Received: 24 May 2021 | Revised: 25 January 2022 | Accepted: 7 February 2022

DOI: 10.1002/acp.3929

WILEY

SHORT PAPER

The effect of hint strength on the benefits of retrieval practice

Kalif E. Vaughn[1] | Grant Fitzgerald[1] | Dasia Hood[1] | Karlee Migneault[1] | Kelly Krummen[2]

[1]Department of Psychological Science, Northern Kentucky University, Highland Heights, Kentucky, USA

[2]University of the Cumberlands, Williamsburg, Kentucky, USA

Correspondence
Kalif E. Vaughn, Department of Psychological Science, Northern Kentucky University, Highland Heights, KY, USA.
Email: vaughnk1@nku.edu

Abstract

Recent work has shown that retrieval hints can make test trials more enjoyable without sacrificing learning. We investigated the extent to which this effect was moderated by hint strength. Both experiments utilized a within-participants design. In Experiment 1 ($n = 41$), participants studied skeletal charts highlighting a bone region. After study, participants received one test trial with either consonants-only (e.g., _cc_p_t_l), vowels-only (e.g., o__i_i_a_), or no hint (e.g., _____). Retrieval latencies were fastest for the consonants-only items, suggesting they provided the strongest hints. Although the consonants-only trial produced the worst final test performance, participants rated that trial as the most effective, the most fun, and the trial they would use from now on. Experiment 2 ($n = 32$) replicated and extended these results using a criterion learning paradigm. These experiments show that participants prefer extremely strong hints during test trials, even when such hints impair performance.

KEYWORDS
hint strength, metacognition, retrieval practice

1 | THE EFFECT OF HINT STRENGTH ON THE BENEFITS OF RETRIEVAL PRACTICE

We have known for decades that retrieval practice improves memory (Roediger & Butler, 2011), yet little is understood about factors that might modify the benefits of retrieval practice. Recently, Vaughn and Kornell (2019) investigated how hints during test trials influenced learning. They manipulated hint strength by providing more (stronger hints) versus fewer (weaker hints) correct letters of the target word. In general, they found that providing hints did not decrease learning from test trials. Furthermore, learners strongly preferred the hints and avoided test trials when no hint was available. These results suggested that providing hints during test trials could increase rates of self-testing without impairing learning.

Although the results from Vaughn and Kornell (2019) offer an encouraging path to increasing self-testing rates among students, there was an important caveat concerning their materials. Vaughn and Kornell (2019) used carefully crafted word pairs (e.g., idea: seeker) so that even stronger hints (e.g., se__er) would not allow the target to be guessable (as *sender*, *server*, *sexier* and *seller* are all possible answers with that hint). What if the targets are guessable? In their Experiment 4, they used weakly related word pairs (e.g., whip-punish), which meant that stronger hints (whip-pu__sh) might render the target guessable. In this case, stronger hints diminished the benefits of retrieval practice, with final performance favoring items tested with weaker hints. Should hints during retrieval practice be avoided if the target might be guessable?

We continued this line of research by varying hint strength during test trials with real-world materials (medical terminology for bone structures). We hoped to answer several key questions. First, when real-world materials are used, do stronger hints reduce the benefits of retrieval practice? Second, if strong hints do diminish the benefits of retrieval practice, are students sensitive to this fact? Third, if strong hints do decrease the benefits of retrieval practice with real-world stimuli, can a criterion-learning paradigm counteract these negative effects? Although this research was mainly exploratory, there are several theoretical accounts, which are relevant to the current research (outlined below).

1.1 | Retrieval effort hypothesis

Pyc and Rawson (2009) posited that effortful but successful retrieval attempts enhance learning more than easy but successful retrieval attempts. This is known as the retrieval effort hypothesis (REH) and suggests that hints might reduce retrieval practice benefits.

Prior research appears mixed on the utility of hints provided during retrieval practice. Carpenter and DeLosh (2006) found that fewer letters provided resulted in better final memory performance, suggesting that the added retrieval effort may enhance the benefit of a retrieval attempt (Bjork & Bjork, 2011). Additionally, Carpenter (2009) found that final retention was best after testing with weakly related cues (e.g., basket–bread) versus strongly-related cues (e.g., toast–bread). van den Broek et al. (2019) found that providing orthographic hints after a failed retrieval attempt did not enhance delayed post-test performance compared to standard show-answer feedback (see also Kornell et al., 2015). These results suggest that strong hints during retrieval practice are either ineffective or harmful.

In contrast, other researchers have found utility in providing hints during practice tests. As opposed to van den Broek et al. (2019), Finn and Metcalfe (2010) found that providing hints following a failed retrieval attempt (scaffolded feedback) boosted performance. Of interest, Finn and Metcalfe (2010) initially provided only the first letter of the target following a failed retrieval attempt (with subsequent letters revealed one at a time via *accumulating feedback*), whereas van den Broek et al. (2019) provided the first and last letter of the target following a failed retrieval attempt. Although this discrepancy suggests that weaker hints distributed over time might be more potent than stronger hints provided all at once, other researchers have found benefits for stronger hints upfront via *diminishing cues*. Diminishing cues refer to practice trials in which the target word is first provided in full (e.g., dust–apyuq). During each subsequent trial, one letter is removed from the target (e.g., dust–ap_uq), with the final trial reflecting a standard test trial without hints (e.g., dust–_____). Finley et al. (2011) found that best final performance resulted from practice trials with diminishing versus accumulating cues (see also Fiechter & Benjamin, 2018).

In short, prior evidence is mixed regarding the effects of hints on retrieval practice. Although there are many potential reasons why hints might (or might not) be effective in a particular learning scenario, the amount of test practice appears to be an important factor. Note that whenever a hint is given during a test trial, there is a trade-off between retrieval effort and retrieval success: Any effort to maximize initial retrieval effort (e.g., weak initial hints via accumulating feedback) will decrease retrieval success, and any effort to improve retrieval success (e.g., strong initial hints via diminishing feedback) will reduce initial retrieval effort. If the amount of test practice is low (e.g., one test trial), retrieval effort seems to matter, as overly strong hints tend to hurt final performance (e.g., Carpenter, 2009). In contrast, if multiple test trials are offered, then both strong and weak initial hints can be effective (e.g., Finley et al., 2011; Finn and Metcalfe, 2010), although such benefits are not guaranteed (e.g., van den Broek et al., 2019).

In our experiments, we manipulated retrieval effort by manipulating hint strength. According to the REH, strong hints reduce retrieval effort and impair learning. We also explored whether the amount of test practice influences the efficacy of retrieval practice hints. In Experiment 1, participants received a single practice test trial, whereas a criterion-learning paradigm was utilized in Experiment 2. We are unaware of research exploring hints within a criterion-learning paradigm. If the amount of test practice matters, we might see harmful effects from strong hints in Experiment 1 but not in Experiment 2.

1.2 | Hints on metacognitive beliefs

By manipulating hint strength, we also manipulated retrieval fluency (the speed and ease with which information is retrievable). If information is quickly and easily retrieved, students tend to predict that they will better remember that information on a future test (Benjamin et al., 1998). In contrast, if information is not retrievable, or slowly comes to mind, students judge their learning as much worse (Koriat & Ma'ayan, 2005).

In our experiments, the strongest hints revealed all of the consonants of the target word (e.g., cl_v_cl_ for the word *clavicle*). These strong hints made retrieval much easier compared to providing weak hints (the vowels, or __a_i__e for *clavicle*) or no hint at all (e.g., _____). These strong hints likely rendered the target guessable.[i] We wondered whether students would prefer testing with such strong hints, or if they might make the test trials too easy, decreasing attention and promoting boredom.

In addition to hint preferences, we also assessed which hint was endorsed as the best for learning. In the real world, students make study decisions based on their own judgments of how well they have learned the content. Unfortunately, these judgments are often inaccurate (Bjork et al., 2013). If fluency of acquisition influences judgments of learning (Benjamin et al., 1998; Koriat & Ma'ayan, 2005), then students might (a) enjoy testing with strong hints, and (b) endorse strong hints as best for learning.

1.3 | Hints with authentic learning materials

We used both realistic study materials and realistic study paradigms. In Experiment 1, participants studied skeletal charts depicting a bone structure (e.g., ulna) and each participant received one practice test trial for each item (e.g., reflecting students who have limited study time). In Experiment 2, we had participants learn these medical terms to a criterion of one correct recall (a popular study strategy; see Kornell & Bjork, 2008).

1.4 | Overview

The current work explored the impact of strong versus weak hints during retrieval practice with authentic learning materials (medical

terminology). We attempted to answer three key questions. First, do strong hints reduce the benefits of retrieval practice with authentic materials? Second, how will learners react to these strong hints in terms of their enjoyment and perceptions of learning? Third, can equating retrieval success (via criterion learning) override the effects of hints on learning from retrieval practice?

2 | EXPERIMENT 1

2.1 | Methods

2.1.1 | Participants

Participants were recruited from a midsized university. To achieve 80% power to detect within-factors effects in a repeated-measures ANOVA, G*Power 3.1.9.7 recommends a sample size of 28 participants [assuming three measurements, $f = 0.25$, and $\alpha = 0.05$; Faul et al., 2007]. Participants were excluded if they did not finish the experiment ($n = 23$), previously participated in the experiment ($n = 11$), wished to be excluded ($n = 5$), or had zero copy performance during initial encoding ($n = 4$). In total, we had 41 participants (Mage = 18.41 years old, age range: 17–22; 33 females, four males, and four who did not list their gender) complete the experiment.

2.1.2 | Materials

Twenty-seven graphic images of the human skeleton were used. On each image, the skeleton could be seen along with an arrow pointing to a specific region (e.g., ulna).

2.1.3 | Design

Trial type was manipulated within-participants: nine images were each randomly assigned to each possible trial type (no hint, vowels-only, or consonants-only). The order of items was randomized during each phase of the experiment.

2.1.4 | Procedure

All procedures were implemented via a web-based study. Participants were informed that they would be learning anatomy terms and their corresponding location on the human body. They were instructed to type in the correct answer in the textbox and click "submit." After reading the instructions, the study phase began.

During the study phase, each skeletal image was presented on the left side of the screen. On the right side of the screen, there was a textbox and a "submit" button. Above the textbox, the correct answer was displayed. Participants were to type in the correct answer and hit "submit."

After all 27 images and terms had been studied, the test phase began. During the test phase, test trials were administered to the participant. As before, the skeletal image was presented on the left side of the screen with a textbox and a submit button on the right side of the screen. There were three possible trial types during the test phase: no-hint, vowels-only, or consonants-only. During a no-hint trial, no hint was given above the textbox. During a vowels-only trial, only the vowels of the correct answer would be displayed (e.g., o__i_i_a_). During a consonants-only trial, only the consonants of the correct answer would be displayed (e.g., _cc_p_t_l). After they had submitted their response, they were given correct-answer feedback and could advance to the next trial by clicking a button. Once all items received one test trial, the test phase ended.

After playing Tetris for 2 min, the final test phase began. The final test phase functioned the same as the other phases: the skeletal image was presented on the left side of the screen and participants were instructed to type in the correct answer in the textbox. No hints were administered during the final test phase. After all items had been tested, participants answered survey and demographic questions before the experiment terminated.

2.2 | Results

Items that were incorrectly copied during the encoding phase were discarded from all analyses.

2.2.1 | Retrieval fluency

Retrieval fluency was measured via first key press latencies, which reflect the amount of time that passes between the presentation of the test item and the typing of the first letter of their response. Stronger hints should make the retrieval attempt easier and result in faster first key press latencies.

First key press latencies[ii] by trial type are plotted in Figure 1. A repeated-measures ANOVA revealed a significant difference in latencies by trial type, $F(2,66) = 9.39$, $\mathrm{MSE} = 2{,}699{,}173.26$, $p < .001$, $\widehat{\eta}_p^2 = .222$. Post hoc tests revealed that retrieval was significantly faster with consonants trials versus vowels trials ($p = .003$) and no hint trials ($p < .001$), whereas there was no significant difference between vowels and no hint trials ($p = .481$). These results confirm that the consonants trials provided the strongest hint for participants.

2.2.2 | Performance

Performance as a function of test phase and trial type is plotted in Figure 1. A 2 (Test phase: initial or final) × 3 (trial type: consonants, no hint, or vowels) repeated-measures ANOVA revealed no significant difference in test performance by test phase, $F(1,39) = 3.85$, $\mathrm{MSE} = 0.02$,

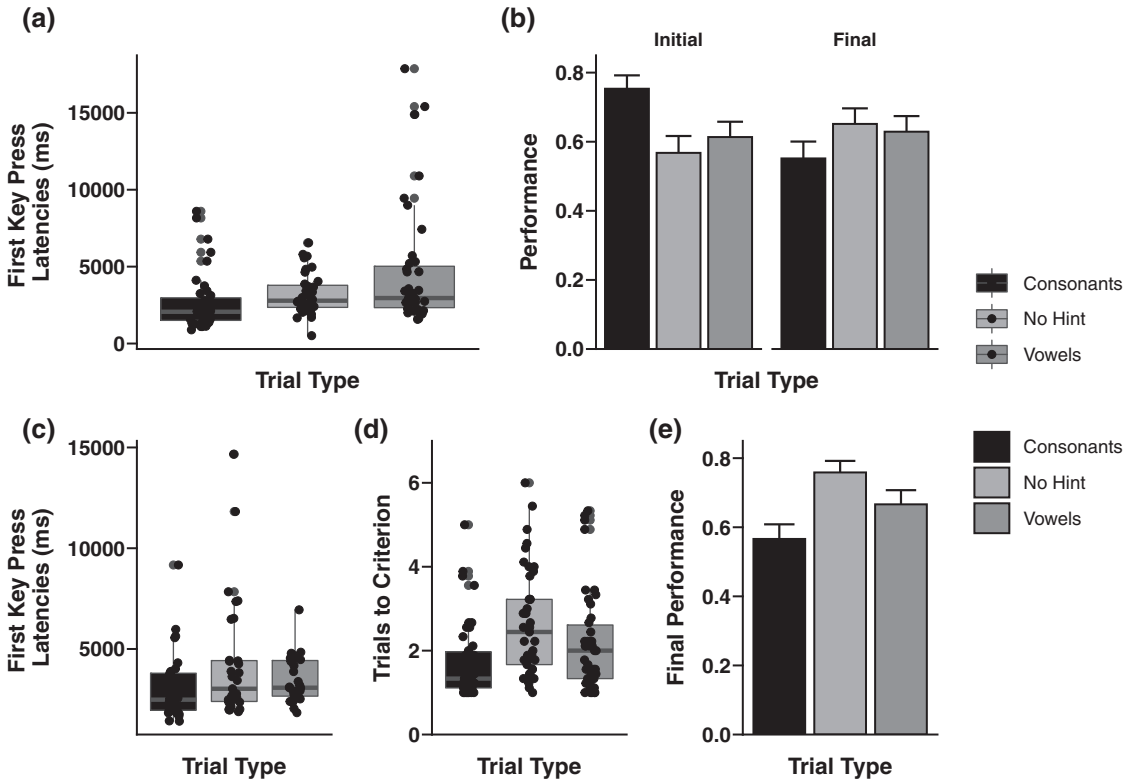

FIGURE 1 Fluency and performance in Experiment 1 (top row) and Experiment 2 (bottom row). First key press latencies are plotted in (a,c). Memory performance is plotted in (b,e). Trials to criterion (only relevant in Experiment 2) is plotted in (d). Error bars report standard error of the mean

TABLE 1 Survey results as a function of question type and trial

Experiment	Question	Consonants	No hint	Vowels
Experiment 1	Which trial type would you use from now on (assuming you could only choose one)?	28	7	6
	Which trial type was the most fun?	27	8	6
	Which trial type helped you learn the most?	24	11	6
Experiment 2	Which trial type would you use from now on (assuming you could only choose one)?	28	1	3
	Which trial type was the most fun?	22	5	5
	Which trial type helped you learn the most?	25	3	4

Note: Values reflect total number of participants. Choices were presented as either 'consonants only (e.g., cr_n__m),' 'vowels only (e.g., __a_iu_),' or 'standard test trial (no hints given).

$p = .057$, $\widehat{\eta}_p^2 = .090$ or trial type $F(2,78) = 1.97$, MSE = 0.03, $p = .146$, $\widehat{\eta}_p^2 = .048$. However, the interaction was significant, $F(2,78) = 29.37$, MSE = 0.02, $p < .001$, $\widehat{\eta}_p^2 = .430$.

A follow-up ANOVA revealed a significant difference in performance by trial type on the initial test, $F(2,78) = 14.78$, MSE = 0.03, $p < .001$, $\widehat{\eta}_p^2 = .275$.Post hoc tests revealed that the consonants trial type outperformed both the vowels and no hint trial types (*p*s < .001). There was no significant difference between the vowels and no hint trial type ($p = .385$).

A significant difference in performance also emerged on the final test, $F(2,78) = 6.00$, MSE = 0.02, $p = .004$, $\widehat{\eta}_p^2 = .133$. Post hoc tests revealed that performance was significantly lower for the consonants trial compared to the vowels and no hint trials (*p*s < .015). There was no significant difference in performance between the vowels and no hint trial ($p = .889$).

2.2.3 | Survey results

Survey results are reported in Table 1. Participants overwhelmingly endorsed the consonants trial as the best for learning [$\chi^2(2, n = 41) = 12.63$, $p = .002$], the most fun [$\chi^2(2, n = 41) = 19.66$, $p < .001$], and the trial they would use from now on [$\chi^2(2, n = 41) = 22.59$, $p < .001$].

3 | EXPERIMENT 1—SUMMARY

During practice, strong hint items (i.e., providing the consonants) were retrieved faster and had higher rates of retrieval success compared to their weak hint (or no hint) counterparts. Participants strongly endorsed the consonants trial as the most effective, most fun, and the trial they would utilize from now on, despite the consonants trial showing the worst final test performance. These results suggest that strong hints decrease the benefits of retrieval practice with real-world materials. Furthermore, participants were not only unaware that these strong hints impaired their learning, but instead believed them to be more effective.

In Experiment 2, we investigated whether a criterion-learning paradigm could result in more equivalent performance on the final test. Prior studies have shown that strong hints can be effective if multiple practice tests are administered (e.g., Finn and Metcalfe, 2010). If retrieval success is equated across items, will the strong hints still reduce the benefits of retrieval practice? Furthermore, will participants still endorse the strong hints as best for learning, or will a criterion learning paradigm alter their metacognitive beliefs?

4 | EXPERIMENT 2

4.1 | Methods

4.1.1 | Participants

Participants were recruited from a midsized university. To achieve 80% power to detect within-factors effects interaction in a repeated-measures ANOVA, G*Power 3.1.9.7 recommends a sample size of 28 participants [assuming three measurements, $f = 0.25$, and $\alpha = 0.05$; Faul et al., 2007]. Participants were excluded if they did not finish the experiment ($n = 11$), previously participated in the experiment ($n = 2$), wished to be excluded ($n = 1$), or had zero copy performance during initial encoding ($n = 3$). In total, we had 32 participants (*M*age = 25.04 years old, age range: 18–67; 19 females, 8 males, and five who did not list their gender) complete the experiment.

4.1.2 | Materials, design, and procedure

The materials, design, and procedure were identical to Experiment 1 with one key exception regarding the practice test phase. Instead of receiving one practice test trial (as in Experiment 1), items were repeatedly tested until they reached a criterion of one correct recall. If an item was not answered correctly during a test trial, the correct answer was provided for 4 s. After the feedback was displayed, the item was placed at the end of the stack of items not yet recalled successfully (to be tested again). If an item was correctly recalled, it was dropped from practice until the final test phase. Once all items had been correctly recalled one time, the criterion-learning phase ended and participants proceeded to the final test phase.

4.2 | Results

4.2.1 | Retrieval fluency

First key press latencies by trial type are plotted in Figure 1. A repeated-measures ANOVA revealed a significant difference in latencies by trial type, $F(2,52) = 8.12$, MSE = 964,040.02, $p = .001$, $\widehat{\eta}_p^2 = .238$. Post hoc tests revealed that retrieval was significantly faster with consonants trials versus vowels trials ($p < .001$) and no hint trials ($p = .015$). There was no difference between the no hint and vowels trials ($p = .978$).

4.2.2 | Trials to criterion

Mean trials to criterion as a function of trial type are plotted in Figure 1. Values beyond three standard deviations from the mean were excluded (3/128 = 2.34%). A repeated-measures ANOVA confirmed significant differences in trials to criterion by trial type, $F(2,78) = 28.77$, MSE = 0.26, $p < .001$, $\widehat{\eta}_p^2 = .425$. Post hoc tests revealed that the consonants trial required fewer trials to criterion compared to the no hint condition ($p < .001$) and the vowels condition ($p = .007$). The no hint condition required more trials to reach criterion compared to the vowels condition ($p < .001$).

4.2.3 | Performance

Performance as a function of trial type is plotted in Figure 1. A repeated-measures ANOVA revealed a significant difference in final test performance by trial type, $F(2,62) = 8.81$, MSE = 0.03, $p < .001$, $\widehat{\eta}_p^2 = .221$. Post hoc tests revealed that the consonants trial was significantly worse than the no hint trial ($p < .001$), but not worse than the vowels trial ($p = .113$). There was no difference in final performance between the no hint and vowels trials ($p = .130$).

4.2.4 | Survey

Survey results are reported in Table 1. As in Experiment 1, participants overwhelmingly endorsed the consonants trial as the best for learning [$\chi^2(2, n = 32) = 28.94$, $p < .001$], the most fun [$\chi^2(2, n = 32) = 18.06$, $p < .001$], and the trial they would use from now on [$\chi^2(2, n = 32) = 42.44$, $p < .001$].

5 | EXPERIMENT 2—SUMMARY

Once again, the strong hint items (i.e., the consonants items) were retrieved more quickly than the weak hint items (i.e., the vowels items) and no-hint items. The consonants trial also required fewer trials to criterion. As in Experiment 1, participants overwhelmingly endorsed the consonants trial as the most effective, most fun, and the trial they would utilize from now on, despite the consonants trial showing the worst final test performance.

6 | OTHER METRICS OF HINT STRENGTH

6.1 | Percentage of the target revealed

For each trial, we calculated the percentage of the target revealed. For instance, for metacarpals, more of the target is revealed via a consonants (7/11 = 63.6%) versus vowels (4/11 = 36.4%) hint. The higher the percentage of the target revealed, the stronger the hint.

First key press latencies and performance as a function of the percentage of the target revealed is plotted in Figure 2. A negative relationship was observed: As the percentage of the target revealed increased, latencies and final performance tended to decrease. Regarding latencies, this relationship was significant in both Experiment 1 [$b = -45.86$, 95% CI [−78.43,−13.29], $t(403) = -2.77$, $p = .006$] and Experiment 2 [$b = -44.77$, 95% CI [−79.01,−10.54], $t(336) = -2.57$, $p = .011$]. Regarding performance, this relationship was not statistically significant in Experiment 1 [$b = -3.02$, 95% CI [−7.00,0.95], $t(444) = -1.49$, $p = .136$], but was in Experiment 2 [$b = -7.81$, 95% CI [−12.49,−3.12], $t(351) = -3.28$, $p = .001$]. Overall, these results are consistent with the prior analyses showing that stronger hints tend to decrease latencies and impair learning.

6.2 | First letter visibility

For each item (e.g., metacarpals), we determined whether the hint displayed the first letter of the target (e.g., consonants—yes; vowels—

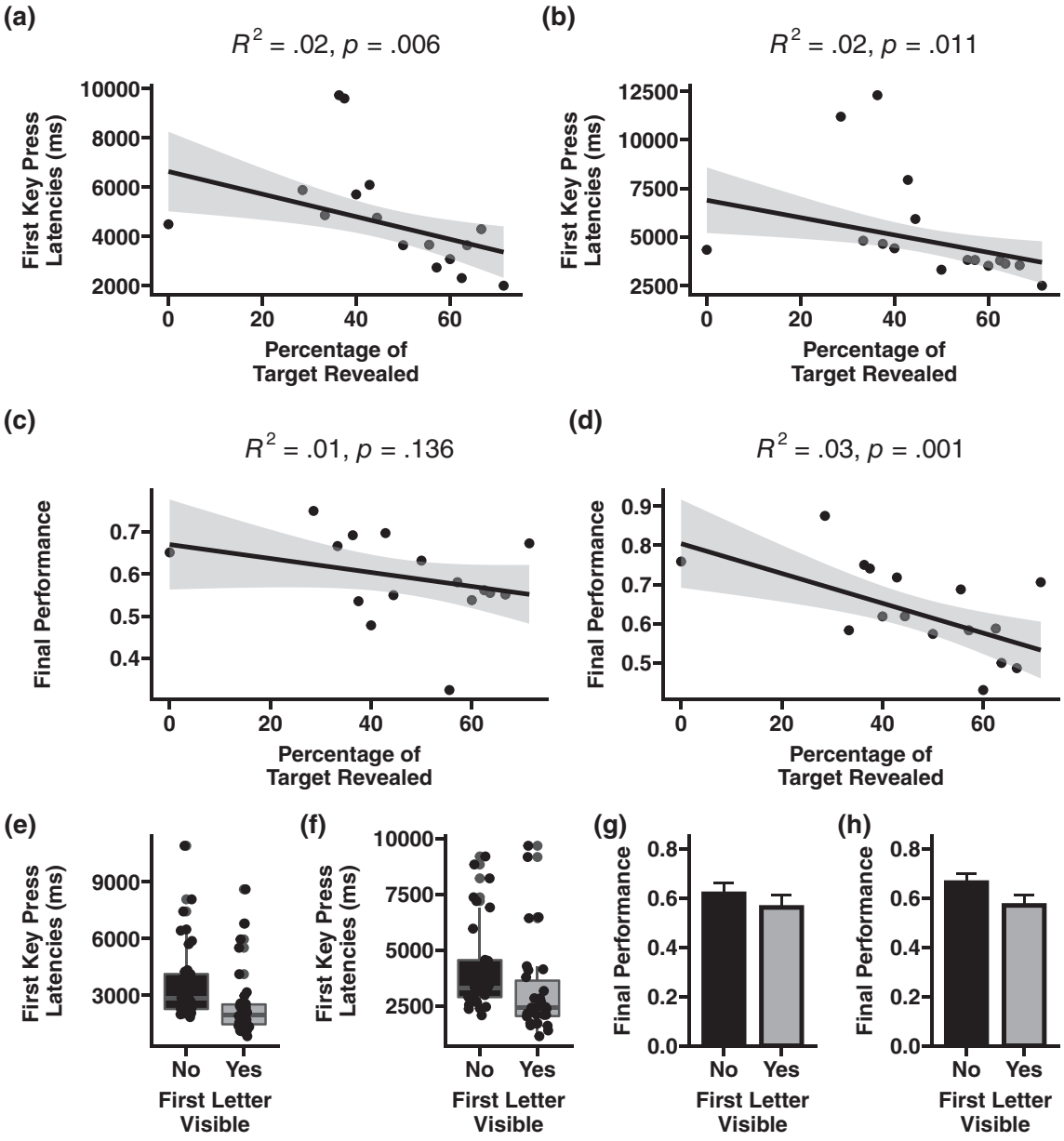

FIGURE 2 First key press latencies and final test performance as a function of the percentage of the target revealed in Experiment 1 (a,c) and Experiment 2 (b,d), respectively. First key press latencies and final test performance as a function of first letter visibility in Experiment 1 (e,g) and Experiment 2 (f,h), respectively. Error bars report standard error of the mean

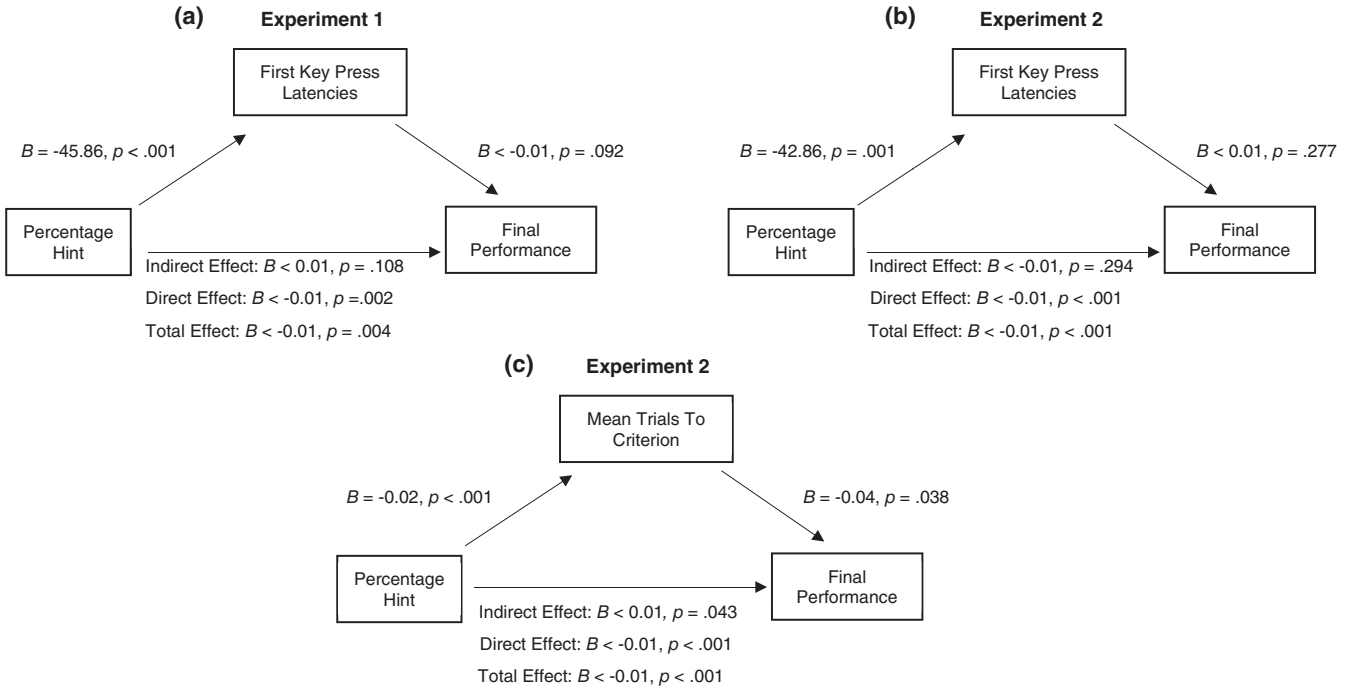

FIGURE 3 Mediation models testing whether the percentage hint on final performance was mediated by first key press latencies in Experiment 1 (a) and Experiment 2 (b), or by mean trials to criterion in Experiment 2 (c)

no). Research suggests that knowing the first letter of the solution can be advantageous (e.g., Witte & Freund, 2001). If first letter visibility provides a strong clue to the target, it might be observed via faster first key press latencies. To explore this, we calculated mean first key press latencies split by first letter visibility (see Figure 2). Outliers, undefined values, and participants without a latency value for both first letter conditions (visible or not visible) were removed from the analyses. Participants recalled items significantly faster when the first letter was visible in Experiment 1 [M_d = 1190.24, 95% CI [897.01,1483.48], $t(36 = 8.23)$, $p = .001$] and Experiment 2 [M_d = 988.35, 95% CI [409.45,1567.25], $t(29 = 3.49)$, $p = .002$]. Additionally, participants recalled items worse on a final test if they were shown the first letter in Experiment 1 [M_d = 0.06, 95% CI [0.00,0.11], $t(39 = 2.22)$, $p = .032$] and Experiment 2 [$M_d = 0.09$, 95% CI [0.02,0.117], $t(31 = 2.49)$, $p = .018$] (see Figure 2).

These outcomes are consistent with our prior results: First letter visibility provided a stronger clue to the target, which decreased retrieval latencies and impaired final test performance.

7 | MEDIATION ANALYSIS

We conducted an exploratory mediation analysis to observe whether percentage hint on final performance was mediated by first key press latencies (see Figure 3).[iii] Results suggest that percentage hint influenced final performance, but this effect was not mediated by first key press latencies. In Experiment 2, the criterion paradigm caused variability in the mean number of test trials, which partially mediated percentage hint on final performance (see Figure 3). Although consistent with our primary analyses showing that strong hints reduced final performance, these results suggest that differences in retrieval effort might not always be reflected by differences in first key press latencies.

8 | GENERAL DISCUSSION

Across two experiments, participants learned proper names for bone structures in human anatomy (e.g., cranium). They either received a fixed amount of practice (Experiment 1) or learned the items to a criterion of one correct recall (Experiment 2). Across both experiments, the consonants trial was consistently the most fluent trial, producing the best initial test performance but the worst final test performance. Despite this pattern, participants consistently endorsed the consonants trial as the most fun, the most effective, and the trial that they would choose from now on. Secondary analyses explored other metrics of hint strength and showed that performance tended to decrease (a) when a greater percentage of the target was revealed during the hint trial and (b) when the first letter of the target was known versus unknown. Overall, these results confirm that strong hints impair learning with real-world materials. Metacognitively, participants are completely unaware of the dangers of testing with strong hints, and instead endorsed the strong hints as best for learning. Lastly, these results suggest that equating retrieval success via a criterion learning paradigm does not override or otherwise alter the weaker benefits gained from testing with strong hints.

8.1 | Implications

These results are important for several reasons. We used real-world materials that a medical student might encounter early in their learning career. With these real-world items, strong hints not only diminished the benefits of retrieval practice, but also produced strong illusions of learning. These results provide a caveat to Vaughn and Kornell (2019). Vaughn and Kornell (2019) found that providing strong hints improved performance while making the test trial more enjoyable; however, the items were unguessable (in Experiments 1–3) due to carefully crafted materials. In Experiment 4, Vaughn and Kornell (2019) showed that strong hints can be harmful with weakly related word pairs. Our results conceptually replicated these findings with real-world materials. Students who find the material challenging might be tempted to test themselves with hints. Our results suggest that such hints must be used with caution, as strong hints will likely reduce retrieval practice benefits.

In addition, a criterion-learning paradigm did not alter our pattern of results. This is somewhat surprising given that other researchers have shown that strong initial hints can be beneficial when multiple retrieval opportunities are provided Fiechter and Benjamin (2018). One key difference between studies investigating diminishing feedback and our criterion learning paradigm is that with diminishing feedback, retrieval effort increased over time (as letters were removed from the target word), whereas retrieval effort stayed more constant in our experiment (as the hints never changed). If we had utilized repeated retrieval practice (e.g., by setting a higher criterion level), a different pattern might have emerged.

This pattern of results is consistent with several theoretical accounts. Participants strongly endorsed the strong hints as better for learning, which replicates prior work showing that learners rely heavily on retrieval fluency to gauge their own learning (Benjamin et al., 1998; Koriat & Ma'ayan, 2005). As such, students who test themselves with hints might be especially prone to make sub-optimal study decisions (see Rivers, 2021 for discussion). Additionally, testing was more beneficial when the retrieval attempts were more challenging, which is consistent with the Retrieval Effort Hypothesis (Pyc & Rawson, 2009).

8.2 | Limitations

There were several limitations in our experiments. First, we used real-world materials, and students might have been familiar with some of the items (e.g., ulna), although we would argue that most were likely unknown to the participants (e.g., trochlea, maxilla, and calcaneus). Even if participants had heard of some of these terms before, they likely did not know the precise location of that bone structure on a human anatomy chart. Furthermore, hint types were randomly assigned to the items, meaning that these potentially familiar items had an equal chance of receiving no hint, a weak hint, or a strong hint. Even if the strong hints made certain items even easier due to prior knowledge, participants still endorsed those strong hints as best for learning.

A second limitation is that final performance was assessed after a 2-min delay within the same learning session. Would a different pattern emerge with longer retention intervals, or would the reduced potency of testing with strong hints be exacerbated further across longer retention intervals? We argue that it would likely depend upon the specific implementation of those strong hints, as researchers have found long-term retention benefits after testing with strong hints using a diminishing cues paradigm (e.g., Fiechter & Benjamin, 2018).

A third limitation is that we found no evidence of first key press latencies mediating the relationship between the percentage hint on final performance. We primarily used latencies as our metric for fluency during the test trials, and we found evidence that revealing more of the target increased fluency. However, the lack of mediation suggests that variations in retrieval effort might not correspond directly to variations in latencies. Future research could examine this relationship more closely.

A fourth limitation is that we had moderate levels of dropout across both experiments. This is perhaps more common in web-based (versus laboratory-based) experiments. Importantly, even after excluding these individuals, both studies were sufficiently powered to detect our effects.

9 | CONCLUSION

Prior work has shown that testing with hints can make retrieval more enjoyable without decreasing learning (Vaughn & Kornell, 2019), but that same work provided preliminary evidence that hints might harm learning if the target becomes guessable (see their Experiment 4). We continued this line of research and found evidence consistent with that claim. Although retrieval practice is a beneficial study strategy, students learning real-world course material might want to avoid testing with hints, especially strong hints. Further research is needed to discover optimal hints with real-world materials.

CONFLICT OF INTEREST

The authors report no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Kalif E. Vaughn https://orcid.org/0000-0003-0710-7647

ENDNOTES

[i] Although we did not explicitly measure guessability, we provide additional analyses later which quantified the strength of each hint in terms of the percentage of the target revealed. Arguably, the higher the percentage of the target revealed, the more guessable the target becomes. Supporting the notion that the consonants trials provided the strongest

hints, a higher percentage of the target was revealed (on average) with consonant versus vowel hints [$\Delta M = 14.88$, 95% CI [10.42,19.33], $t(52.00) = 6.70$, $p < .001$].

[ii] To minimize the influence of aberrant values, median first key press latencies were obtained for each participant for each factor of interest (e.g., trial type). Next, outliers (values greater than three standard deviations from the mean of those medians) were removed for Experiment 1 (6/119 = 5.04%) and Experiment 2 (6/96 = 6.25%). Finally, mean values were computed (a mean of the medians) across participants.

[iii] We removed latencies three SD beyond the mean in Experiment 1 (12/417 = 2.88%) and Experiment 2 (14/353 = 3.97%). We also removed extreme trials to criterion in Experiment 2 (4/353 = 1.13%).

REFERENCES

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55–68.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In: Gernsbacher, M. A., Pew, R. W., Hough, L. M., & Pomerantz, J. R., (eds.), *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* (pp. 59–64). New York, NY: Worth.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. In *Annual review of psychology* (Vol. 64, pp. 417–444). Annual Reviews Inc.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition*, *34*(2), 268–276.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fiechter, J. L., & Benjamin, A. S. (2018). Diminishing-cues retrieval practice: A memory-enhancing technique that works when regular testing doesn't. *Psychonomic Bulletin and Review*, *25*(5), 1868–1876.

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*(4), 289–298.

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory and Cognition*, *38*(7), 951–961.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, *52*(4), 478–492.

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—And costs—Of dropping flashcards. *Memory*, *16*(2), 125–136.

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447.

Rivers, M. L. (2021). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review*, *33*, 823–862.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

van den Broek, G. S., Segers, E., Van Rijn, H., Takashima, A., & Verhoeven, L. (2019). Effects of elaborate feedback during practice tests: Costs and benefits of retrieval prompts. *Journal of Experimental Psychology: Applied*, *25*(4), 588–601.

Vaughn, K. E., & Kornell, N. (2019). How to activate students' natural desire to test themselves. *Cognitive Research: Principles and Implications*, *4*(1), 35.

Witte, K. L., & Freund, J. S. (2001). Single-letter retrieval cues for anagram solution. *The Journal of General Psychology*, *128*(3), 315–328.

How to cite this article: Vaughn, K. E., Fitzgerald, G., Hood, D., Migneault, K., & Krummen, K. (2022). The effect of hint strength on the benefits of retrieval practice. *Applied Cognitive Psychology*, 1–9. https://doi.org/10.1002/acp.3929