


# Children's Retrieval of Science Facts: The Role of Hints and Confidence

Elisabeth C. McLane & Diana Selmecky


To cite this article: Elisabeth C. McLane & Diana Selmecky (25 Sep 2024): Children's Retrieval of Science Facts: The Role of Hints and Confidence, Memory, DOI: [10.1080/09658211.2024.2406312](https://doi.org/10.1080/09658211.2024.2406312)

To link to this article: <https://doi.org/10.1080/09658211.2024.2406312>


 [View supplementary material](#)


 Published online: 25 Sep 2024.


 [Submit your article to this journal](#)

 [View related articles](#)

 [View Crossmark data](#)

 This article has been awarded the Centre for Open Science 'Open Data' badge.

 This article has been awarded the Centre for Open Science 'Open Materials' badge.

 This article has been awarded the Centre for Open Science 'Preregistered' badge.



## Children's Retrieval of Science Facts: The Role of Hints and Confidence

Elisabeth C. McLane and Diana Selmeczy

Department of Psychology, University of Colorado, Colorado Springs, CO, USA

### ABSTRACT

The effortful process of retrieving information from memory has been established as an effective strategy for improving student learning. However, we have a limited understanding of the development of retrieval practice in children, including contexts that may scaffold its benefit. In the current pre-registered study, we examined whether the use of hints during retrieval practice improved free recall in an online science learning task in 8- to 13-years-olds ( $N=77$ ,  $N_{\text{females}}=36$ ). We found partial evidence supporting the provision of hints as boosting the benefit of retrieval practice. Children's long-term retention of science facts was higher when they received hints during an earlier practice test compared to restudying information, but not compared to a test only condition without hints. Furthermore, we found similar effects across both age and levels of confidence, suggesting that retrieval practice remains stable across these factors.



### KEYWORDS


Testing effect; retrieval practice; children; hints; confidence

The effortful process of retrieving previously learned information, termed retrieval practice, is a highly beneficial strategy that improves long-term retention of information (Fazio & Marsh, 2019; Roediger & Butler, 2011). Extensive research demonstrates the effectiveness of retrieval practice in college-aged adults (see Roediger & Butler, 2011 for a review). The majority of this literature demonstrates that test-taking is a particularly effective form of retrieval practice and leads to much greater retention than simply restudying information (i.e., the testing effect) (Eisenkraemer et al., 2013; Roediger & Karpicke, 2006; Rowland, 2014) in both laboratory and classroom settings (Brame & Biel, 2015). Although testing is clearly beneficial in adults, we have relatively limited knowledge of the development of the testing effect in children and adolescents (see Fazio & Marsh, 2019 for a review), including what contexts may scaffold its benefit. Given the relevance of the testing effect for educational contexts, including primary education (Bouwmeester & Verkoeijen, 2011; Goossens et al., 2014; Jaeger et al., 2014; Karpicke et al., 2014, 2016; Lipowski et al., 2014; Ma et al., 2020), it is important to understand how its benefit may change across development. In the current study, we examined the development of retrieval practice in children ages 8- to 13-years-old during the learning of science facts. Critically, we examined whether scaffolding the retrieval process through the use of hints improved retention of science facts compared to test and study conditions without hints. Additionally, we examined if the benefit of testing changed as a

function of children's subjective assessments of their memory quality (i.e., confidence). Understanding how the testing effect improves throughout development and what factors contribute to these changes is vital to supporting children's knowledge acquisition and educators' ability to implement age-appropriate learning strategies.

The testing effect has been observed in children as young as preschoolers (Fritz et al., 2007; Kliegl et al., 2018), as well as elementary through high-school students (Bouwmeester & Verkoeijen, 2011; Carpenter et al., 2009; Goossens et al., 2014; Jaeger et al., 2014; Karpicke et al., 2014, 2016; Lipowski et al., 2014; Ma et al., 2020; McDaniel et al., 2011, 2013; McDermott et al., 2014; Roediger et al., 2011). However, previous research often focuses on one developmental period (e.g., Carpenter et al., 2009; Fritz et al., 2007; Goossens et al., 2014; Jaeger et al., 2014; Karpicke et al., 2014, 2016; McDaniel et al., 2011, 2013; Roediger et al., 2011) and therefore the developmental progression of the testing effect is less well-known (Bouwmeester & Verkoeijen, 2011; Halperin, 1974; Kliegl et al., 2018; Lipowski et al., 2014; McDermott et al., 2014). Critically, the testing effect is not always evident in children (Bouwmeester & Verkoeijen, 2011; Karpicke et al., 2014; Kliegl et al., 2018) with research suggesting the benefit of testing may be constrained by task difficulty or retrieval demands. For example, in preschool and elementary students the testing effect is much larger or only observed under contexts that scaffold retrieval, such as when provided with immediate feedback and using recognition or

**CONTACT** Diana Selmeczy  [diana.selmeczy@uccs.edu](mailto:diana.selmeczy@uccs.edu)  Department of Psychology, University of Colorado, 1420 Austin Bluffs Parkway, Colorado Springs, CO 80918, USA

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/09658211.2024.2406312>.

© 2024 Informa UK Limited, trading as Taylor & Francis Group

cued-recall tests (Fritz et al., 2007; Kliegl et al., 2018; Lipowski et al., 2014; Ma et al., 2020). However, when the context is too demanding and very little information can be retrieved, younger children often do not exhibit the testing effect (Karpicke et al., 2014; Kliegl et al., 2018). In contrast, high-school students (Cranney et al., 2009; McDermott et al., 2014) and adults (Kang et al., 2007; Karpicke & Roediger, 2007; Pashler et al., 2005; Roediger & Karpicke, 2006) exhibit the testing effect under a wider range of circumstances, including more challenging conditions such as free recall tests and tests without feedback (Brame & Biel, 2015). Critically, adults are also less likely to benefit from testing when retrieval is easy or performance is high (Carpenter & Delosh, 2006; Pyc & Rawson, 2009), suggesting that testing benefits only occur when learners must engage in a sufficient level of effortful retrieval.

Taken together, the above findings suggest differing retrieval demands may impact the development of the testing effect. This prediction is consistent with retrieval effort accounts of the testing effect which posit that successful retrieval under difficult contexts leads to better retention than successful retrieval under easy contexts or unsuccessful retrieval (Bjork, 1999; Pyc & Rawson, 2009). Thus, one reason young children may exhibit limited benefits of testing compared to older children and adults is because certain difficult contexts (e.g., when no feedback is provided or limited retrieval cues are available) may lead to more unsuccessful retrievals (Ghetti & Angelini, 2008; Halperin, 1974). Thus, supporting successful retrieval while still maintaining effortful engagement is likely critical to observing the testing effect particularly in younger children. Importantly, one effective way to support retrieval may be through building semantic associations. Previous research demonstrates that the testing effect occurs, in part, because testing can generate semantically related content (Carpenter et al., 2009) which can then be used as a retrieval cue during a final test (Pyc & Rawson, 2010). For example, when asked to recall how cats are superior to humans, retrieving related content during a practice test (e.g., cats have very large eyes relative to their head size) can then be used as a retrieval cue when recalling the target fact on a final test (e.g., cats have superior peripheral vision compared to humans). Consistent with this idea, Karpicke et al. (2014) showed that when children's retrieval is effortful but scaffolded through semantic cues (e.g., questions maps), they experience greater retention after a short delay relative to restudying information. However, this previous study investigated a small age range, did not directly examine test conditions in the absence of cues, and used a short delay period during which testing effects are less likely to occur (Roediger & Karpicke, 2006).

In addition to hints impacting the testing effect, the testing effect may also vary across other conditions such as levels of confidence. Confidence levels can serve as an important indicator of when retrieval might feel difficult.

Thus, lower confidence may serve as instances during which retrieval practice may be particularly beneficial. Indeed, confidence is positively associated with retrieval success in both adults (DeSoto & Roediger, 2014; Kleitman & Stankov, 2007) and children (Ghetti et al., 2002; Kleitman & Gibson, 2011; Roebbers, 2002; Roebbers et al., 2014). Furthermore, confidence levels are linked to retrieval effort with research showing that retrieval fluency (measured as faster response times) is associated with higher confidence in both adults (Hu et al., 2022) and children (Koriat & Ackerman, 2010). However, different theories of the testing effect suggest potentially differing predictions regarding how retrieval benefits may interact with levels of confidence. On the one hand, the retrieval effort account suggests the testing effect may be larger when successful retrieval occurs during low compared to high confidence due greater retrieval difficulty during lower confidence (Pyc & Rawson, 2009). On the other hand, more retrieval cues may be generated during high confidence responses since accuracy is typically higher and more details are retrieved (Roediger et al., 2012). Thus, it is also possible that the testing effect may be larger under higher confidence. Research on the interaction between confidence and the testing effect has been limited and only conducted with adults, with results demonstrating that the testing effect is larger during high confidence recognition responses, but not free recall responses, of word-pairs (Zhang et al., 2019). Since the relation between memory and confidence improves with development (Fandakova et al., 2017), it is critical to examine whether similar patterns would emerge with school-aged children. Thus, we examined the role of confidence on the testing effect in children to help further our understanding of the contexts that influence retrieval practice across development.

In the current pre-registered study (<https://osf.io/43ynf>), we investigated whether providing children ages 8- to 13-years-old with retrieval support, in the form of semantic hints, led to better long-term retention of animal science facts than testing without hints or simply restudying facts. Hints indicated the general category of learned animal facts (i.e., eyes, moves, or eats; see Shields et al., 2024) based on previous research suggesting that weakly related semantic cues can encourage elaborative retrieval (Carpenter et al., 2009). The provision of category cues are also effective at improving recall in children (Kobasigawa, 1974) and the use of helpful hints or clues are an effective pedagogical tool for promoting learning (Karabenick & Gonida, 2017). By using broad categorical hints during a practice test, we hoped to increase retrieval success while still encouraging effortful retrieval since the hints were only weakly associated to the target fact and associated with multiple target facts. Overall, we predicted that children's long-term retention on a final cued-recall test (~24 hours after learning) would be higher under conditions when they engaged in retrieval during an earlier practice test, including being tested on information with hints or without

hints, compared to restudying information. This prediction was based on previous research demonstrating benefits of retrieval practice during middle childhood (Fazio & Marsh, 2019). Critically, we predicted earlier presented hints would be beneficial particularly for younger children's retention, since younger children may have greater difficulty spontaneously generating helpful retrieval cues (Ackerman, 1982; Kobasigawa, 1974). Thus, we expected that younger children would demonstrate higher final test accuracy on hint compared to test conditions, while older children would demonstrate similar performance across these two conditions. Finally, we examined whether retrieval benefits would differ as a function of children's confidence level. We predicted that the effect of practice test condition (hint, test, vs. study) would be more robust for lower levels of confidence, with the expectation that retrieval demands would be most difficult under lower confidence conditions. Given the novelty of this research, our pre-registration simply predicted that children's final test performance during lower confidence would be greatest during conditions that required retrieval practice (hint and test conditions) compared to restudy.

## Methods

### Transparency and Openness

The study methods and analyses on final test performance were preregistered (<https://osf.io/43ynf>). Data and materials are openly available at <https://osf.io/rx32s/>. The task is openly available online at <https://app.gorilla.sc/openmaterials/661279>. R version 4.2.0 was used for data analyses.

### Participants

The target sample size of 73 participants was determined using an a-priori power analysis ( $\alpha = .05$ ,  $1 - \beta = .90$ ) assuming a medium effect size for the interaction between age and condition ( $f^2 = .15$ , total number of predictors = 3, number of predictors tested = 1 interaction term). Given limited research on the testing effect across this age range, an expected effect size was difficult to estimate. Therefore, we determined the proposed medium effect size to be the smallest effect of interest for our current study. Furthermore, post-hoc power analyses showed that the current sample size could detect a two-tailed within-participant difference (e.g., study vs. test) of Cohen's  $d$  equal to  $\sim .30$  with 80% power.

Eligibility to participate required that children be 8- to 13-years-old, have no developmental disorders, and live in the United States. Our final sample after exclusions consisted of 77 children ( $N_{females} = 36$ ,  $N_{males} = 39$ ,  $N_{notreported} = 2$ ,  $M_{age} = 10.66$ ,  $SD_{age} = 1.67$ ). Data from an additional 6 children were collected and excluded due to not completing the task.

Participants' race was reported as 66.2% White, 2.6% African American, 15.6% Asian, 13% multi-racial, and 2.6% not reported. Ethnic background was reported as 9.1% Hispanic, 89.6% not Hispanic, and 1.3% not reported. Participants' reported family income was distributed as less than \$15,000 (3.9%), more than \$15,000 and less than \$25,000 (0%), more than \$25,000 and less than \$40,000 (3.9%), more than \$40,000 and less than \$60,000 (10.4%), more than \$60,000 and less than \$90,000 (18.2%), more than \$90,000 (57.1%), and not reported (6.5%).

Participants were recruited with paid online Facebook advertisements distributed to U.S. residents and flyers distributed at community events and in local public schools. Participants were compensated with a \$15.00 online electronic gift card per one-hour session, for a total of \$30.00.

## Materials

### Stimuli

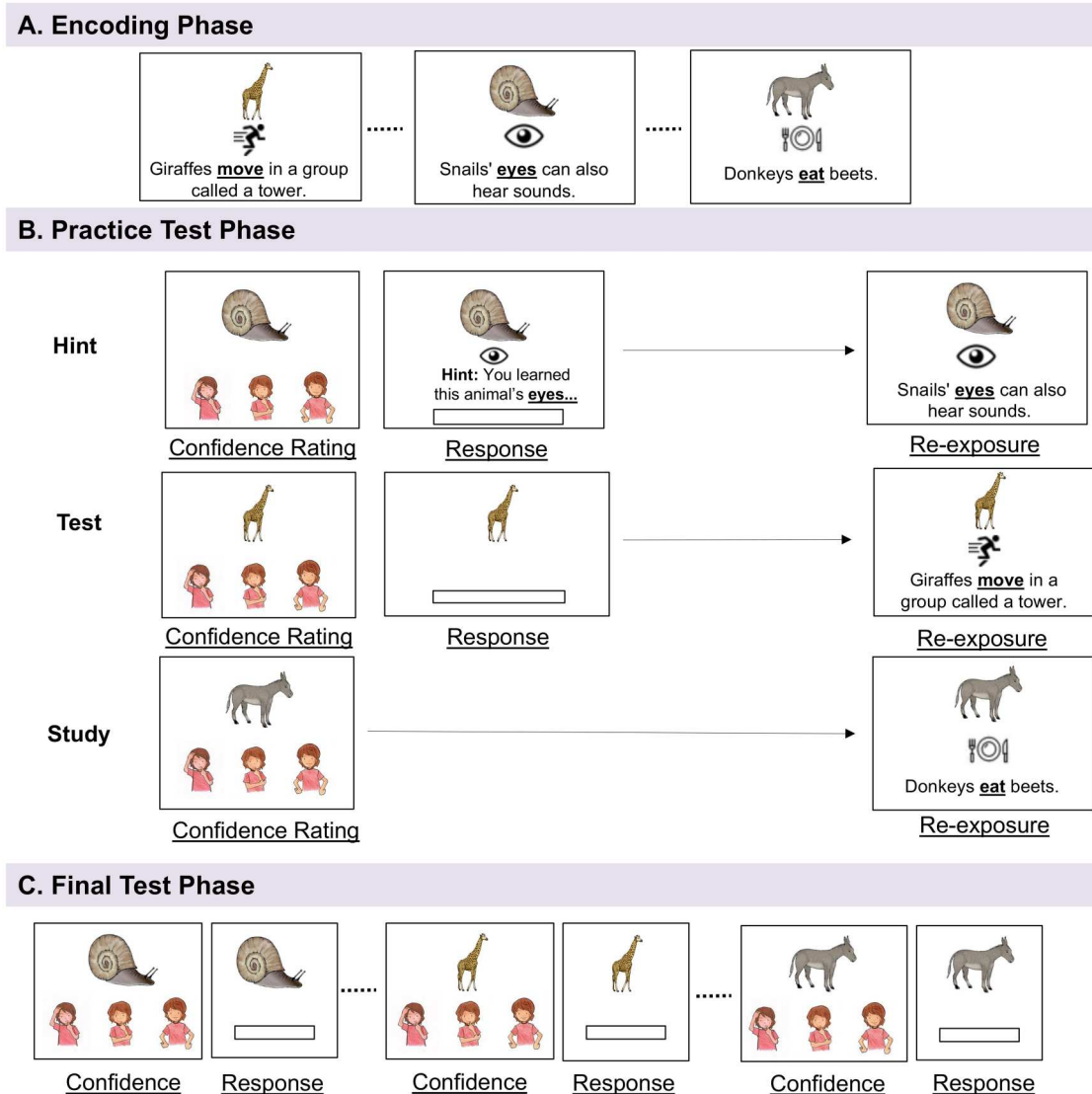
Stimuli included 42 sentences about animals (6 practice trials, 36 task trials; Shields et al., 2024). Each sentence consisted of the animal's name, fact category, and fact information, and was presented with a colour image of the animal (Rossion & Pourtois, 2004; Snodgrass & Vanderwart, 1980). Facts included information regarding animals' eyes (e.g., Snails' eyes can also hear sounds), how animals move (e.g., Giraffes move in a group called a tower), or what animals eat (e.g., Donkeys eat beets). Facts across each category (move, eyes, eat) were created for each animal and then counterbalanced across 3 orders that were randomly assigned to participants. Each participant was randomly presented with 36 unique animal facts (12 facts in each category) which were randomly assigned to condition (i.e., study, test, hint).

### Receptive Vocabulary

The NIH Toolbox Picture Vocabulary Task (<https://nihtoolbox.org/domain/cognition/>) measured receptive vocabulary skills and was used as a descriptive measure of our sample. During the Picture Vocabulary Task children were asked to select which picture corresponded with a word (presented via recorded audio) out of 4 possible options. The task automatically adapted based on children's age and response accuracy.

### Procedure

Participants completed two remote one-hour long sessions held  $\sim 24$ -hours apart via Zoom online conferencing. The programming software, Gorilla Experiment Builder (<https://gorilla.sc>) was used to present the task online (Anwyl-Irvine et al., 2020). Participants were asked to share their screen with the researcher while completing the task to ensure participants were attentive and had no technical errors. After introductions and consent, researchers turned off their audio and video feed during



**Figure 1.** Experimental task. Note. (A) During the encoding phase, animal facts were serially presented across three categories (move, eyes, and eat). (B) During the practice test phase, participants provided confidence judgements about whether they could correctly recall the animal fact and then recalled the earlier presented animal fact with hints (hint trials) or without hints (test trials) or restudied the animal fact (study trials). After providing a response, participants were presented with the full animal fact again during hint and test trials. (C) Approximately 24 hours later, participants completed a final memory test without hints.

the memory task and only communicated with children if they had questions or were off task. All instructions for the task were audio recorded to enable participants to complete the task independently. The main memory task consisted of three phases, including an encoding phase and practice test phase during Session 1 and a final test phase during Session 2.

The measure of receptive vocabulary (i.e., NIH Toolbox Picture Vocabulary) was completed at the end of Session 1 to characterise our sample. The task was administered using an iPad faced at the web camera and researchers selected answers based on the child's verbal responses. Additional measures of executive function were completed at the end of Session 2. These executive function measures were collected for exploratory analyses and not reported.

### Session 1 Memory Task

During the encoding phase, participants were presented with thirty-six 6100 ms trials containing the picture of an animal, a category cue (moves, eyes, or eats), and a corresponding animal fact along with a recorded audio message reading each fact (See Figure 1A). Participants then reported if they had already known the fact prior to the study session to maintain engagement. The yes/no option screen was displayed for the last 2500 ms of the trial, followed by a 250 ms fixation before moving on to the next trial. Participants completed six practice trials to ensure their understanding of the task.

Following encoding, participants completed a practice test (See Figure 1B). Participants were informed this phase was a practice test that would help prepare them for a final test the next day. During the practice test,



trials were randomly assigned to the study, test, or hint condition. Each trial began by presenting a picture of an animal and asking participants to provide a confidence rating indicating their level of certainty in recalling the associated animal fact using a 3-point pictorial confidence scale (i.e., not so sure, kind of sure, or really sure; Hembacher & Ghetti, 2014). Confidence was solicited prior to a response in order to examine confidence on all trials, including the study condition where no recall response was provided. In the study condition, after providing a confidence rating participants were presented with the full animal fact before moving on to the next trial. In the test condition, participants were prompted to recall the fact by typing their answer in a response box. Participants had unlimited time to type their response and would press a continue button to move on to the next trial. In the hint condition, participants were given a category cue. The category cue consisted of the category image (moves, eyes, or eats) along with a recorded audio message (e.g., you learned this animal eats). After the category cue, participants were prompted to recall the fact by typing in their answer. Once again, participants had unlimited time to type their response and would press a continue button to move onto the next trial. After submitting a response, during both the test and hint conditions participants were presented with the full correct animal fact to ensure all facts were presented the same number of times.

### Session 2 Memory Task

The second session occurred approximately 24-hours after the first session and included the final test phase (See Figure 1C). Participants were once again prompted to report how sure they were they remembered the animal fact using the 3-point confidence scale. Participants were then prompted to recall the fact by typing in their answer and would press a continue button to move onto the next trial. No hints were provided, and the correct answer was not presented after submitting a response.

### Data Processing

#### Response Coding

Free response answers were coded as either correct (1) or incorrect (0) (See Supplementary Materials). Responses were coded as correct when the fact was accurately described even if the wording was not identical to the originally presented fact (e.g., Snails move in a wave motion was coded as correct for the fact *Snails move their body in waves*). Two researchers independently coded blinded response data for accuracy (inter-rater reliability of  $k=0.90$ ,  $p<.001$ ). When discrepancies occurred, researchers held a discussion to come to an agreed-upon final accuracy code for each response. After coding, we noticed that during the practice test when hint trials occurred participants would sometimes provide their answer as a completion to the hint (e.g., hint: you learned this animal eats, typed answer: moss).

A third independent rater assessed all hint trials and coded responses that provided the accurate information in the form of a completion to the hint as correct.

## Results

### Preliminary Analyses

#### Standardised Vocabulary Scores

Age corrected standardised receptive vocabulary scores had a mean of 108.75 ( $SD=15.10$ ) and were significantly higher than the standardised population score of 100,  $t(69)=4.85$ ,  $p<.001$ . Thus, our sample exhibited higher than average receptive vocabulary.

### Main Analyses

#### Practice Test Accuracy

First, we verified that hints were beneficial and improved performance during the practice test. Using a multilevel regression analysis, we predicted average practice test accuracy using random effects of participant and fixed effects of condition (0 = Test, 1 = Hint) and age (continuous). There was a significant effect of age such that accuracy increased with age,  $b=.02$ ,  $SE=.01$ ,  $p=.038$ . There was also a significant effect of condition,  $b=.05$ ,  $SE=.02$ ,  $p=.045$ , such that accuracy was higher in the hint ( $M=.57$ ,  $SD=.19$ ) compared to test ( $M=.52$ ,  $SD=.19$ ) condition (Cohen's  $d=.23$ ). Adding the interaction between condition and age did not significantly increase model fit,  $p>.60$ . Thus, accuracy significantly improved when hints were presented during the practice test and this effect was similar across age. These results suggest that children benefitted from hints during the practice test.

#### Final Test Accuracy

Next, we examined whether performance would differ on the final test as a function of the practice test conditions (hint, test, and study). We used a multilevel regression analysis to predict average final test accuracy using random effects of participant and fixed effects of condition (dummy coded relative to the study condition) and age (continuous) (See Table 1 and Figure 2). The effect of age approached significance,  $b=.03$ ,  $SE=.01$ ,  $p=.052$  such that accuracy increased with age. There was also an effect of condition such that accuracy was significantly higher in the hint ( $M=.61$ ,  $SD=.24$ ) relative to study ( $M=.56$ ,  $SD=.22$ ) condition,  $b=.05$ ,  $SE=.02$ ,  $p=.01$ , Cohen's  $d=.29$ . However, the hint and test ( $M=.59$ ,  $SD=.21$ ) conditions did not significantly differ,  $b=.02$ ,  $SE=.02$ , Cohen's  $d=.12$ ,  $p=.32$ . This difference between hint and test conditions was also not significant when only examining responses that were successfully recalled during the practice test ( $M_{diff}=.003$ ,  $p=.91$ , Cohen's  $d=.01$ ). Finally, the difference between test and study conditions was also not significant,  $b=.03$ ,  $SE=.02$ ,  $p=.11$ , Cohen's  $d=.18$ . Adding the condition by age interaction

**Table 1.** Multilevel regression results predicting final test accuracy.

Predictors	Final test accuracy				
	Estimates	Std. Error	CI	Statistic	<i>p</i>
(Intercept)	0.28	0.14	0.00–0.57	1.99	<b>0.050</b>
Age	0.03	0.01	0.00–0.05	1.98	<b>0.052</b>
Condition [Test vs. Study]	0.03	0.02	−0.01–0.07	1.59	0.115
Condition [Hint vs. Study]	0.05	0.02	0.01–0.09	2.59	<b>0.010</b>
<b>Random effects</b>					
$\sigma^2$ 0.02; $\tau_{00subj}$ 0.03; ICC 0.66; $N_{subj}$ 77; Observations 231					

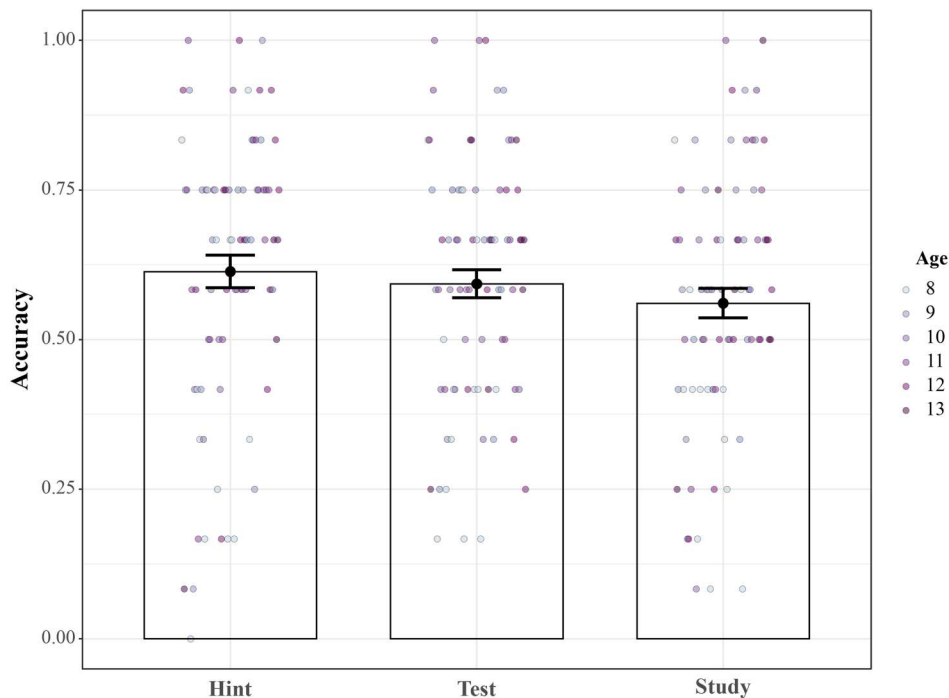
did not significantly increase model fit,  $p = .93$ . As noted in our pre-registration, we also conducted a mixed 3 X 2 ANOVA including a within-participant factor of condition (hint, test, and study) and between subject factor of age group (median split into younger and older children). The results were similar to the multilevel analyses and demonstrated significant main effects of age group  $F(1,75) = 5.09$ ,  $p = .03$ ,  $\eta_p^2 = .06$  and condition  $F(2,150) = 7.69$ ,  $p = .04$ ,  $\eta_p^2 = .04$ , and a non-significant interaction,  $p = .94$ . Exploratory analyses also demonstrated that condition effects did not interact with whether participants reported knowing the fact during the encoding phase (See Supplementary Results).

Overall, participants improved their final test accuracy when presented with hints during the practice test relative to simply restudying information. Test condition final performance was numerically in between the hint and study conditions. Performance was not significantly higher in the hint condition relative to the test condition,

suggesting that earlier presented hints do not significantly improve performance to a greater extent than testing without hints. Interestingly, we also did not find strong evidence for the typical testing effect and performance was only numerically higher in the test compared to study condition. Finally, we did not observe an age by condition interaction, suggesting the effects of retrieval practice did not vary across our developmental age range.

### Final Test Accuracy by Confidence

We examined whether final test accuracy differed as a function of practice test condition and confidence. We used a multilevel trial-wise logistic regression analysis, and predicted final test accuracy (0 = incorrect, 1 = correct) using random effects of participant and fixed effects of age (continuous), condition (dummy coded relative to study), and confidence (continuous), and the interaction between condition and confidence (See Table 2 and Figure 3). The main effect of age once again approached significance,  $OR = 1.10$ ,  $SE = .06$ ,  $p = .083$ , such that accuracy increased with age. The main effect of confidence was significant,  $OR = 2.89$ ,  $SE = .27$ ,  $p < .001$ , such that final test accuracy increased with higher levels of practice test confidence. Critically, the interaction between condition and confidence was not significant ( $ps > .26$ ), suggesting the effects of condition did not depend on practice test confidence levels. Importantly, we confirmed that when using a simpler trial-wise model with only main effects of age, condition, and confidence, the condition differences were similar to those reported previously (Hint vs. Study,  $OR = 1.35$ ,  $SE = .15$ ,  $p = .006$ ;



**Figure 2.** Final test accuracy as a function of condition. Note. Final test accuracy for hint, test, and study conditions. Points represent individual participants. Error bars represent  $\pm 1$  SE around the mean.

**Table 2.** Multilevel trial-wise logistic regression predicting final test accuracy.

Predictors	Final test accuracy				
	Estimates	Std. Error	CI	Statistic	<i>p</i>
(Intercept)	0.05	0.03	0.01–0.16	−4.99	<0.001
Condition [Test vs. Study]	1.19	0.36	0.66–2.15	0.57	0.568
Condition [Hint vs. Study]	1.36	0.40	0.76–2.42	1.04	0.299
Confidence	2.89	0.27	2.41–3.47	11.34	<0.001
Age	1.10	0.06	0.99–1.22	1.73	0.083
Condition [Test vs. Study] * Confidence	0.99	0.13	0.77–1.28	−0.06	0.953
Condition [Hint vs. Study] * Confidence	1.00	0.13	0.78–1.28	−0.02	0.982

**Random effects**  
 $\sigma^2$  3.29;  $\tau_{00subj}$  0.45; ICC 0.12;  $N_{subj}$  77; Observations 2772

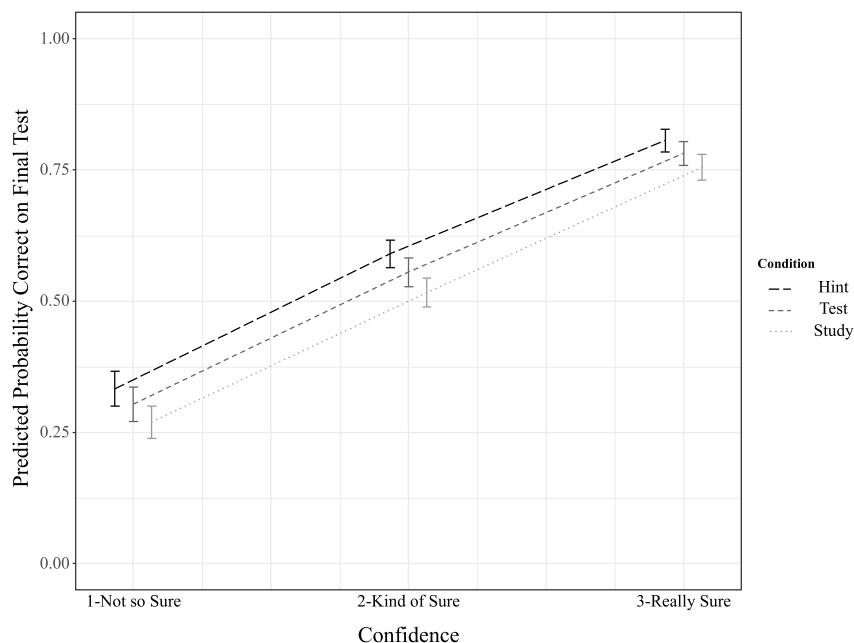
Hint vs. Test,  $OR = 1.15$ ,  $SE = .13$ ,  $p = .19$ , Test vs. Study,  $OR = 1.17$ ,  $SE = .13$ ,  $p = .15$ ). Similar results were observed when using final test confidence as opposed to practice test confidence (See Supplementary Results). Overall, these results suggest that both practice confidence and condition explained unique variance in final test recall and retrieval practice differences were similar across all levels of confidence.

## Discussion

In the current study we investigated whether the provision of hints increased the benefit of retrieval practice relative to test only and study conditions in 8- to 13-year-old children. We found partial support for our hypotheses, such

that children's long-term retention of science facts was higher when they received hints during an earlier practice test relative to restudying facts. This finding is consistent with previous research suggesting that scaffolded retrieval improves retention compared to study only conditions in children (Karpicke et al., 2014). Although the effect of hints compared to the study condition was relatively small (~5% increase in test accuracy), when considering practical applications (Funder & Ozer, 2019; McCartney & Rosenthal, 2000) this increase could result in half a letter grade change (e.g., changing from a B+ to A- or A) in common U.S. grading systems. However, hints did not result in significantly better retention than the test only condition, suggesting that the added benefit of hints on retrieval practice was limited. Furthermore, test performance was only numerically higher than study performance and this effect did not reach significance. Finally, in contrast to our predictions, we found the effects of retrieval practice were similar across age and levels of confidence, demonstrating that these effects may be robust against changes in development and subjective assessments of memory performance.

The provision of hints in our study was helpful and appropriate. Hints significantly improved practice test performance while maintaining below ceiling performance, and also improved final test accuracy compared to study conditions. However, it is possible because the hints were categorical and not uniquely related to each fact, children may have struggled to generate additional retrieval cues that could aid final test recall relative to test only conditions. In previous studies using cue-target pairs, the cue is often uniquely related to the target item (e.g., basket-bread) (Carpenter et al., 2006; Carpenter & Delosh,



**Figure 3.** Final test accuracy as a function of confidence and condition. Note. The predicted probability correct on the final test for hint, test, and study conditions across levels of confidence. Error bars represent  $\pm 1$  SE around the mean.



2006; Carrier & Pashler, 1992). Elaborate retrieval and mediator accounts of the testing effect posit that cues activate related semantic information during initial testing which can then be used as helpful additional retrieval cues on a final test (Carpenter et al., 2009; Pyc & Rawson, 2010). In our study, we intentionally chose categorical cues in order to keep the structure of facts similar, the task sufficiently difficult, and the length of the sessions appropriate to maintain interest by children while allowing for a sufficient number of trials. However, the use of categorical cues may have resulted in the generation of overlapping retrieval cues (e.g., the cue *eat* for the target fact *Kangaroos eat carrots* may generate additional cues such as kangaroos eat moss, flowers, and insects which could have been relevant food items to other animals on the list). Thus, future research should investigate whether unique semantic hints benefit retrieval practice in children to a greater extent than categorical hints. Furthermore, it is possible that after experiencing the hints during the practice test, children may have also attempted to retrieve the categorical hints on other trials without hints. Although, we observed significant differences between study and hint conditions, it is possible this effect was dampened if children engaged in this strategy. Previous research suggests young children do not spontaneously engage in categorisation strategies when learning related single-items for a free recall task (Kobasigawa, 1974). However, future research could further investigate children's learning strategies using think aloud protocols during encoding and retrieval of related information to better determine children's spontaneous strategy use and semantic association processes during learning of more complex information.

In the current study we also prompted for confidence ratings prior to a potential recall response in order to assess confidence across all conditions including when facts were restudied. However, providing confidence ratings likely encouraged participants to engage in an initial retrieval attempt and could have limited the benefits of hints and test-taking. Recent research suggests that soliciting metamemory judgments can improve retention in children (Zhao et al., 2022) and therefore one intriguing question for future research is whether confidence ratings promote retrieval practice in children and how their benefit compares to the typical testing effect. Research in adults demonstrates that delayed judgements of learning can lead to similar benefits as test-taking when learning difficult material (Akdoğan et al., 2016), suggesting that retrieval processes during testing may be similar to those during metacognitive judgments under certain contexts. Critically, our results demonstrated retrieval benefits of testing with hints across levels of confidence, suggesting that testing effects and confidence provide unique contributions to memory recall in children. Future research should directly examine whether the testing effect is minimised in children under contexts where metacognitive assessments are present vs. absent. Furthermore, since confidence and accuracy were

correlated in our study, we show that children engaged in effective metacognitive monitoring and that testing effects were stable across different mean levels of recall accuracy, replicating previous research in adults (Zhang et al., 2019). Future research could also examine the impact of hints using a younger age range (e.g., 5- to 7-year-olds) during which greater developmental changes in metacognition are typically observed (Destan et al., 2014; Selmezy & Ghetti, 2019).

In contrast to our predictions, retrieval practice was stable across development. The lack of developmental differences is consistent with previous research suggesting that the testing effect emerges as early as preschool age (Fritz et al., 2007; Kliegl et al., 2018) and is observed in elementary aged children when support such as feedback is provided (Goossens et al., 2014; McDaniel et al., 2011, 2013). In the current study, children were provided with feedback through re-exposure to the facts after providing a response. Thus, it is possible that developmental differences may emerge under more challenging contexts such as free recall in the absence of feedback. Additionally, previous research shows that children's ability to spontaneously engage in or recognise the benefit of testing improves with development (Tullis & Maddox, 2020). However, it is possible that when children are required to take tests, retrieval benefits may not heavily rely on additionally implemented strategic processes (Alamri & Higham, 2022; Fritz et al., 2007). Furthermore, our sample had higher than average receptive vocabulary and relatively limited diversity, which may have restricted our ability to detect age related effects. Finally, our sample size was not highly powered to detect small to very small effect sizes. Thus, it is possible that significant age effects or differences between study and test conditions or confidence levels would emerge with larger sample sizes.

In conclusion, our findings indicate positive benefits of testing on children's retention of science facts, with partial evidence suggesting that the provision of hints may boost the benefit of retrieval practice. Furthermore, we demonstrate that the effects of testing were similar across age and levels of confidence, suggesting that retrieval practice can benefit memory across a range of ages and contexts. Overall, our research suggests that primary school educators can benefit student learning by providing students assignments that support retrieval practice, including tests with helpful hints that include feedback for additional learning opportunities.

### Open Scholarship



This article has earned the [Center for Open Science](#) badges for Open Data, Open Materials and Preregistered. The data and materials are

openly accessible at <https://osf.io/rx32s/>, <https://app.gorilla.sc/open-materials/661279> and <https://osf.io/43ynf>.

## Acknowledgements

We would like to thank Michelle Shields for her helpful contributions to this project.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## References

- Ackerman, B. P. (1982). Retrieval variability: The inefficient use of retrieval cues by young children. *Journal of Experimental Child Psychology*, 33(3), 413–428. [https://doi.org/10.1016/0022-0965\(82\)90056-X](https://doi.org/10.1016/0022-0965(82)90056-X)
- Akdoğan, E., Izaute, M., Danion, J.-M., Vidailhet, P., & Bacon, E. (2016). Is retrieval the key? Metamemory judgment and testing as learning strategies. *Memory (Hove, England)*, 24(10), 1390–1395. <https://doi.org/10.1080/09658211.2015.1112812>
- Alamri, A., & Higham, P. A. (2022). The dark side of corrective feedback: Controlled and automatic influences of retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(5), 752–768. <https://doi.org/10.1037/xlm0001138>
- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1), 388–407. <https://doi.org/10.3758/s13428-019-01237-x>
- Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). The MIT Press.
- Bouwmeester, S., & Verkoeijen, P. P. J. L. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language*, 65(1), 32–41. <https://doi.org/10.1016/j.jml.2011.02.005>
- Brame, C. J., & Biel, R. (2015). Test-Enhanced learning: The potential for testing to promote greater learning in undergraduate science courses. *CBE—Life Sciences Education*, 14(es4), 1–12. <https://doi.org/10.1187/cbe.14-11-0208>
- Carpenter, S. K., & Delosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, 34(2), 268–276. <https://doi.org/10.3758/BF03193405>
- Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, 23(6), 760–771. <https://doi.org/10.1002/acp.1507>
- Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, 13(5), 826–830. <https://doi.org/10.3758/BF03194004>
- Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, 20(6), 633–642. <https://doi.org/10.3758/BF03202713>
- Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*, 21, 919–940. <https://doi.org/10.1080/09541440802413505>
- DeSoto, K. A., & Roediger, H. L. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science*, 25(3), 781–788. <https://doi.org/10.1177/0956797613516149>
- Destan, N., Hembacher, E., Ghetti, S., & Roebbers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5- to 7-year-old children. *Journal of Experimental Child Psychology*, 126, 213–228. <https://doi.org/10.1016/j.jecp.2014.04.001>
- Eisenkraemer, R. E., Jaeger, A., & Stein, L. M. (2013). A systematic review of the testing effect in learning. *Paidéia (Ribeirão Preto)*, 23(56), 397–406. <https://doi.org/10.1590/1982-43272356201314>
- Fandakova, Y., Selmecky, D., Leckey, S., Grimm, K. J., Wendelken, C., Bunge, S. A., & Ghetti, S. (2017). Changes in ventromedial prefrontal and insular cortex support the development of metamemory from childhood into adolescence. *Proceedings of the National Academy of Sciences of the United States of America*, 114(29), 7582–7587. <https://doi.org/10.1073/pnas.1703079114>
- Fazio, L. K., & Marsh, E. J. (2019). Retrieval-based learning in children. *Current Directions in Psychological Science*, 28(2), 111–116. <https://doi.org/10.1177/0963721418806673>
- Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *Quarterly Journal of Experimental Psychology*, 60(7), 991–1004. <https://doi.org/10.1080/17470210600823595>
- Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2), 156–168. <https://doi.org/10.1177/2515245919847202>
- Ghetti, S., & Angelini, L. (2008). The development of recollection and familiarity in childhood and adolescence: Evidence from the dual-process signal detection model. *Child Development*, 79(2), 339–358. <https://doi.org/10.1111/j.1467-8624.2007.01129.x>
- Ghetti, S., Qin, J., & Goodman, G. S. (2002). False memories in children and adults: Age, distinctiveness, and subjective experience. *Developmental Psychology*, 38(5), 705–718. <https://doi.org/10.1037/0012-1649.38.5.705>
- Goossens, N. A. M., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition*, 3(3), 177–182. <https://doi.org/10.1037/h0101800>
- Halperin, M. S. (1974). Developmental changes in the recall and recognition of categorized word lists. *Child Development*, 45(1), 144–151. <https://doi.org/10.2307/1127760>
- Hembacher, E., & Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Science*, 25(9), 1768–1776. <https://doi.org/10.1177/0956797614542273>
- Hu, X., Yang, C., & Luo, L. (2022). Retrospective confidence rating about memory performance is affected by both retrieval fluency and non-decision time. *Metacognition and Learning*, 17(2), 651–681. <https://doi.org/10.1007/s11409-022-09303-0>
- Jaeger, A., Eisenkraemer, R. E., & Stein, L. M. (2014). Test-enhanced learning in third-grade children. *Educational Psychology*, 35(4), 513–521. <https://doi.org/10.1080/01443410.2014.963030>
- Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, 19(4/5), 528–558. <https://doi.org/10.1080/09541440601056620>
- Karabenick, S. A., & Gonida, E. N. (2017). Academic help seeking as a self-regulated learning strategy: Current issues, future directions. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 421–433). Routledge.
- Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-Based learning: Positive effects of retrieval practice in elementary school children. *Frontiers in Psychology*, 7(350), <https://doi.org/10.3389/fpsyg.2016.00350>
- Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014). Retrieval-based learning: The need for guided retrieval in elementary school children. *Journal of Applied Research in Memory and Cognition*, 3(3), 198–206. <https://doi.org/10.1037/h0101802>

- Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(4), 714–719. <https://doi.org/10.1037/0278-7393.33.4.704>
- Kleitman, S., & Gibson, J. (2011). Metacognitive beliefs, self-confidence and primary learning environment of sixth grade students. *Learning and Individual Differences*, 21(6), 728–735. <https://doi.org/10.1016/j.lindif.2011.08.003>
- Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes—ScienceDirect. *Learning and Individual Differences*, 17(2), 161–173. <https://doi.org/10.1016/j.lindif.2007.03.004>
- Kliegl, O., Abel, M., & Bäuml, K.-H. T. (2018). A (preliminary) recipe for obtaining a testing effect in preschool children: Two critical ingredients. *Frontiers in Psychology*, 9(1446), <https://doi.org/10.3389/fpsyg.2018.01446>
- Kobasigawa, A. (1974). Utilization of retrieval cues by children in recall. *Child Development*, 45(1), 127–134. <https://doi.org/10.2307/1127758>
- Koriat, A., & Ackerman, R. (2010). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science*, 13(3), 441–453. <https://doi.org/10.1111/j.1467-7687.2009.00907.x>
- Lipowski, S. L., Pyc, M. A., Dunlosky, J., & Rawson, K. A. (2014). Establishing and explaining the testing effect in free recall for young children. *Developmental Psychology*, 50(4), 994–1000. <https://doi.org/10.1037/a0035202>
- Ma, X., Li, T., Duzi, K., Li, Z. Y., Ma, X., Li, Y., & Zhou, A.-B. (2020). Retrieval practice promotes pictorial learning in children aged six to seven years. *Psychological Reports*, 123(6), 2085–2100. <https://doi.org/10.1177/0033294119856553>
- McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, 71(1), 173–180. <https://doi.org/10.1111/1467-8624.00131>
- McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-Enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, 103(2), 399–414. <https://doi.org/10.1037/a0021782>
- McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, 27, 360–372. <https://doi.org/10.1002/acp.2914>
- McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, 20(1), 3–21. <https://doi.org/10.1037/xap0000004>
- Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1), 3–8. <https://doi.org/10.1037/0278-7393.31.1.3>
- 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, 437–447. <https://doi.org/10.1016/j.jml.2009.01.004>
- Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002), 335–335. <https://doi.org/10.1126/science.1191465>
- Roebbers, C. M. (2002). Confidence judgments in children's and adult's event recall and suggestibility. *Developmental Psychology*, 38(6), 1052–1067. <https://doi.org/10.1037/0012-1649.38.6.1052>
- Roebbers, C. M., Krebs, S. S., & Roderer, T. (2014). Metacognitive monitoring and control in elementary school children: Their interrelations and their role for test performance. *Learning and Individual Differences*, 29, 141–149. <https://doi.org/10.1016/j.lindif.2012.12.003>
- Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, 17(4), 382–395. <https://doi.org/10.1037/a0026252>
- 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. <https://doi.org/10.1016/j.tics.2010.09.003>
- Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science*, 17(3), 181–269. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Roediger, H. L., Wixted, J. H., & DeSoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. In L. Nadel & W. P. Sinnott-Armstrong (Eds.), *Memory and law* (pp. 84–118). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199920754.003.0004>
- Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, 33(2), 217–236. <https://doi.org/10.1068/p5117>
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432–1463. <https://doi.org/10.1037/a0037559>
- Selmecky, D., & Ghetti, S. (2019). Here is a hint! How children integrate reliable recommendations in their memory decisions. *Journal of Experimental Child Psychology*, 177, 222–239. <https://doi.org/10.1016/j.jecp.2018.08.004>
- Shields, M., Calabro, G., & Selmecky, D. (2024). Active help-seeking and metacognition interact in supporting children's retention of science facts. *Journal of Experimental Child Psychology*, 237, 105772. <https://doi.org/10.1016/j.jecp.2023.105772>
- Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2), 174–215. <https://doi.org/10.1037/0278-7393.6.2.174>
- Tullis, J. G., & Maddox, G. B. (2020). Self-reported use of retrieval practice varies across age and domain. *Metacognition and Learning*, 15, 129–154. <https://doi.org/10.1007/s11409-020-09223-x>
- Zhang, M., Chen, X., & Liu, X. L. (2019). Confidence in accuracy moderates the benefits of retrieval practice. *Memory (Hove, England)*, 27(4), 548–554. <https://doi.org/10.1080/09658211.2018.1529796>
- Zhao, W., Li, B., Shanks, D. R., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., Yin, Y., Luo, L., & Yang, C. (2022). When judging what you know changes what you really know: Soliciting metamemory judgments reactively enhances children's learning. *Child Development*, 93(2), 405–417. <https://doi.org/10.1111/cdev.13689>