufficient to increase the frequency of replications; conducting “didactic replications” in our classes—as we did here—is another option (Frank & Saxe, 2012; Gernsbacher, 2018; Grahe et al., 2012; Hawkins et al., 2018). Frank and Saxe (2012) argue, for example, that students in research-methods courses often must conceive, design, and conduct studies in just a few weeks, often with little enthusiasm. Mentoring publication-worthy replication studies instead may simultaneously inspire curiosity and motivation in students and generate value outside the classroom. Mechanisms such as the Collaborative Replications and Education Project (Wagge et al., 2019) can support these efforts.

Conclusion

Our direct replication of Mueller and Oppenheimer's Study 1 showed that, relative to longhand note taking, laptop note taking boosted word count and verbatim overlap with lecture content, but it did not reduce knowledge of the lecture material after a brief delay with no opportunity to study. Results, thus, did not support the idea that longhand note taking improves performance via better encoding of information.

When original and replication studies find different results, there are three interpretations: (a) There was a problem with the replication, (b) there was a problem

with the original research, and (c) the phenomenon under study is not enduring or universal (i.e., there is a constraint on generality). These interpretations are not mutually exclusive. In fact, all three apply here. Situation noise was a problem with our replication. Weak evidence (large p value, Bayesian evidence favoring the null hypothesis) and a small sample size were problems with Mueller and Oppenheimer's original study. And a difference in preferences of note-taking medium between the two may represent a constraint on generality.

Meta-analytic work can help to distill the conclusions we should draw from a body of studies. Our exploratory mini meta-analyses of studies that used similar same-day laboratory experimental procedures failed to support longhand superiority for retention of lecture material. A recent meta-analysis across a larger, more heterogeneous set of classroom studies revealed a small effect that supported longhand superiority (Allen et al., 2020). Neither of these meta-analyses considered the effect of publication bias or the extent to which the opportunity to study the notes or preference of note-taking medium moderates findings.

Until future research determines whether and when note-taking media influence academic performance, we conclude that students and professors who are concerned about detrimental effects of computer note taking on encoding information to be learned in lectures may not need to ditch the laptop just yet. However, there is more work to be done using methods that more closely mimic actual educational contexts and that evaluate the impact of changing note-taking preferences.

Transparency

Action Editor: D. Stephen Lindsay

Editor: D. Stephen Lindsay

Author Contributions

We conducted this study as part of an undergraduate experimental psychology course (PSY 32, Experimental Psychology) at Tufts University in the Spring 2017 semester. H. L. Urry is the professor who taught the course. C. S. Crittle, V. A. Floerke, M. Z. Leonard, and C. S. Perry, III, were graduate student members of the teaching team, listed alphabetically. The remaining authors were undergraduate students in the course, listed alphabetically. All the authors contributed to the study design. The undergraduate authors collected and analyzed the data and wrote their own empirical report in partial fulfillment of course requirements. The graduate authors facilitated the research in weekly lab sections. Formal contributions to this work according to the Contributor Roles Taxonomy (CRediT; <https://casrai.org/credit>) were as follows: All authors conceptualized the ideas and provided resources. H. L. Urry acquired funds; curated the data; conducted

formal analysis, visualization (preparation of figures), and validation (verification of analytic reproducibility); and wrote the original manuscript draft. H. L. Urry, C. S. Crittle, V. A. Floerke, M. Z. Leonard, and C. S. Perry, III, developed methodology, handled project administration, and provided supervision. All undergraduate authors contributed to investigation (data collection). H. L. Urry, V. A. Floerke, C. S. Crittle, R. S. Brody, J. P. Jimbo, E. M. Kahn, M. S. Lauzé, M. G. Lyons, A. D. Moser, C. A. Mujica, S. M. Vervoordt, and D. T. Zarrella reviewed and edited the manuscript. A number of undergraduate student authors could not be reached for inclusion in the review and editing process and to approve the final manuscript for submission; they are not, thus, listed as authors on this version of the manuscript. All the listed authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by a Faculty Research Awards Committee grant from Tufts University.

Open Practices

All data, materials, and analysis scripts have been made publicly available via OSF and can be accessed at <https://osf.io/tr868/>. The design and analysis plans for this study were preregistered at <https://osf.io/qe3wb/wiki/home/>. Deviations from the preregistration are discussed in the Supplemental Material. This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iD

Heather L. Urry  <https://orcid.org/0000-0003-4915-1785>

Acknowledgments

We are grateful to Morton Ann Gernsbacher for discussion and feedback on an earlier version of this manuscript.

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/0956797620965541>

References

- References marked with an asterisk indicate studies included in the meta-analyses.
- Allen, M., LeFebvre, L., LeFebvre, L., & Bourhis, J. (2020). Is the pencil mightier than the keyboard? A meta-analysis comparing the method of notetaking outcomes. *Southern Communication Journal*, 85(3), 143–154. <https://doi.org/10.1080/1041794X.2020.1764613>

- Aust, F., & Barth, M. (2018). *papaja: Prepare APA journal articles with R Markdown*. <https://github.com/crsh/papaja>
- Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306–307.
- Di Vesta, F. J., & Gray, G. S. (1972). Listening and note taking. *Journal of Educational Psychology*, *63*(1), 8–14. <https://doi.org/10.1037/h0032243>
- Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*(6), 600–604.
- Frantz, Z., Morling, B., & Radu, N. (2018). Conceptual replication of Mueller and Oppenheimer (2014). *PsyArXiv*. <https://doi.org/10.31234/osf.io/gkjsz>
- Gernsbacher, M. A. (2018). Three ways to make replication mainstream. *Behavioral and Brain Sciences*, *41*, Article e129. <https://doi.org/10.1017/S0140525X1800064X>
- Grahe, J. E., Reifman, A., Hermann, A. D., Walker, M., Oleson, K. C., Nario-Redmond, M., & Wiebe, R. P. (2012). Harnessing the undiscovered resource of student research projects. *Perspectives on Psychological Science*, *7*(6), 605–607.
- Hawkins, R. X. D., Smith, E. N., Au, C., Arias, J. M., Catapano, R., Hermann, E., Keil, M., Lampinen, A., Raposo, S., Reynolds, J., Salehi, S., Salloum, J., Tan, J., & Frank, M. C. (2018). Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science*, *1*(1), 7–18. <https://doi.org/10.1177/2515245917740427>
- Holstead, C. E. (2015). The benefits of no-tech note taking. *The Chronicle of Higher Education*. <https://www.chronicle.com/article/The-Benefits-of-No-Tech-Note/228089>
- Jansen, R. S., Lakens, D., & IJsselstein, W. A. (2017). An integrative review of the cognitive costs and benefits of note-taking. *Educational Research Review*, *22*, 223–233.
- *Kirkland, K. M. (2016). *The effect of note taking media and preference on the cognitive processes involved in learning* [Undergraduate honors thesis]. https://scholar.colorado.edu/honr_theses/1244
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemailcar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability: A “Many Labs” replication project. *Social Psychology*, *45*(3), 142–152. <https://doi.org/10.1027/1864-9335/a000178>
- Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. <https://doi.org/10.1177/1948550617697177>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269.
- *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*, 947–971. <https://doi.org/10.1007/s11251-018-9458-0>
- *Mitchell, A., & Zheng, L. (2017). *Examining longhand vs. laptop debate: Evidence from a replication*. In Americas Conference on Information Systems (AMCIS) Proceedings. <https://aisel.aisnet.org/amcis2017/Replication/Presentations/2/>
- *Morehead, K., Dunlosky, J., & Rawson, K. A. (2019). How much mightier is the pen than the keyboard for note-taking? A replication and extension of Mueller and Oppenheimer (2014). *Educational Psychology Review*, *31*, 753–780. <https://doi.org/10.1007/s10648-019-09468-2>
- *Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, *25*(6), 1159–1168. <https://doi.org/10.1177/0956797614524581>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. doi:10.1126/science.aac4716
- Phillips, N. (2017). *yarr: A companion to the e-book “YaRrr!: The Pirate’s Guide to R.”* <https://CRAN.R-project.org/package=yarr>
- R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.2) [Computer software]. <https://www.R-project.org/>
- Re, A. C. D. (2013). *compute.es: Compute effect sizes*. <http://cran.r-project.org/web/packages/compute.es>
- RStudio Team. (2020). *RStudio: Integrated development environment for R*. <http://www.rstudio.com/>
- Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *The Journal of Open Source Software*, *1*(3), Article 37. <https://doi.org/10.21105/joss.00037>
- Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2019). *afex: Analysis of factorial experiments*. <https://CRAN.R-project.org/package=afex>
- Torchiano, M. (2019). *effsize – a package for efficient effect size computation*. <https://doi.org/10.5281/zenodo.1480624>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3). <https://doi.org/10.18637/jss.v036.i03>
- Wage, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggins, B., & Grahe, J. E. (2019). Publishing research with undergraduate students via replication work: The Collaborative Replications and Education Project. *Frontiers in Psychology*, *10*, Article 247. <https://doi.org/10.3389/fpsyg.2019.00247>
- Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman Hall/CRC.