Two routes to the same place: learning from quick closed-book essays versus open-book essays

Kathleen M. Arnold, Emmaline Drew Eliseev, Alexandria R. Stone, Mark A. McDaniel & Elizabeth J. Marsh

Published online: 01 Apr 2021.

Submit your article to this journal

View related articles

View Crossmark data

Routledge
Taylor & Francis Group

Check for updates

Two routes to the same place: learning from quick closed-book essays versus open-book essays

Kathleen M. Arnold[a], Emmaline Drew Eliseev[b], Alexandria R. Stone[b], Mark A. McDaniel[c] and Elizabeth J. Marsh[b]

[a]Department of Psychology, Radford University, Radford, VA, USA; [b]Department of Psychology and Neuroscience, Duke University, Durham, NC, USA; [c]Department of Psychological and Brain Sciences, Washington University in St. Louis, Saint Louis, MO, USA

ABSTRACT
Knowing when and how to most effectively use writing as a learning tool requires understanding the cognitive processes driving learning. Writing is a generative activity that often requires students to elaborate upon and organise information. Here we examine what happens when a standard short writing task is (or is not) combined with a known mnemonic, retrieval practice. In two studies, we compared learning from writing short open-book versus closed-book essays. Despite closed-book essays being shorter and taking less time, students learned just as much as from writing longer and more time intensive open-book essays. These results differ from students' own perceptions that they learned more from writing open-book essays. Analyses of the essays themselves suggested a trade-off in cognitive processes; closed-book essays required the retrieval of information but resulted in lower quality essays as judged by naïve readers. Implications for educational practice and possible roles for individual differences are discussed.

We should think less about teaching students to write, and more about how we might use writing in our classrooms in the interest of learning. (Bernhardt, n.d.)

We believe that most educators and employers would agree that writing well is an important skill that can facilitate many goals, both academic and professional. Less attention, however, has been given to the idea that writing is also a strategy for learning information (*writing-to-learn*). That is, writing may help the writer to identify gaps in his/her understanding (thus serving a metacognitive function) and facilitate connections between the to-be-learned material and one's personal experiences and prior knowledge (Bangert-Drowns et al., 2004; Emig, 1977). At the same time, writing is often viewed as effortful and time-consuming by both students and educators alike – highlighting the need to understand when and how very brief writing assignments support learning, even if never graded. However, such short writing assignments are not typically the focus of academic study; a

survey of 2000 articles on writing classified less than 1% as dealing with very brief writing assignments (Stewart et al., 2010). After briefly reviewing this literature and noting some of the ways writing tasks differ, we focus on and systematically compare two versions of a quick writing task: a short essay written from memory (closed-book) versus one written with access to to-be-learned information (open-book). As developed below, we chose these tasks because closed-book essays have the potential to maximise learning by combining two active learning strategies known to promote learning in other contexts: retrieval practice (e.g. Roediger & Butler, 2011; Roediger & Karpicke, 2006), and generative learning (Fiorella & Mayer, 2015, 2016). That is, closed-book essays require the learner to retrieve information from memory and additionally to make sense of that information, elaborating upon it and organising it into an essay.

It is clear that short writing-to-learn exercises improve learning in a wide range of classes, including psychology (Butler et al., 2001; Gingerich et al.,

2014), ecology (Balgopal & Wallace, 2009), and computer science (Papadopoulos et al., 2011). The form of the writing exercise varies greatly across studies, including "minute papers" in which students reflect on what they learned in class (Stead, 2005), answering brief questions about course material (Gingerich et al., 2014), writing reflective journal entries (Connor-Greene, 2000), as well as many other forms (Friend, 2002; Voss & Wiley, 1997). In most of these studies, the comparison was to a no-writing control, or a pre–post writing comparison, as opposed to isolating key functional factors of the interventions.

Across studies, we noted variation in a factor that was seldom discussed, namely whether or not students had access to their course materials during the writing exercise. We found examples of both open-book (Balgopal & Wallace, 2009) and closed-book writing assignments (Papadopoulos et al., 2011); in some cases, we could not determine if the writing exercise was open-book or closed-book (Butler et al., 2001). We believe this factor to be an important one, and one worth drawing attention to given how little notice it receives in the literature. It is also theoretically important as closed-book (but not open-book) essays combine retrieval practice with generative activities – a combination that has yielded mixed results in past work with other tasks (Roelle & Berthold, 2017; Roelle & Nuckles, 2019; Waldeyer et al., 2020).

To date there is much more work on open- versus closed-book *testing*, as opposed to writing open-versus closed-book *essay-writing*. However, this literature has yielded contradictory results as to the relative benefits of open- versus closed-book testing on subsequent memory (see Durning et al., 2016 for a review). In the laboratory, student learning of texts (as measured by performance on delayed exams) is similar regardless of whether initial test questions were answered in open-book or closed-book fashion (e.g. Agarwal & Roediger, 2011; Agarwal et al., 2008). In the classroom, some studies show no difference between open- and closed-book exams (Gharib et al., 2012; Pauker, 1974), whereas others suggest that closed-book exams may lead to more learning (Moore & Jensen, 2007; Rummer et al., 2019; of course, these classroom findings could result from students studying more for closed-book tests, an indirect benefit of test format).

When making predictions about learning from writing open-book (not requiring retrieval) versus closed-book (dependent on retrieval) essays, we can draw on the more general theoretical and empirical literature on the effects of incorporating retrieval practice into generative learning activities. One view is that requiring retrieval should optimise the benefits of a generative learning task (we term this the *optimizing* view). This view stems straightforwardly from the large literature demonstrating the power of retrieval practice: Taking a test or otherwise retrieving information from memory (i.e. recalling the information) is more powerful for learning than a chance to restudy content material (for a review, see Roediger & Butler, 2011). The optimising view suggests that generative tasks involve elaboration and organisation, which in turn foster the construction of a coherent mental model (e.g. Waldeyer et al., 2020), thereby leading to good learning. Retrieval practice strengthens the representation (e.g. McDaniel & Masson, 1985) — perhaps through enhancing retrieval routes (Bjork, 1988) or through consolidation (Waldeyer et al., 2020) — further augmenting (optimising) the learning benefits of the generative task. Considering that essay writing is a highly generative task (encouraging elaboration, reorganisation and requiring the learner to go beyond the to-be-learned information), the optimising view clearly anticipates that closed-book essays (which require retrieval) should produce better performance on a final test than open-book essays.

Some initial evidence supports the optimising view. For instance, Roelle and Berthold (2017) compared learning from open-book versus closed-book adjunct questions (questions incorporated into a text). Both required learners to summarise their learning, but performance on a final delayed test was better in the closed-book condition. More relevant to our question is the finding that closed-book generative study prompts (that required students to identify and elaborate on the main points of texts) improved comprehension more than did answering the same prompts in open-book fashion (Roelle & Nuckles, 2019, Experiment 1).[1] It is noteworthy that learners' responses to the generative prompts contained similar numbers of idea units in the closed- and open-book conditions.

However, other findings disfavour the optimising view. With more complex adjunct questions that

---

[1]This contrast was not reported in Roelle and Nuckles (2019), but a t-test computed from the tabled values revealed a significant difference.

required inferences, learners did better on a final test when they had answered these questions in an open-book manner rather than a closed-book manner (Roelle & Berthold, 2017). Similarly, with relatively low-coherence texts, generative study prompts (described above) produced better final test performance when the prompts were answered with access to the original materials (Roelle & Nuckles, 2019, Experiment 2; see also Ebersbach, 2020, for a similar finding when students were asked to generate questions about to-be-learned material, either with or without access to the text). To explain their outcome, Roelle and Nuckles (2019) proposed that retrieval hurdles can mitigate the effectiveness of a generative study activity. According to this *retrieval-hurdles* view, closed-book prompts may reduce the number of ideas that receive generative processing, because of learners' inability to retrieve some of the information from the text. In line with the retrieval-hurdles view (see also Anderson & McDaniel, in press, Experiment 2), learners' open-book responses included more idea units from the passage than did the closed-book responses (Roelle & Nuckles, 2019, Experiment 2; in contrast to their Experiment 1 finding described above). The open-book responses were also more organised than the closed-book responses, suggesting that the cognitive processes required for retrieval can interfere with the organisational and elaborative processing fostered by the generative task. Returning to the question of essay writing, the retrieval-hurdle view anticipates that closed-book essays will yield *less learning* to the extent that closed-book essays contain less information than their open-book counterparts.

Finally, a more nuanced view is that the kind of generative activity may determine whether retrieval hurdles penalise (closed-book) generative learning activities (Waldeyer et al., 2020). Generative activities vary, for example, in how much they focus learners on specific idea units (e.g. explain how concept A is an example of X) versus allowing learners to decide which aspects of the text to focus on (pick a concept and explain why it is an example of X). The latter generative task allows learners to shift to retrievable idea units in the service of the generative activity, and as a consequence may reduce retrieval-hurdle penalties in closed-book conditions. Waldeyer et al. (2020) explored this possibility using generative prompts like "what is the most important content", "try to highlight the most important content and connections," and "try to illustrate the

most important content by giving your own examples." In two experiments learning was nearly identical in closed-book and open-book conditions. The authors suggested that open-book prompts allowed learners to cover more material (relative to closed-book) but that this advantage was balanced by the benefits of retrieval practice in the closed-book condition. For ease of exposition we term this the *balance* view. On the assumption that essays do not require the learner to focus on a particular set of idea units (at least, not more so than the prompts in Waldeyer et al.), the balance view anticipates that open-book and closed-book essays will reveal similar final test performances. Further this pattern would be accompanied by increased content (i.e. longer essays) in the open-book than closed-book essays, which would theoretically make up for the lack of retrieval practice.

To recapitulate, to our knowledge the question of whether open- or closed-book essays promote better learning has not been investigated. Nevertheless, the literature on incorporating retrieval practice (closed-book conditions) into generative learning activities offers three intriguing theoretical views that pertain to that question. These views provide reason to expect any one of three outcomes: closed-book essays will produce better learning (final test performance) than open-book essays (the optimising view), closed-book essays will penalise performance relative to open-book essays (the retrieval-hurdle view), or closed-book and open-book essays will produce equivalent learning (the balance view). To inform these possibilities, we conducted two experiments to compare the effects of writing open- versus closed-book essays in response to the same prompt. To provide insights into the underlying dynamics proposed by each of the three theoretical views, we analysed each essay for markers of cognitive processes (as described below) and then linked those characteristics to the learning observed two days later on final measures of learning (cf. Roelle & Nuckles, 2019; Waldeyer et al., 2020). We also created a metacognitive measure to capture students' perceptions of their learning, given that we know students are often unaware of the benefits of difficult learning tasks (e.g. Kornell & Bjork, 2008). In other words, we wanted to see if students were aware of any benefits observed from writing closed-book (or open-book) essays.

Several other features of this study warrant mention. First, we included a measure (based on

the Multi-Media Comprehension Battery [MMCB]; Gernsbacher & Verner, 1988) of individual differences in structure-building, which is broadly defined as people's ability to extract coherent mental structures of events or texts (Gernsbacher et al., 1990). Briefly, low-ability structure builders routinely extract fragmented and less cohesive mental structures, which in turn reduces memory for that information and leads to poorer performance on a range of learning measures (Arnold et al., 2017; Arnold et al., 2016; Bui & McDaniel, 2015; Callender & McDaniel, 2007; Lin et al., 2018; Martin et al., 2016). While correlated with reading comprehension, structure building is a separate skill that goes beyond just reading and is integral for comprehension of multiple modalities (McDaniel et al., 2002; McDaniel et al., in press). We wanted to evaluate whether any conclusions about open- versus closed-book essays were driven by low-structure builders struggling to retrieve information in the closed-book condition; specifically, low-structure builders' retrieval hurdles, if present, could lead to more prominent benefits of open-book essays (relative to closed-book) for these learners. To do so, we used an extreme-groups design in Experiment 1 (recruiting high and low structure-builders) and treated structure-building as a continuous variable in Experiment 2.

Second, students in a real class would likely be informed whether their essay-writing assignment was open- or closed-book. To approximate this authentic context and thereby glean more applied value from the study, we likewise informed the participants prior to reading the text passages on the nature of the essay task (open- or closed-book). When students know they are going to be given a closed-book exam, they tend to prepare more extensively (relative to preparing for an open-book exam), which can lead to better test performance (see Durning et al., 2016). The same may be true for closed- versus open-book essay preparation, which might reduce the retrieval hurdles and favour learning from closed-book versus open-book essays (the optimising view).

Third, a limitation of the evidence favouring the balance view (equivalent learning from open- and closed-book generative study activity) is that those experiments were not powered to detect differences that were less than large effects (as acknowledged by the authors; Waldeyer et al., 2020). To provide a more

sensitive test of possible differences, we tested samples that achieved high power to detect medium size effects of the open–closed book manipulation. Further, to provide a more general test we manipulated this variable both within- (Experiment 1) and between-subjects (Experiment 2) (see McDaniel & Bugg, 2008, for an array of memory effects that change across within- and between-subjects manipulations). Because the two experiments were so similar, we report them together.

Method

Design

Both experiments compared open-book vs. closed-book essays, but this factor was manipulated within-participants in Experiment 1 (with order counterbalanced across participants) and between-participants in Experiment 2. Furthermore, structure-building was manipulated in an extreme-groups design in Experiment 1 but treated continuously in Experiment 2.

Participants

Experiment 1

Fifty-four undergraduates from Washington University in St. Louis participated in exchange for course credit or monetary compensation. These data were collected as part of a larger study that included a separate note-taking condition ($n = 54$) that will not be discussed here given our focus on understanding differences between open- and closed-book essays.[2] All participants were invited to participate based on their scores on the Multi-Media Comprehension Battery (MMCB; Gernsbacher & Verner, 1988); eligible participants scored either a 31 or less (low-ability structure-builders; $n = 26$) or a 36 or higher (high-ability structure-builders; $n = 28$; these cutoffs follow from prior work, Callender & McDaniel, 2007; Martin et al., 2016). This sample size was chosen to be sufficient to detect a medium effect size for the main effect of open/ closed book condition, with power greater than .80. The current sample size provided power of .95 to detect both a medium effect size for the main effect of open/closed book condition and its interaction with structure-building ability, f = .25.

---

[2]Although the note-taking condition is not the focus of our study, the interested reader can find analyses of learning in this condition in Appendix A.

Experiment 2

One hundred and fourteen undergraduates from Washington University in St. Louis ($n = 50$) and Duke University ($n = 64$) participated in exchange for course credit or monetary compensation. This sample size was larger than in Experiment 1 because the essay condition was manipulated between-participants in this study (versus within-participants in Experiment 1). This choice also reflects our decision to collect data over a three-month period, with a goal of achieving power greater than .80 to detect a medium effect size using regression analysis with three between-subjects predictors (two main effects and an interaction term; current sample size provided power of .94 to detect a medium effect size, $f^2 = .15$).

Participants were randomly assigned to either an open-book ($n = 57$) or closed-book condition ($n = 57$). Unlike Experiment 1, participants were not pre-selected based on their MMCB score; the MMCB was administered during the experimental session, yielding a set of participants with a continuous range of MMCB scores. There was no difference in MMCB scores across open-book and closed-book conditions in Experiment 2 ($M = 32.4$ vs. 33.0, $t < 1$).

Materials

Passages

Participants in both studies read two passages about astronomy; one described the scientific search for extraterrestrial life (928 words; 11.8 Flesch-Kincaid grade level), and the other described different forms of solar activity (810 words; 9.9 Flesch-Kincaid grade level). Both were used in Arnold et al. (2017), and were originally created using information from an undergraduate level astronomy textbook (Karttunen et al., 2006).

Test questions

For each passage, there were eight multiple-choice questions and four problem-solving short answer questions (Arnold et al., 2017). The answers to the multiple-choice questions were either stated verbatim in the texts or required very simple inferences ("What is the size of the magnetic fields in sunspots?"). In contrast, solving the short answer problems required participants to draw connections across facts and make inferences about information in the passage (similar to Mayer & Gallini, 1990). When solving the problems about detecting life in outer space, participants were asked to imagine

they were researchers who strongly believed in the existence of extraterrestrial life. For the solar activity problems, participants were asked to imagine they were astronomers watching solar activity from their backyard. For example, the solar activity passage problem-solving questions included the following problem:

> You want to show a friend solar activity in the sky, but you do not have access to a telescope at the moment. Which of these solar activities (sunspots, faculae, eruptive prominences, solar flares) would you be most likely to be able to see? Please give two reasons to explain your answer.

This question could earn a maximum of 3 points: 1 point for correctly identifying sunspots as the most visible solar activity and 1 point for each correctly identified explanation for their answer.

Metacognitive questions

Participants answered five (Experiment 1) or four (Experiment 2) metacognitive questions using a scale from 1–5. These questions asked participants to evaluate their experience writing the essays and to compare their experience writing open- vs. closed-book essays (Experiment 1) or to predict how their experience would have differed in the other condition (Experiment 2; see Appendix B for questions).

Structure building

In both studies, we used the reading portion of the MMCB (Gernsbacher & Verner, 1988) to measure structure-building ability (following Arnold et al., 2016; Arnold et al., 2017; Bui & McDaniel, 2015; Callender & McDaniel, 2007; Callender & McDaniel, 2009; Martin et al., 2016). The MMCB contains 4 narratives (ranging from 538 to 957 words) with 12 corresponding multiple-choice questions that ask about key details in the story.

Procedure

Both studies consisted of two sessions. In Experiment 1, participants completed the MMCB task in the first session of the experiment. For this task they read four narratives at their own pace, with one to two paragraphs on the screen at a time. After each narrative, participants answered multiple-choice questions about the story. The MMCB was then scored, and, following predetermined cutoffs from prior studies (Callender & McDaniel,

2007; Martin et al., 2016; McDaniel et al., 2002). Participants with high (36 or higher) or low (31 or lower) scores were invited to participate in the rest of the experiment. Participants with scores falling in the middle of the designated range were invited to participate in a separate study not reported here.

Participants in both studies were asked to read two scientific passages. They were explicitly instructed that they would write essays to help them learn the material. All participants first read and wrote about the scientific search for life in outer space before reading and writing about different forms of solar activity. Participants read paper copies of the passages and typed their essays on the computer. Both passage-reading and essay-writing were self-paced.

In the open-book condition, participants retained the copy of the passage and could refer to it while writing their essays, whereas in the closed-book condition, participants returned the passage to the experimenter before writing their essay. Participants were informed <u>before</u> reading each passage whether they would write an open- or closed-book essay. In Experiment 1, one essay was written in the open-book condition and one in the closed-book condition, with the order of conditions counterbalanced across participants. In Experiment 2, both essays were written either in the open-book or closed-book condition, depending on random assignment.

For each essay, participants were given an essay prompt (Arnold et al., 2017). The prompt for the detecting life in outer space passage was as follows:

> Write an essay describing the indicators of life that may be used to detect other intelligent civilizations and how we have attempted to communicate with these possible civilizations. Be as clear, detailed, and thorough as possible so that a high school student who has not read the text could understand. Your essay should have an introduction and a clear thesis, and you should make sure to back up your points with supporting details.

Essays about the passage describing different forms of solar activity were written in response to the following prompt:

> Write an essay describing the different types of solar activity, including their properties, their relationships with one another, and their effects on Earth. Be as clear, detailed, and thorough as possible so that a high school student who has not read the text could understand. Your essay

should have an introduction and a clear thesis, and you should make sure to back up your points with supporting details.

After completing both passages and essays, participants answered a set of metacognitive questions (see Appendix B).

In both experiments, participants returned to the lab two days later. In this session, participants first answered multiple-choice and problem-solving questions for the passage about detecting life in outer space and then answered the questions corresponding to the passage about solar activity. After answering these questions, in Experiment 2, participants completed the MMCB task. Participants in both experiments were then debriefed and thanked for participating.

Results

Overview

We first compare final test performance across open-book and closed-book essay conditions. To preview, we found no differences in learning of the information from the scientific texts (consistent with the balance view). However, open-book and closed-book essays themselves varied in several ways, as captured in a separate section of the results. Finally, we connect these essay characteristics to test performance to speculate about the cognitive processes underlying test performance in the two conditions.

Test performance

For multiple-choice and problem-solving questions, Experiment 1 was analysed using a 2 (open- vs. closed-book) X 2 (low ability, high ability structure building) mixed analysis of variance (ANOVA), with access to the passage (open, closed) as a within-subjects factor and structure-building ability as the between-subjects factor. Because structure building was treated as a continuous variable in Experiment 2, these data were analysed using hierarchical regression, with Model 1 including two main effects: open- vs. closed-book (dummy coded: closed-book (0) and open-book (1)), and MMCB score (mean-centered) and Model 2 adding the interaction term.

Multiple-Choice questions

For Experiment 1, performance on the multiple-choice questions (see Table 1) did not differ as a

Table 1. Performance on the final test as a function of essay condition in both Experiment 1 and Experiment 2.

| | Experiment 1 | | | | Experiment 2 | | | |
| | Open-Book | | Closed-Book | | Open-Book | | Closed-Book | |
Question Type	M	SD	M	SD	M	SD	M	SD
MC questions	.60	.21	.53	.24	.60	.13	.61	.14
Problem-solving	.36	.16	.33	.14	.41	.14	.41	.12

function of access to the passage (open-book, closed-book), $F(1, 52) = 2.14$, $p = .15$, $\eta_p^2 = .04$. Similar results were obtained in Experiment 2[3]; multiple-choice test performance did not differ across open-book and closed-book conditions, $\beta = -.02$, $t < 1$.

In both experiments, structure building was associated with higher scores on the MC test. High ability structure builders outperformed low ability structure builders in Experiment 1 ($M = .62$ vs. 51), $F(1, 52) = 6.48$, $p = .01$, $\eta_p^2 = .11$. Similar results were obtained in Experiment 2, $\beta = .37$, $t(111) = 4.18$, $p < .001$. These benefits of structure-building did not depend upon essay condition; there was no interaction between essay condition and structure-building ability in either Experiment 1, $F < 1$, or in Experiment 2, where adding the interaction term did not increase the amount of variance explained, $\Delta R^2 = .01$, $F < 1$.

Problem Solving

Two independent coders scored responses to problem-solving questions, using a scoring rubric to award each problem zero to four points (depending on the question, the maximum score was two to four points; the scoring rubric was the same as that used in Arnold et al., 2017). Reliability between coders was very good (Experiment 1: Cohen's κ = .85; Experiment 2: Cohen's $\kappa_2 = .82$) and discrepancies were resolved through discussion. The final points were summed across questions and divided by the total possible points to create a problem-solving performance score.

For Experiment 1, similar to the multiple-choice question results, performance did not differ as a function of access to the passage $F(1, 52) = 2.02$, $p = .16$, $\eta_p^2 = .04$, and high-ability structure builders outperformed low-ability structure builders ($M = .40$ vs. .29), $F(1, 52) = 16.73$, $p < .001$, $\eta_p^2 = .24$. These factors did not interact, $F(1, 52) = 1.58$, $p = .22$, $\eta_p^2 = .03$.

Experiment 2 paralleled these results; performance did not differ across open-book and closed-book conditions[4], $\beta = .02$, $t < 1$, but performance increased with increasing structure building ability, $\beta = .29$, $t(111) = 3.21$, $p = .002$. These factors did not interact; adding the interaction term to the model did not increase the amount of explained variance, $\Delta R^2 = .001$, $F < 1$.

Characteristics of open-book and closed-book essays

Consistent with the balance view, open-book and closed-book essays resulted in similar levels of learning. This view posits that the retrieval hurdles that reduce the impact of generative learning in the closed-book condition are balanced by the enhancing effects of retrieval processing. A predicted consequence of this trade-off in cognitive processing is that open-book essays should contain more content and that participants were able to engage in more organisational and elaborative processing when given access to the passage. To test this hypothesis, in this section, we examine several characteristics of the essays (summarised in Table 2) to gain insight into how participants approached each type of essay, with implications for how the learning processes may have differed. In the following analyses, paired-samples t-tests were used for Experiment 1 data and independent samples t-tests were used for Experiment 2 data.

Task time

We examined that amount of time each participant spent reading and writing. It was not possible to separate time spent reading vs. writing in the open-book condition (as participants were allowed to go back and forth between reading and writing as much as they wanted), and thus the analyses are on the total amount of time engaged across the reading and essay-writing activities, to allow the open- and closed-book conditions to be directly compared. Participants spent more time on the learning activities (reading plus writing) in the open-book condition than when they read the passages, returned them, and wrote essays from memory (closed-book condition). This finding was observed in both Experiment 1 [$M = 20.8$ min vs. 16.3 min; $t(53) = 2.84$, $p = .006$, $d = .40$] and

---

[3]Model 1 was significant, $R^2 = .14$, $F(2, 111) = 8.79$, $p < .001$.
[4]Model 1 was significant, $R^2 = .29$, $F(2, 111) = 5.15$, $p = .007$.

Table 2. Characteristics of the open-book and closed-book essays in Experiments 1 and 2.

| Measures | Experiment 1 | | | | Experiment 2 | | | |
| | Open-Book | | Closed-Book | | Open-Book | | Closed-Book | |
	M	SD	M	SD	M	SD	M	SD
Task Time (min)	20.8	12.3	16.3	7.17	24.3	7.0	21.0	7.0
Content								
Word Count	303.9	103.5	233.1	93.5	370.8	122.8	310.8	115.1
Prp Content	0.25	0.08	0.17	0.06	0.32	0.10	0.19	0.05
MTurk Quality ratings	3.85	0.45	3.43	0.65	3.86	0.42	3.62	0.50

Note. For both essay conditions, task time reflects the time to read the passage combined with the time spent writing the essay, as it is not possible to separate these times in the open-book condition. Prp Content is defined as the proportion of content words from the original passage that were present in a participant's essay. MTurk Quality ratings were made on a scale from 1 (poor) to 5 (excellent).

Experiment 2 [$M = 24.3$ min vs. 21.0 min; $t(112) = 2.51$, $p = .01$, $d = .47$].

Content

Open-book essays were longer, consisting of significantly more words than closed-book essays in both Experiment 1 [$M = 303.9$ vs. 233.1 words; $t(53) = 4.89$, $p < .001$, $d = .67$] and Experiment 2 [$M = 370.8$ vs. 310.8 words; $t(112) = 2.69$, $p = .01$, $d = .50$]. To understand if the longer essays contained more scientific content, we used Linguistic Inquiry and Word Count (LIWC) software (Pennebaker et al., 2015) to count the presence of key content words in each essay, defined as all nouns, action verbs, and non-determiner adjectives from the original texts that relayed meaningful information (e.g. astronomy, transmit, ultraviolet). Our past work shows that these simple counts are highly correlated ($r = .88$) with trained scorers' estimates of the inclusion of key scientific information from the passages (Arnold et al., 2017). That is, in a previous experiment using the same materials (passages, essay prompts, and final test questions) two human coders scored each of a similar set of essays for the 73 pieces of content needed to answer the multiple-choice and problem-solving questions, and these scores were highly correlated with LIWC counts of content words. Thus the LIWC count is an excellent proxy for measuring the amount of relevant scientific content.

Content word scores (as measured by LIWC) indicated that open-book essays contained more scientific content. That is, open-book essays included a higher proportion of content words than closed-book essays in both Experiment 1 [$M = .25$ vs. .17; $t(53) = 5.17$, $p < .001$, $d = .70$] and Experiment 2 [$M = .32$ vs. .19; $t(85.9) = 8.71$, $p < .001$, $d = 1.63$], indicating that not only were open-book essays longer but that they also referenced more scientific content from the original text.

Relationship to Test Performance and Cognitive Processes.

Can this content measure be mapped onto cognitive processes? In our past work with closed-book writing tasks, we used the proportion of included passage content as a proxy for the process of *retrieving* information from memory. Retrieval is known to boost memory, which appeared to explain the advantage of closed-book writing tasks over tasks like note-taking and highlighting (Arnold et al., 2017). In contrast, the predictions are less clear for the open-book condition; even though there were minimal retrieval requirements, content might still be a marker of elaboration or a more complete essay. In the balance view, content indicates the amount of to-be-learned information that could be a target of generative processing in both open- and closed-book conditions, with more content indicating more opportunities for generative processing. In addition, in the closed-book condition only, content is an indicator of the amount of retrieval processing. To preview, the results align with this framework: Content mattered for both types of essays, but more so for closed-book essays.

To maximise power, we combined the results from the two experiments. To account for the design differences between experiments, we combined the results from the first essay written by participants in Experiment 1 with the data from Experiment 2. In this way, open-book and closed-book essays were manipulated between-participants for all included data. For multiple-choice and problem-solving performance, separate regression analyses were conducted to examine the impact of content on learning.

Regression analyses revealed that the scientific references (proportion of content words) interacted with essay condition (open-book/closed-book), for both multiple-choice [$\beta = .34$, $t(164) = 3.27$, $p = .001$] and problem-solving questions [$\beta = .27$, $t(164) =$

Table 3. Results from the regression analyses examining the effects of open- and closed-book condition and the proportion of content words included on multiple-choice and problem-solving performance using combined data from Experiments 1 and 2.

	Multiple-Choice				Problem-Solving			
Predictor	B	SE B	β	Adj. R^2	B	SE B	β	Adj. R^2
Overall Model	–	–	–	.14***	–	–	–	.13***
Open/Closed-Book	.09*	.03	.26**	–	.07*	.03	.24*	–
Content Words	.37*	.18	.21*	–	.38*	.15	.26*	–
Open/Closed-Book X Content Words	1.20**	.37	.34**	–	.79*	.30	.27*	–

Note. $N = 168$. Open-book (0) and closed-book (1) conditions were dummy coded. Content was mean-centered. *$p < .05$. ** $p < .01$. ***$p < .001$.

2.61, $p = .01$], suggesting content was a more powerful predictor in the closed-book than open-book condition (see Table 3 for full results). However, content also independently predicted performance [multiple-choice: $\beta = .21$, $t(164) = 2.09$, $p = .04$; problem-solving: $\beta = .26$, $t(164) = 2.60$, $p = .01$] suggesting that even in the open-book condition the amount of content included in an essay was somewhat associated with later test performance. Follow-up correlational analyses showed that content predicted performance in both the open-book [multiple-choice: $r = .23$, $p = .04$; problem-solving: $r = .25$, $p = .02$] and closed-book [multiple-choice: $r = .47$, $p < .001$; problem-solving: $r = .48$, $p < .001$] conditions, although, as indicated by the significant interaction, this relationship was stronger in the closed-book condition. Finally, the regression analyses also revealed that when content words were taken into account, there was a benefit of closed-book essays over open-book essays, as indicated by significant main effects for the open/closed-book condition in both multiple-choice [$\beta = .26$, $t(164) = 2.78$, $p = .001$] and problem-solving analyses [$\beta = .24$, $t(164) = 2.49$, $p = .01$]. This is consistent with the balanced view; when amount of content available for generative processing is statistically controlled, the benefits of having to retrieve that content in the closed-book condition emerged.

Quality ratings

As another measure of generative processing, we also examined organisation. An essay should be written in well-organised prose that effectively communicates information to a reader, and creating such a well-written essay likely involves generative processing. While it is almost impossible to completely separate content and writing quality, we attempted to do so by turning quality ratings over to a group of judges who did not have any basis on which to evaluate the scientific content.

That is, we used crowdsourcing to measure the quality of the essays (see Arnold et al., 2017 for use of

a similar technique). Five hundred and fifty "workers" on Amazon Mechanical Turk (MTurk) each read 5 randomly selected essays from Experiment 1 and 1,115 MTurk "workers" each read 5 randomly selected essays from Experiment 2. For each MTurk worker, the five randomly selected essays were written about the same passage. Participants were naïve in that they (1) did not know about the different experimental conditions and (2) they had never seen the original passages nor did we expect them to know much about the science described in the essays. The MTurk judges were given the following instructions:

> Imagine you are a judge in a student essay contest. All of the essays you will read were written by students after they read a passage about [different types of solar activity; the scientific search for extraterrestrial life]. They were instructed to make their essays clear, so that they could be easily read by people who did not read the same passage.
>
> Each student wrote his/her essay in one sitting and did not have an opportunity to go back to it later for proofreading and revision.
>
> You will read several essays and rate them on quality. A high quality essay should explain [different types of solar activity/the search for extraterrestrial life] in a way that is coherent, organized, interesting, and easy to read. **Please rate the essays on a scale from 1 to 5, where 1 is poor and 5 is excellent.**
>
> When making your ratings, please try to ignore the accuracy of individual facts as well as spelling and grammatical errors – just focus on the overall writing quality.

Each essay was rated by an average of 25.1 (Experiment 1) or 24.9 (Experiment 2) MTurk workers. Using a random sample of 22 (Experiment 1) or 21 (Experiment 2) ratings per responses (the minimum number of ratings for any given response), a high degree of reliability was found in both experiments. Using a one-way random effects model, the average measure intraclass correlation (ICC) was .89 with a 95% confidence interval from .86 to .92, $F(107, 2268) = 9.21$, $p < .001$ for

Table 4. Results from the regression analyses examining the effects of open- and closed-book condition and the quality of the essays on multiple-choice and problem-solving performance using combined data from Experiments 1 and 2.

Predictor	Multiple-Choice				Problem-Solving			
	B	SE B	β	Adj. R^2	B	SE B	β	Adj. R^2
Overall Model	–	–	–	.07**	–	–	–	.16***
Open/Closed-Book	.004	.03	.01	–	.01	.02	.04	–
MTurk Quality	.011***	.03	.30***	–	.12***	.02	.42***	–

Note. $N = 165$. Open-book (0) and closed-book (1) conditions were dummy coded. Quality was mean-centered. ***$p < .001$.

Experiment 1 and .87 with a 95% confidence interval from .84 to .89, $F(221, 4440) = 7.59$, $p < .001$ for Experiment 2. Ratings were averaged to create a quality score for each essay response, which, in Experiment 2, were then averaged together to create an overall quality rating for each participant.

Overall there was a difference in judged quality, such that open-book essays were rated as higher quality than closed-book essays in both Experiment 1 [$M = 3.85$ vs. 3.43; $t(53) = 5.98$, $p < .001$, $d = .8$] and Experiment 2[5] [$M = 3.86$ vs. 3.62; $t(109) = 2.81$, $p = .006$, $d = .53$].

One possible concern about these data is that open-book conditions allowed for plagiarism (direct copying of the original passage), and that heavily plagiarised essays might drive the higher ratings observed in the open-book condition. To examine this possibility, we used an open-source programme called WCopyfind (Version 4.1.5; Bloomfield, 2008) to mark word strings (of six or more words) in essays that were identical to language in the original passage. This programme provides a measure of plagiarism by calculating the percentage of total words in a writing sample that are identical (defined as part of a string of six or more matching words) to the original passage. These strings were identified based on programme parameters, which were determined using Bloomfield's (2008) recommendations and two human coders who judged the results to maximise the identification of plagiarism. The "most imperfections to allow" parameter was set at nine, meaning that the programme would connect perfectly matching strings of words separated by up to nine nonmatching words. The "minimum % of matching words" parameter was set at 70%, meaning the programme would report a string as matching if 70% or more of the prose matched the original document. We then used these plagiarism estimates to determine if more heavily plagiarised essays were rated as higher quality.

To preview, our analyses do not suggest that plagiarism drove the higher quality ratings in the open-book condition. Our analyses collapsed across experiments to maximise power and included all open-book essays (1 essay per participant in Exp. 1 and 2 essays per person in Exp. 2; $n = 168$ essays). Overall, plagiarism levels were relatively low ($M = 18.6\%$, $SD = 19.2$). More importantly, plagiarism was not significantly correlated with quality ratings, $r = -.13$, $p = .09$. In fact, there was a slight, yet non-significant, trend for a negative relationship between these variables. This result suggests that plagiarised essays were not necessarily viewed as being of higher quality, and therefore plagiarism does not appear to be the reason open-book essays were rated as higher quality.

Relationship to Test Performance and Cognitive Processes. Writing a good quality essay may involve reorganising passage information into a new structure and connecting ideas across sentences and paragraphs to create an easy-to-follow logical flow. Successfully accomplishing this task likely requires both organisational and elaborative processes, both of which are types of generative processing that benefit learning (Galbraith & Baaijen, 2018; Levin, 1988; Mandler, 1967; Rawson et al., 2015). To the degree that these processes are engaged, this learning benefit is likely to occur in both open-book and closed-book essays. Although the quality ratings are not a direct measure of the engagement of these cognitive processes, we use them here as a proxy to determine if better quality essays, as rated by naïve readers, were associated with better test performance. As with content, we combined the experiments to maximise power, using only the first essay written by each participant in Experiment 1 so that essay type would be manipulated between-participants only.

Regression analyses support our predictions; higher essay quality (as determined by MTurk

---

[5]In Experiment 2, three participants' essays were not rated for quality due to computer or experimental error.
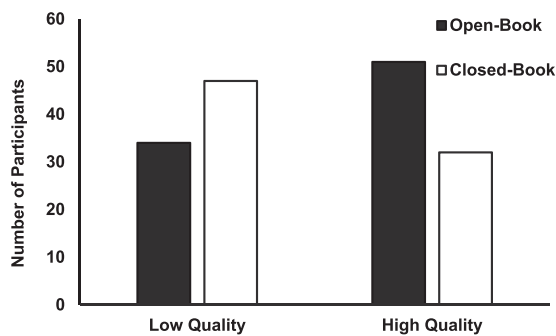
Figure 1. This histogram illustrates the frequency of participants with low- and high-quality essays ratings (by naïve MTurk judges) as a function of open-book and closed-book essay conditions collapsed across Experiment 1 and Experiment 2. Quality is divided at the median such that low-quality ratings are those below the median and high-quality ratings are those above the median. Only the first essays in Experiment 1 are included.

ratings) predicted better performance on both multiple-choice [$\beta = .30$, $t(162) = 3.80$, $p < .001$] and problem-solving questions [$\beta = .42$, $t(162) = 5.72$, $p < .001$] (see Table 4 for full results). Adding an interaction term did not increase the amount of variance explained for either multiple-choice ($\Delta R^2 = .003$, $p = .48$) or problem-solving ($\Delta R^2 = .003$, $p = .42$) questions, suggesting that this relationship was equivalent in both conditions; writing a better quality open-book and closed-book essay both benefited learning.

Instead, what differed between conditions was the *likelihood* of writing a high-quality essay, which led to the reported difference in quality ratings as a function of condition. That is, as shown in Figure 1, essays were rated across the quality spectrum, regardless of condition, but there were more high-quality essays in the open-book conditions than in the closed-book conditions. This is consistent with the balanced view; having access to the passage allowed participants to engage in more organisational processing.

Metacognitive questions

Participants were asked a series of metacognitive questions to gauge their impressions and beliefs regarding writing the essays. The majority of the questions required participants to directly compare their experiences (or their predicted experiences) writing an open-book vs. a closed-book essay, with the typical scale ranging from 1 (more closed-book) to 5 (more open-book). In

other words, a rating of 3 (the midpoint) meant the participant felt the same way about the two types of essays (see Appendix B). We quantified whether participants had a significant preference for one type of essay over the other by comparing the average rating for each question to the midpoint with one-sample t-tests. In Experiment 1, these analyses collapsed over structure-building ability, as participants with high and low structure-building ability gave similar ratings to all questions [largest difference was for the degree to which essays were helpful in identifying important information, $t(52) = 1.15$, $p = .25$, $d = .31$]. In Experiment 2, these analyses collapsed over the two essay conditions, as participants in these two conditions rated everything similarly with one exception: participants in the open-book condition rated writing essays as more helpful in identifying important information than those in the closed-book condition, $t(112) = 2.92$, $p = .004$, $d = .54$. For completeness, we report the full set of data in Table 5 for the interested reader.

Participants clearly believed that they wrote (or would have written) better quality essays with access to the original passages; in both experiments the average ratings were significantly above the midpoint of the scale [Experiment 1: $t(53) = 5.91$, $p < .001$, $d = .81$; Experiment 2: $t(113) = 22.26$, $p < .001$, $d = 2.09$]. These metacognitive beliefs can be considered accurate to the extent that they are consistent with the MTurk quality ratings.

Participants also associated a number of other positives with open-book essays. They judged them to be easier to write than closed-book essays, in both Experiment 1, $t(53) = 10.00$, $p < .001$, $d = 1.36$, and Experiment 2, $t(113) = 26.17$, $p < .001$, $d = 2.44$. Participants in Experiment 1 (who wrote both types of essays) thought they learned more from writing open-book essays, $t(53) = 2.88$, $p = .006$, $d = .39$, although this preference was not significant for participants in Experiment 2 (who only wrote one type of essay), $t(113) = 1.18$, $p = .24$, $d = .11$. Not surprisingly, given these positive perceptions, participants in Experiment 1 indicated an overall preference for open-book essays, $t(53) = 4.45$, $p < .001$, $d = .60$ (this question was not asked in Experiment 2).

Discussion

These experiments demonstrate that writing short open-book essays and short closed-book essays

Table 5. Metacognitive responses in Experiments 1 and 2 as a function of whether participants wrote open-book (OB) essays, closed-book (CB) essays, or both types of essays prior to making their responses.

Questions	Experiment 1: Within-Subjects				Experiment 2: Between-Subjects			
	Wrote both OB and CB Essays		Wrote OB Essays Only		Wrote CB Essays Only		Collapsed over Essay Condition	
	M	SD	M	SD	M	SD	M	SD
Quality of essay 1 (Higher for CB) to 5 (Higher for OB)	4.02	1.27	4.56	0.95	4.79	0.62	4.68	0.80
Difficulty of writing essays 1 (Harder to write for CB) to 5 (Harder to write OB)	1.65	0.99	1.33	0.61	1.44	0.71	1.39	0.66
Essay-writing aided Identification of important points 1 (Not at all) to 5 (Absolutely)	3.50	0.95	3.61	0.94	3.09	0.99	3.35	1.00
Amount Learned from writing 1 (More from CB) to 5 (More from OB)	3.54	1.37	3.39	1.41	2.93	1.43	3.16	1.43
Preference for type of essay 1 (Strongly prefer to write CB) to 5 (Strongly prefer to write OB)	3.81	1.35	–	–	–	–	–	–

Note. For the exact wording of questions and scale labels, see Appendix B.

can be equally effective learning activities, consistent with the balance view. Closed-book essays yield the benefit of retrieval practice: The more content learners retrieved and included in their essays, the better they did on the final test. In contrast, the open-book condition removed the retrieval hurdle, presumably allowing more effort to be devoted to elaborative and organisational processing, both of which benefit learning (Bui & McDaniel, 2015; Einstein et al., 1990; Galbraith & Baaijen, 2018; Glogger et al., 2012; Wiley & Voss, 1999). Though we do not have a direct measure of those processes, our MTurk rating of "quality" likely captured some characteristics of the final products that reflect the degree to which participants engaged with the material. This interpretation is supported by the finding that the essay quality ratings predicted learning in both open-book and closed-book conditions. Learners were better able to write highly-rated essays when they had access to the source material, suggesting that these two types of essays may present a trade-off (see Waldeyer et al., 2020, for similar findings with another generative task); closed-book essays benefit students via retrieval but students are limited in the extent to which they can engage with the material to do additional cognitive processing. Open-book essays allow students to more extensively engage with material but do not provide as clear an opportunity for students to benefit from retrieval.

Our conclusions about open- versus closed-book essays did not depend on students' structure-building ability, in either experiment. Structure-building (as measured by performance on Gernsbacher's MMCB) was very predictive of learning, consistent with past research both in the laboratory (Arnold

et al., 2017; Bui & McDaniel, 2015; Callender & McDaniel, 2007; Lin et al., 2018; Martin et al., 2016) and the classroom (Arnold et al., 2016; see McDaniel et al., in press, for a review). Structure-building predicts basic retention of textual information (and thus likely correlates with the size of the retrieval hurdle experienced in the closed-book condition) – but structure-building is also correlated with the ability to make connections, something that would be important when writing an open-book essay. Thus structure-building ability likely matters for both types of essays, albeit for different reasons.

What is clear, though, is that closed-book essays had an efficiency advantage. Students spent considerably more time reading and working on open-book essays than closed-book essays. Despite spending this additional time working with the material, students did not do any better on the final test relative to the closed-book condition. This result matters, as learning quickly has many potential advantages given that time is limited, both in and outside of the classroom. The present results suggest that closed-book essays may be the superior learning activity because students can learn the same amount of material in less time.

However, this efficiency advantage notwithstanding, students and educators alike may need convincing that closed-book essays can be effective learning activities. In Experiment 1, where participants wrote both an open-book and a closed-book essay, they rated the open-book essay as helping them learn the material better and said that they preferred writing the open-book essay. These viewpoints are consistent with that of the MTurk raters who judged the open-book essays as

being of better quality, but are concerning in that easier does not always translate into better learning (in fact, more difficult study activities often lead to superior learning, a pattern frequently referred to as desirable difficulties; Bjork, 1994). Students often misjudge these more difficult learning activities and think that because the activity is difficult, they must not be learning the material well when in fact they are (Kirk-Johnson et al., 2019; Kornell & Bjork, 2008). In the case of essays, both open-book and closed-book essays are potentially more effortful and more difficult than some other types of study activities. For this reason, students and educators may need convincing that essays can be effective learning activities at all, even though prior work has shown that they are more effective than arguably easier learning activities such as note-taking and highlighting (Arnold et al., 2017). The present results suggest students and educators may be especially skeptical of the more difficult closed-book essays, even though they appear to be more efficient if not more effective.

Although not a direct comparison, the present results are consistent with the limited prior work on open-book versus closed-book exams. Just as the present experiments found no learning differences between open-book and closed-book essays, multiple prior studies found no differences on a delayed final exam after students had taken either an initial open-book or closed-book test (Agarwal et al., 2008; Gharib et al., 2012). A common pattern found in these experiments is that students do better on the initial test when it is open-book, similar to how participants in the present studies included more content and wrote better quality essays in the open-book condition. For both open-book tests and open-book essays, having the material present allowed participants to access more content, either for purposes of answering test items or for inclusion in their essay. However, in both this study and prior studies, this particular advantage did not translate into better performance on the delayed test, likely because open-book conditions, unlike closed-book conditions, do not encourage retrieval processing, which has a powerful mnemonic benefit on the content retrieved (Roediger & Karpicke, 2006). Yet, reducing the demands of retrieval with open-book essays seemed to improve the quality of the essays (perhaps because more information was available, perhaps because cognitive resources were not needed for retrieval, or both). These observations suggest a trade-off between benefits of retrieval and the penalties imposed by retrieval demands on the other cognitive processes involved in writing to learn (the balance view).

Understanding the differential contributions to learning of various cognitive processes can provide insight into what kinds of tasks will lead to the best learning in different circumstances. Here we have found that open-book and closed-book essays can lead to equivalent learning because the relative benefits of the differentially engaged cognitive processes were equated, but that need not always be the case. For example, if students struggle with retrieving material when writing, they may not be as able to benefit from a closed-book essay. This would parallel testing-effect research, which has shown that when initial retrieval levels are low (and no feedback is provided), testing may not benefit learners (relative to restudy; McDaniel & Masson, 1985). This could occur, for instance, if the students do not have sufficient levels of initial learning prior to writing. The present studies illustrated the importance of successful retrieval when writing closed-book essays; the more content participants included in their essays (the more retrieval success they had), the better they did on the final test. Similarly, there may be conditions under which learning from closed-book essays surpasses learning from open-book essays. For example, a longer retention interval may reveal that the benefits of engaging in retrieval practice in the closed-book essay outpace the additional generative processes engaged in the open-book essay. The effect of retrieval practice increases over time (Roediger & Butler, 2011; Rummer et al., 2017), and therefore waiting longer before administering a final test may reveal an advantage for the closed-book essay, consistent with the optimising view.

Our approach to exploring the effects of writing-to-learn emphasises cognitive processes. That is, we asked not just which learning activity was better for learning, but speculated as to which cognitive processes led to learning in each writing task. This approach can provide insight for educators, allowing them to make informed decisions as to how to improve the learning impact of their writing assignments. For example, essay prompts that encourage more elaborative processing may enhance learning from essays (Voss & Wiley, 1997; Wiley & Voss, 1999). To isolate the process of retrieval, we had all students respond to the same essay prompt, regardless

of whether they were assigned to the open-book or closed-book condition. Outside of the laboratory the type of prompt or test may co-vary with open-versus closed-book essays/exams, in that open-book exams are sometimes justified as a way to get students to go beyond retention and reason at a higher level in Bloom's taxonomy (e.g. Eilertsen & Valdermo, 2000) – in other words, open-book tests are often harder (e.g. Feller, 1994) and that extra difficulty could boost learning more than any boost from retrieval practice. We cannot rebut this possibility from the present studies, but we note that the benefits of retrieval practice can hold even after the burden of retention is removed (Black-Maier, Butler, Casimir, & Marsh, in prep). However, in general we would expect the real-world instantiation of open- versus closed-book exams to differ in numerous ways, including in both teacher-led and student-directed study activities (see Ioannidou, 1997).

Although learning was found to be equivalent in the present experiments, the cognitive process approach to writing-to-learn suggests that closed-book essays have the potential to provide more learning than open-book essays. In both types of essay-writing tasks, students can learn from the cognitive processes that are involved in constructing a good quality essay. However, closed-book essays have the added advantage of retrieval processing. This additional powerful processing can only benefit students, though, if they are able to successfully retrieve the material. Anything that increases successful retrieval should help increase the learning potential of closed-book essays, including increasing students' base knowledge before assigning the essay or providing scaffolding such as outlines that would help cue students as they engage in retrieval. With these kinds of support, students should retrieve more content and thus benefit more from retrieval processing. Further, providing these supports may decrease the trade-off seen in the present experiments; by minimising the burden of retrieval, students may be able to increase the learning benefit of the other cognitive processes involved in writing an essay.

Acknowledgements

Author contributions

M. A. McDaniel and E. J. Marsh developed the initial study concept and Experiment 1 design. K. M. Arnold, M. A. McDaniel, and E. J. Marsh designed Experiment 2 and supervised data collection for both experiments. K. M. Arnold and E. D. Eliseev scored the data and collected essay quality data on Amazon Mechanical Turk. A. Stone assessed the essays for plagiarism, checked data for accuracy, and analysed data for and drafted Appendix A. K. M. Arnold analysed the data. K. M. Arnold, E. D. Eliseev, M. A. McDaniel, and E. J. Marsh drafted the manuscript.

Disclosure statement

Funding

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open-and closed-book tests. *Applied Cognitive Psychology*, *22*(7), 861–876. https://doi.org/10.1002/acp.1391

Agarwal, P. K., & Roediger, H. L. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory*, *19*(8), 836–852. https://doi.org/10.1080/09658211.2011.613840

Anderson, F. T., & McDaniel, M. A. (in press). Restudying with the quiz in hand: When correct-answer feedback is no better than minimal feedback. *Journal of Applied Research in Memory and Cognition*. Advanced online publication, 2020. https://doi.org/10.1016/j.jarmac.2020.10.004.

Arnold, K. M., Daniel, D. B., Jensen, J., McDaniel, M. A., & Marsh, E. J. (2016). Structure building predicts grades in college psychology and biology. *Applied Cognitive Psychology*, *30*(3), 454–459. https://doi.org/10.1002/acp.3226

Arnold, K. M., Umanath, S., Thio, K., Reilly, W. B., McDaniel, M. A., & Marsh, E. J. (2017). Understanding the cognitive processes involved in writing to learn. *Journal of*

Experimental Psychology: Applied, *23*(2), 115–127. https://doi.org/10.1037/xap0000119

Balgopal, M. M., & Wallace, A. M. (2009). Decisions and dilemmas: Using writing to learn activities to increase ecological literacy. *The Journal of Environmental Education*, *40*(3), 13–26. https://doi.org/10.3200/joee.40.3.13-26

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, *74*(1), 29–58. https://doi.org/10.3102/00346543074001029

Bernhardt, S. A. (n.d.). Writing as instructional practice [Blog post]. http://www.nea.org/home/34959.htm

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory II* (pp. 396–401). Wiley.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Black-Maier, A. C., Butler, A. C., Casimir, E., & Marsh, E. J. (in prep). Why does retrieval practice produce superior transfer?

Bloomfield, L. (2008). WCopyfind [computer software]. http://www.plagiarism.bloomfieldmedia.com.

Bui, D. C., & McDaniel, M. A. (2015). Enhancing learning during lecture note-taking using outlines and illustrative diagrams. *Journal of Applied Research in Memory & Cognition*, *4*(2), 129–135. https://doi.org/10.1016/j.jarmac.2015.03.002

Butler, A., Phillmann, K. B., & Smart, L. (2001). Active learning within a lecture: Assessing the impact of short, in-class writing exercises. *Teaching of Psychology*, *28*(4), 257–259. https://doi.org/10.1207/s15328023top2804_04

Callender, A. A., & McDaniel, M. A. (2007). The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology*, *99*(2), 339–348. https://doi.org/10.1037/0022-0663.99.2.339

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*(1), 30–41. https://doi.org/10.1016/j.cedpsych.2008.07.001

Connor-Greene, P. A. (2000). Making connections: Evaluating the effectiveness of journal writing in enhancing student learning. *Teaching of Psychology*, *27*(1), 44–46. https://doi.org/10.1207/s15328023top2701_10

Durning, S. J., Dong, T., Ratcliffe, T., Schuwirth, L., Artino, A. R., Boulet, J. R., & Eva, K. (2016). Comparing open-book and closed-book examinations: A systematic review. *Academic Medicine*, *91*(4), 583–599. https://doi.org/10.1097/ACM.0000000000000977

Ebersbach, M. (2020). Access to the learning material enhances learning by means of generating questions: Comparing open-and closed-book conditions. *Trends in Neuroscience and Education*, *19*, 100130. https://doi.org/10.1016/j.tine.2020.100130

Eilertsen, T. V., & Valdermo, O. (2000). Open-book assessment: A contribution to improved learning? *Studies in Educational Evaluation*, *26*(2), 91–103. https://doi.org/10.1016/S0191-491X(00)00010-9

Einstein, G. O., McDaniel, M. A., Owen, P. D., & Coté, N. C. (1990). Encoding and recall of texts: The importance of material appropriate processing. *Journal of Memory and Language*, *29*(5), 566–581. https://doi.org/10.1016/0749-596X(90)90052-2

Emig, J. (1977). Writing as a mode of learning. *College Composition and Communication*, *28*(2), 122–128. https://doi.org/10.2307/356095

Feller, M. (1994). Open-book testing and education for the future. *Studies in Educational Evaluation*, *20*(2), 235–238. https://doi.org/10.1016/0191-491X(94)90010-8

Fiorella, L., & Mayer, R. E. (2015). *Learning as a Generative Activity: Eight ways to promote learning*.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Friend, R. (2002). Summing it up. *The Science Teacher*, *69*(4), 40–43.

Galbraith, D., & Baaijen, V. M. (2018). The work of writing: Raiding the inarticulate. *Educational Psychologist*, *53*(4), 238–257. https://doi.org/10.1080/00461520.2018.1505515

Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 430–445. https://doi.org/10.1037/0278-7393.16.3.430

Gernsbacher, M. A., & Verner, K. R. (1988). *The multi-media comprehension battery*. University of Oregon, Institute of Cognitive and Decision Sciences.

Gharib, A., Phillips, W., & Mathew, N. (2012). Cheat sheet or open-book? A comparison of the effects of exam types on performance, retention, and anxiety. *Psychology Research*, *2*(8), 469–478.

Gingerich, K. J., Bugg, J. M., Doe, S. R., Rowland, C. A., Richards, T. L., Tompkins, S. A., & McDaniel, M. A. (2014). Active processing via write-to-learn assignments: Learning and retention benefits in introductory psychology. *Teaching of Psychology*, *41*(4), 303–308. https://doi.org/10.1177/0098628314549701

Glogger, I., Schwonke, R., Holzäpfel, L., Nückles, M., & Renkl, A. (2012). Learning strategies assessed by journal writing: Prediction of learning outcomes by quantity, quality, and combinations of learning strategies. *Journal of Educational Psychology*, *104*(2), 452–468. https://doi.org/10.1037/a0026683

Ioannidou, M. K. (1997). Testing and life-long learning: Open-book and closed-book examination in a university course. *Studies in Educational Evaluation*, *23*(2), 131–139. https://doi.org/10.1016/S0191-491X(97)00008-4

Karttunen, H., Kröger, P., Oja, H., Poutanen, M., & Donner, K. J. (2006). *Fundamental astronomy*. Springer.

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*, Article 101237. https://doi.org/10.1016/j.cogpsych.2019.101237

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Levin, J. R. (1988). Elaboration-based learning strategies: Powerful theory = powerful application. *Contemporary Educational Psychology*, *13*(3), 191–205. https://doi.org/10.1016/0361-476X(88)90020-3

Lin, C., McDaniel, M. A., & Miyatsu, T. (2018). Effects of flashcards on learning authentic materials: The role of detailed versus conceptual flashcards and individual differences in structure building. *Journal of Applied Research in Memory and Cognition*, *7*, 529–539. https://doi.org/10.1016/j.jarmac.2018.05.003

Mandler, G. (1967). Organization and memory. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 1, pp. 327–372). Academic Press. https://doi.org/10.106/S0079-7421(08)60516-2.

Martin, M., Nguyen, K., & McDaniel, M. A. (2016). Structure building differences influence learning from educational text: Effects on encoding, retention, and meta-cognitive control. *Contemporary Educational Psychology*, *46*, 52–60. https://doi.org/10.1016/j.cedpsych.2016.03.005

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand word? *Journal of Educational Psychology*, *82*(4), 715–726. https://doi.org/10.1037/0022-0663.82.4.715

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review*, *15*(2), 237–255. https://doi.org/10.3758/PBR.15.2.237

McDaniel, M. A., Hines, R. J., & Guynn, M. J. (2002). When text difficulty benefits less-skilled readers. *Journal of Memory and Language*, *46*(3), 544–561. https://doi.org/10.1006/jmla.2001.2819

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, *20*(4), 516–522. https://doi.org/10.1111/j.1467-9280.2009.02325.x

McDaniel, M. A., Marsh, E. J., & Gouravajhala, R. (in press). Individual differences in structure building: Impacts on comprehension and learning, theoretical underpinnings, and support for less-able structure builders. *Perspectives on Psychological Science*.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 371–385. https://doi.org/10.1037/0278-7393.11.2.371

Moore, R., & Jensen, P. A. (2007). Do open-book exams impede long-term learning in introductory biology courses? *Journal of College Science Teaching*, *36*(7), 46–49.

Papadopoulos, P. M., Demetriadis, S. N., Stamelos, I. G., & Tsoukalas, I. A. (2011). The value of writing-to-learn when using question prompts to support web-based learning in ill-structured domains. *Educational Technology Research and Development*, *59*(1), 71–90. https://doi.org/10.1007/s11423-010-9167-0

Pauker, J. D. (1974). Effect of open-book examinations on test performance in an undergraduate child psychology course. *Teaching of Psychology*, *1*(2), 71–73. https://doi.org/10.1177/009862837400100205

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *LIWC2015: Linguistic inquiry and word count*. Pennebaker Conglomerates (www.LIWC.net).

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*(4), 619–633. https://doi.org/10.3758/s13421-014-0477-z

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: The complexity of the task matters. *Learning and Instruction*, *49*, 142–156. https://doi.org/10.1016/j.learninstruc.2017.01.008

Roelle, J., & Nuckles, M. (2019). Generative learning versus retrieval practice in learning from text: The cohesion and elaboration of the text matters. *Journal of Educational Psychology*, *111*(8), 1341–1361. https://doi.org/10.1037/edu0000345

Rummer, R., Schweppe, J., Gerst, K., & Wagner, S. (2017). Is testing a more effective learning strategy than note-taking? *Journal of Experimental Psychology: Applied*, *23*(3), 293–300. https://doi.org/10.1037/xap0000134

Rummer, R., Schweppe, J., & Schwede, A. (2019). Open-book versus closed-book tests in university classes: A field experiment. *Frontiers in Psychology*, *10*, 463. https://doi.org/10.3389/fpsyg.2019.00463

Stead, D. R. (2005). A review of the one-minute paper. *Active Learning in Higher Education*, *6*(2), 118–131. https://doi.org/10.1177/1469787405054237

Stewart, T. L., Myers, A. C., & Culley, M. R. (2010). Enhanced learning and retention through "writing to learn" in the psychology classroom. *Teaching of Psychology*, *37*(1), 46–49. https://doi.org/10.1080/00986280903425813

Voss, J. F., & Wiley, J. (1997). Developing understanding while writing essays in history. *International Journal of Educational Research*, *27*(3), 255–265. https://doi.org/10.1016/S0883-0355(97)89733-9

Waldeyer, J., Heitmann, S., Moning, J., & Roelle, J. (2020). Can generative learning tasks be optimized by incorporation of retrieval practice? *Journal of Applied Research in Memory and Cognition*, *9*(3), 355–369. https://doi.org/10.1016/j.jarmac.2020.05.001

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology*, *91*(2), 301–311. Doi:10.1037/0022-0663.91.2.301

Appendices

Appendix A

As reported in the methods section, data collected in Experiment 1 included a separate note-taking condition. This condition is not relevant to our main focus of comparing learning from open-book essays vs. closed-book essays and is therefore not included in the main text. However, for completeness, the methods and results for this condition are reported below. Results include analyses comparing the two types of learning conditions (essays vs. note-taking).

In Experiment 1, an additional 54 Washington University in St. Louis undergraduate students participated in a note-taking condition. Of these, 29 participants were low structure builders (scored either a 31 or less on the Multi-Media Comprehension Battery; MMCB), and 25 participants were high-structure builders (scored a 36 or higher on the MMCB). One participant is not included in the multiple-choice analysis and another participant is not included in the problem-solving analysis due to computer error.

As with the essay condition in Experiment 1, access to the passage (open-book vs. closed-book) was manipulated within-participants. The procedure was identical to the essay condition except that participants were instructed to write notes rather than to write an essay. As in the essay condition, before reading each passage participants were explicitly told whether they would be able to reference the passage when writing their notes, with the order of conditions counterbalanced across participants. All participants read paper copies of the passages and typed their notes on the computer, in a self-paced manner.

Results

Test performance for the multiple-choice and problem-solving questions was analysed using a 2 (essay, note-taking) X 2 (open- vs. closed-book) X 2 (low ability, high ability structure building) mixed analysis of variance (ANOVA), with access to the passage (open, closed) as a within-subjects factor and learning activity and structure-building ability as the between-subjects factors.

Multiple-choice questions

As in the analysis with only the essay condition, there was no main effect of access to the passage ($M_{\text{open-book}} = .58$ vs. $M_{\text{closed-book}} = .56$) on test performance for multiple-choice questions, $F < 1$. Learning was also equivalent across essay ($M = .57$) and note-taking conditions ($M = .58$), $F < 1$. Further, there was no significant interaction between learning activity and access to the passage, $F (1, 103) = 2.28$, $p = .13$, $\eta_p^2 = .02$.

Consistent with the essay-only analyses, high-ability structure builders outperformed low-ability structure builders ($M = .62$ vs. .53), $F (1, 103) = 3.42$, $p = .07$, $\eta_p^2 = .06$. None of the interactions involving structure-building was significant, indicating that the benefits of structure-

building did not depend on access to the passage or learning activity (all Fs < 1).

Problem solving

Similar to performance on multiple-choice questions, there was no main effect of access to the passage ($M_{\text{open-book}} = .35$ vs. $M_{\text{closed-book}} = .35$) or learning activity ($M_{\text{essay}} = .35$ vs. $M_{\text{notes}} = .35$) on test performance on problem-solving questions, both Fs < 1. There was also no significant interaction between access to the passage and learning activity, $F(1, 103) = 2.36$, $p = .13$, $\eta_p^2 = .02$.

Structure-building was associated with performance on the problem-solving questions. Specifically, high-ability structure builders significantly outperformed low-ability structure builders ($M = .41$ vs. .29), $F (1, 103) = 29.28$, $p < .001$, $\eta_p^2 = .22$. Benefits of structure building did not depend on access to the passage or on learning activity as there were no significant interactions involving structure-building (all Fs < 1).

Summary

Learning was equivalent across the notetaking and essay conditions, and this pattern was consistent across both open-book and closed-book conditions. Prior literature that has compared note-taking to tasks similar to the essay condition have produced mixed results, with some having shown writing essays leading to more learning than note-taking (e.g. Arnold et al., 2017), whereas others have shown equivalent performance in both conditions (e.g. contrasting free recall vs. note-taking on problem-solving questions; McDaniel et al., 2009). The similar learning from both activities in the present study may have been in part due to the similar retrieval demands in both closed-book note-taking and essay-writing conditions.

Appendix B

Metacognitive Questions. Questions 1 was identical across studies; questions 2 and 4 were conceptually the same across studies with minor wording changes reflected whether subjects had experienced both types of essays (1) or only one (2). Subjects in both experiments estimated their learning (question 3), but the wording was different across the two studies. Only participants in Experiment 1 answered question 5, as the question was intended for subjects who had experienced both conditions.

Note. Unless otherwise noted, the question and answer scales below reflect the wording used in Experiment 1. Brackets indicate changes in the wording used in Experiment 2.

Q1. Did writing essays help you **identify** important information?

1. Not at all
2. A little
3. Somewhat

4. A lot
5. Absolutely

Q2. Do you think you wrote [would have written] a **better quality essay** when [if] the original passage was present or when [if] you relied [had to rely] on your memory for the passage?

1. Much better when [if] I relied on memory
2. Somewhat better when [if] I relied on memory
3. No difference
4. Somewhat better when [if] passage was present
5. Much better when [if] passage was present

Q3. Experiment 1 version:.
Did you **learn more** about science when you wrote an essay with the original passage present or when you wrote an essay relying on your memory for the passage?

1. Much more when I relied on memory
2. Slightly more when I relied on memory
3. Equivalent learning
4. Slightly more when passage was present
5. Much more when passage was present

Experiment 2 version:.
If you were trying to learn this material for an upcoming test in a class, which do you think would **help you learn the material better**: writing an essay while being able to refer to the passage or writing an essay while relying on your memory of the passage?

1. Relying on memory would teach me the most
2. Relying on memory would teach me somewhat better
3. No difference
4. Using the passage would teach me somewhat better
5. Using the passage would teach me the most

Q4. Which was [would be] more **difficult**: writing an essay with the passage present or relying on your memory to write the essay?

1. Much more difficult to rely on memory
2. Somewhat more difficult to rely on memory
3. No difference
4. Somewhat more difficult when [if the] passage was present
5. Much more difficult when [if the] passage was present

Q5. *Only presented in Experiment 1*: If you were **given the choice**, would you rather write an essay with the passage present or from memory? Note: the goal would be to learn the material, as opposed to receiving a grade on the essay.

1. Strongly prefer to rely on memory
2. Somewhat prefer to rely on memory
3. No preference
4. Somewhat prefer to have passage present
5. Strongly prefer to have passage present