Comparing generating predictions with retrieval practice as learning strategies for primary school children

Paulo F. Carvalho[1] and Karrie Godwin[2]

[1]Human-Computer Interaction Institute, Carnegie Mellon University

[2]Department of Psychology and the Sherman Center for Early Learning in Urban Communities, University of Maryland, Baltimore County

Author Note

Paulo F. Carvalho https://orcid.org/0000-0002-0449-3733

Karrie Godwin https://orcid.org/0000-0003-0127-986X

Both authors contributed equally to this work.

Correspondence concerning this article should be addressed to Paulo F. Carvalho at pcarvalh@cs.cmu.edu and Karrie Godwin at kgodwin@umbc.edu.

Paulo F. Carvalho and Karrie E. Godwin conceptualized and designed the study. Karrie E. Godwin supervised data collection. Paulo F. Carvalho wrote the analysis code and model simulations. Paulo F. Carvalho and Karrie E. Godwin analyzed the data and wrote the manuscript.

All data and analyses code are available at OSF (https://osf.io/qb8ac/).

Abstract

This eye tracking study examines the learning benefits of two common active learning approaches — generating predictions and retrieval — for young children. Both generating predictions and retrieval practice are active learning approaches that involve generating responses and then being provided with the correct information or retrieving previously provided correct information. Participants included 90 children (mean age: 7 years; Female = 46, Male = 42). Parents reported children's race and ethnicity as follows: 2% Asian/Pacific Islander, 5% African American, 74% Caucasian, 3% other, and 6% identified as two or more categories; demographics largely reflective of the county where the data was collected, but nevertheless the generalizability of these findings to more diverse populations may be limited. In this study, young children learned facts about insects (e.g., "insects are hard on the outside") while we measured their attention to the lesson using eye tracking technology. Then their knowledge was assessed on an immediate test. All children were presented with the same materials but the presentation order was modified based on condition assignment. In the generating predictions condition, children saw examples of animals and were asked if an animal was an insect or they saw animals and were asked to identify which one was the insect, followed by the correct response. In the retrieve condition, the presentation order was reversed such that children first saw the correct response and then were asked if the animal was an insect or which of two examples was an insect. Results suggest that although retrieval practice results in overall better learning outcomes, generating predictions increased children's attention to the materials (d = 1.92), and among children who were able to maintain attention, learning outcomes were equal among the two conditions.

Keywords: generation; prediction; retrieval practice; attention; memory

Public Significance Statement

This work examined whether generating predictions or retrieval practice yields better learning outcomes for young children learning science facts. We hypothesized the key difference between the two approaches is the attentional and encoding demands. We found children attended more to the materials when learning by generating predictions, but performed better at posttest when learning through retrieval. Attention varied less for children learning by generating predictions, and, in that condition only, those who attended to the instructional materials most, benefited most. The findings suggest generating predictions can be as effective as retrieval practice of facts for primary school children. However, our model simulations and prior evidence indicate students' attentional control abilities, which increase with age, may be key to this benefit. This work highlights the importance of considering student-level factors when deciding the best learning strategies to implement.

Word count 137/120-150

Students learn better when they actively engage with instructional materials. The benefits of active learning for educational outcomes and attainment are well documented (Freeman et al., 2007, 2014; Haak et al., 2011; Yannier et al., 2021). Children are naturally curious and physically active, suggesting that creating active learning instruction should be easy. Importantly, not all active learning is created equal (Chi, 2009, 2021). For example, evidence has shown that the best learning outcomes are not achieved by allowing for active free exploration or discovery (Klahr & Nigam, 2004; Matlen & Klahr, 2013), but through active engagement with the materials and with the appropriate support (Chi, 2021; Klahr & Nigam, 2004). Thus, an important question is what type of support is more effective to help learners, and young learners, in particular, learn from active learning approaches? What types of active learning activities lead to the best learning outcomes in young children and why? Specifically, we focus on two common ways to learn and teach: generating predictions and retrieving previously presented information. We focus on these two approaches because they both constitute active, constructive learning (Chi, 2009), which should have a positive impact on learning, albeit not as much as interactive learning, but have important differences in terms of attentional demands and memory consequences.

Substantial research has demonstrated that learners remember information better when they generate information compared to reading (e.g., Slamecka & Graf, 1978). This general finding has been described in both young and older adults (Bertsch et al., 2007) and elementary school children (McFarland et al., 1983; Varga & Bauer, 2013). Importantly, the benefits of generation go beyond memorization tasks. For example, Rittle-Johnson and Kmicikewycz (2008) demonstrated that 3rd graders with low prior knowledge were more likely to correctly answer novel questions at posttest if training involved generating responses compared to just reading them. One possible explanation for the generation effect is that generation encourages learners to use strategies during study that are necessary to acquire new information or procedures (McNamara & Healy, 2000). An additional benefit is how the cognitive processes

used during training and test match — if during training learners are trying to generate answers as they are during testing, transfer will be easier (Winstanley et al., 1996).

Another approach, retrieval practice, has also been shown to improve learning and memory with adolescents and adults (Pashler et al., 2007; Roediger & Karpicke, 2006a). For example, Roediger and Karpicke (2006a) compared the effect of retrieval practice (i.e., repeated testing) to restudying on undergraduate students' memory of a text passage. Participants were asked to read a passage and then given a free recall test at various time intervals (5-min delay, 2-day delay, 1-week delay). Although restudying was found to initially benefit participants' memory for the passage more than retrieval, retrieval practice elicited greater recall performance after a delay. This pattern of results was consistent across different delay intervals (2 days and 1 week) suggesting that retrieval practice has benefits for long-term memory. Additionally, the benefits of retrieval practice have been found with a variety of stimuli (e.g., word lists, pictures) and memory assessments (e.g., Karpicke et al., 2016; Pashler et al., 2007; Roediger & Butler, 2011; Roediger & Karpicke, 2006b; Sutterer & Awh, 2016; for review see Rowland, 2014).

Although less is known about the potential benefits of retrieval practice for children (see e.g., Fazio & Marsh, 2019; Marsh et al., 2012), there is growing evidence suggesting this learning strategy may also have benefits for young children (e.g., Karpicke et al., 2016; Ritchie et al., 2013). For example, Karpicke et al. (2016) investigated elementary school students' memory for word lists after engaging in retrieval practice or restudy (manipulated within-participants). The authors found benefits of retrieval practice over restudy for both free recall and recognition performance.

Prior research contrasting the effects of generation and retrieval practice for memory and learning has suggested that the key difference between the two is connected to the act of retrieving information (Karpicke & Zaromb, 2010). In a series of experiments, Karpicke and Zaromb (2010) showed that when participants were asked to use cues to recall information

presented beforehand they remembered the information they retrieved longer than when the same cues were presented but participants were asked to generate the first word that came to mind. Their analyses also showed that retrieval enhanced the retention of item-specific information to a greater extent than did generation. One possible reason why retrieval enhances memory compared to generation is that the act of effortfully retrieving information from memory delays forgetting (e.g., Pyc & Rawson, 2009). For example, Pavlik and Anderson (2008) used an extension of the ACT-R model (Adaptive Control of Thought—Rational model; Anderson et al., 1997; described in more detail below) to demonstrate that retrieval was associated with lower initial activation of information but slower decay of that activation, resulting in a higher likelihood of recall at a delay compared to re-reading. Taken together, these results suggest that as long as retrieval practice leads children to retrieve information from memory, it will have a positive impact on later retrieval, even if children do not attend to the materials as much (a lower initial activation).

However, other researchers (Brod, 2020, 2021; Brod et al., 2018; Brod & Breitwieser, 2019), have suggested that generation can be powerful learning approach, especially for young children. Specifically, they have suggested that generating predictions is likely improve learning across different age ranges and settings compared to other generative learning approaches (Brod, 2020). Consistent with that argument, Brod et al. (2018) showed that asking undergraduate students to generate predictions as opposed to responding to questions after being told the critical information, resulted in better learning of geography knowledge (see also, Carvalho et al. 2018). Brod et al. (2018) compared looking time when the prediction took place before or after a new piece of information was presented. They found that making a prediction, particularly if the subsequent outcome was unexpected or novel, increased pupillary diameter, suggesting greater attention. Similar results have been found with school-age children (Brod & Breitwieser, 2019; Marsh et al., 2012). One hypothesis for the power of generating predictions, particularly in young children, is that predicting increases curiosity and surprise, leading to better

learning (Brod et al., 2018), potentially by increasing attention and encoding of the relevant information compared to retrieval.

Effective attentional control involves both selection, in which only a subset of information is processed, as well as the ability to maintain a state of focused attention over time. The ability to selectively sustain attention to an object or event can be driven by exogenous factors (e.g., novelty or stimulus saliency) or controlled internally by endogenous factors (e.g., cognitive processes such as goals) (for a review see, Bornstein, 1990; Ruff & Rothbart, 2001). Exogenously regulated attention is largely an automatic process evident early in development; in contrast, endogenously regulated attention is thought to emerge later in development (e.g., Diamond, 2006; Luna, 2009; Ruff & Rothbart, 2001). In formal instructional contexts, the ability to endogenously regulate attention may be crucial for learning (Erickson et al., 2015; Fisher et al., 2014) as children need to inhibit a variety of distractions commonly found in classroom environments that compete for attention including distractions from peers, the surrounding visual environment, as well as the propensity to engage in self-distractions (Godwin et al., 2016, 2022).

In sum, both retrieval practice and generating predictions have been found to improve learning outcomes across multiple age ranges. Whereas retrieval practice slows forgetting generating predictions is thought to increase attention and encoding of the information provided following the generation of the prediction. However, this benefit might require attention regulation abilities that younger children (6 to 9 year olds) lack (Brod et al., 2020; for a discussion of the development of attention see Ruff & Rothbart, 2001). In fact, it is possible that the full benefit of generating predictions is only possible with more advanced attentional control (see e.g., Brod et al., 2020). As noted above the ability to maintain attention to the lesson is thought to be important for learning. For example, when generating a prediction children need to selectively attend to the essential elements of the instruction to understand which elements matter and which do not and how the elements fit together while inhibiting elements that do not

fit within the same organizational/conceptual structure. In addition, children must maintain attention to the corrective feedback in order to encode the information, integrate it into their conceptual model and/or correct a misconception. Thus, although young children might be able to generate predictions, and doing so might boost attention and encoding, the full learning benefits from generating predictions might only be attained as children's capabilities to regulate their attention endogenously become more advanced. We do not have an a-priori hypothesis as to whether attending to specific elements, feedback, or some combination of the two is most critical for learning. Thus, in the present study, we elect to use a global measure of attention, proportion of time attending to the lesson. See Method section for additional details.

Model Simulation

Putting aside the possibility that at least some young children might lack the attentional control required to benefit from generating predictions, it is still an open question whether generating predictions can increase learners attention to the materials and, if yes, whether increasing attention can by itself lead to better learning compared to slowing forgetting. We predict that if generating predictions increases attention and encoding and final testing occurs shortly after training, then generating predictions can result in better learning even if it does not delay forgetting. To further demonstrate this hypothesis, we simulated learning through retrieval and generating predictions using the Adaptive Control of Thought—Rational model (ACT-R, Anderson et al., 1997). We simulated a simplified version of the learning task used in the present experiment (described below): a fact is presented at time 0 (e.g., "Insects are hard on the outside"), either through generating a prediction or retrieval practice. We simulated only the memory component, so we simulated how activation created by that learning event decays up until the final test, which occurs eight trials later. ACT-R activation is related to repetition (which we are not manipulating) and attentional changes (e.g., Lovett, Daily, & Reder, 2000). In ACT-R, each time an item is practiced, the activation of the item, B(t), receives a boost in strength that decays (d) away as a power function of time (t_k):

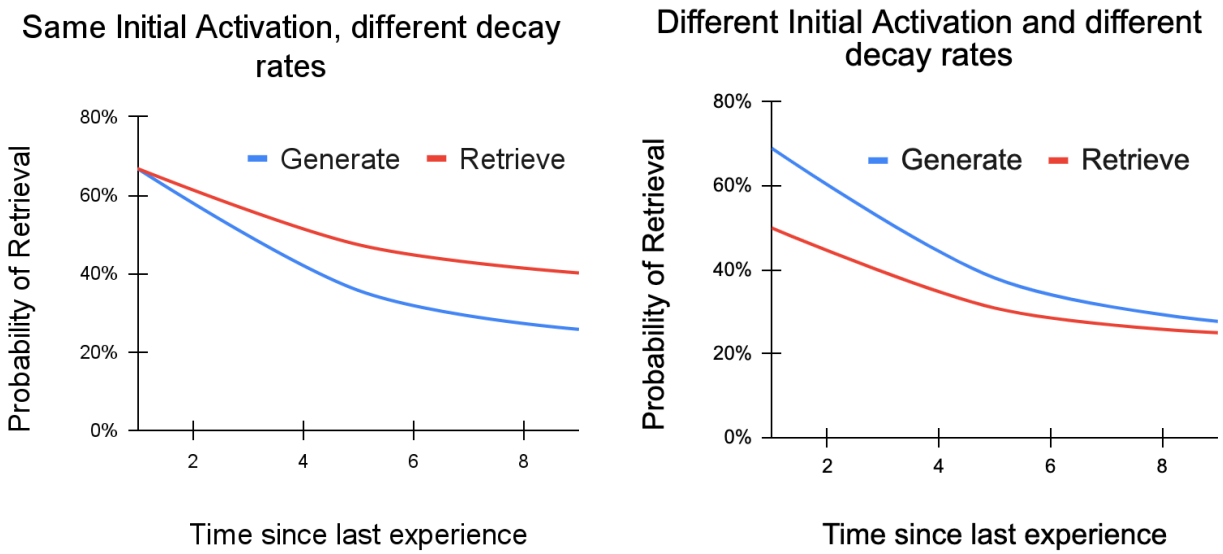
$$B(t) = ln \sum_{k=1} t_k^{-d} + B \quad (1)$$

We implemented the predictions described above in this model by varying the decay rate (-d) and initial activation (B) described in formula (1). Retrieval practice is predicted to delay forgetting, thus resulting in a lower decay rate than generating predictions. Conversely, generating predictions is predicted to increase activation, thus resulting in higher initial activation. We simulated the results of the experiment by running the model and selecting parameters that vary only the decay rate (d = -0.5 for retrieval and d = -0.8 for generating predictions, lower for retrieval because retrieval has been proposed to delay forgetting) and by manipulating both decay rate (d = -0.5 for retrieval and d = -0.8 for generating predictions) and initial activation (B = 0 for retrieval and B = 0.8 for generating predictions, higher for generating predictions because it has been suggested that generating predictions increases attention and encoding, which would result in higher initial activation). Activation is on a log-odds scale so we converted it into the probability of recall.

As can be seen from the simulation results below, if we assume no differences in initial activation between retrieval and generating predictions, the probability of recalling a memory at the end of training is higher for retrieval than for generating predictions (left panel in Figure 1). However, if we consider that generating predictions boosts initial encoding and thus results in higher initial activation, then generating predictions results in a higher probability of recall at the end of training (right panel of Figure 1).

Figure 1

Results of memory activation simulations using ACT-R



Note. Model simulations hypothesize *only* differences in decay rate between retrieval and generating predictions (d = -0.5 for retrieval and d = -0.8 for generation, B = 0.7 for both), (left panel) or differences in both decay rate (d = -0.5 for retrieval and d = -0.8 generating predictions) and initial activation (B = 0 for retrieval and B = 0.8 for generating predictions) between retrieval and generating predictions (right panel).

Thus, previous research and the model simulations above make a simple prediction that we will test in the current study: generating predictions will improve learning compared to retrieval in as much as the initial encoding is increased. This requires children to control and sustain their attention while generating predictions and receiving subsequent feedback. Otherwise, retrieval practice is likely to result in the best learning. One way to measure initial encoding activation and attention is by having measures of attention. As detailed below, in this study we included both online measures (eye tracking to investigate how much children looked at the screen) and performance-based (TrackIt) measures of attention.

Current Study

In this study we investigate whether generating predictions can increase attention during learning when compared to practice testing as suggested, but not directly tested, by previous research (Brod et al., 2020). Moreover we investigate the mechanism by which generating predictions might improve learning by probing whether, for children for whom generating predictions increases attention, it also results in equivalent or better learning outcomes than retrieval practice, as suggested by the literature and our ACT-R simulations above.

Participants completed a short science lesson about insects. During the lesson, participants were presented with pairs of animals as well as negative examples and asked to indicate which animal was an insect (or *if* the animal was an insect). Participants were asked to make their predictions *prior* to learning the diagnostic feature (Generating Predictions condition) or immediately *after* learning the diagnostic feature (Retrieve condition). This subtle manipulation equates the lesson content that children receive in each condition by simply modifying the presentation order without changing anything else about the materials across conditions. Note that the number of learning opportunities is the same across conditions; however, due to inherent differences in the expected accuracy with which participants are expected to generate predictions vs. retrieve a *correct* answer, the conditions differ in the number of expected opportunities children have to encode the *correct* information on each trial. The Retrieve condition provides children with two opportunities to encode the *correct* information per trial: once when the diagnostic information is presented in the lesson and once when children are asked to retrieve the information (assuming they retrieve the correct answer). Although the Generate condition also includes both of these opportunities, because the first step (i.e., prediction) involves guessing, children are less likely to produce accurate predictions and thus more likely to have only one opportunity to encode the correct information per trial, arguably making generating predictions more difficult than retrieval. This manipulation is somewhat similar to previous research comparing the relative benefits of practice/exploration

after or before direct instruction (e.g., Fyfe et al., 2014). In that literature, too, there seem to be interactions and moderators (e.g., Ashman et al., 2020).

Maintaining focused attention is challenging for children and lapses in attention resulting in off-task behavior are common, particularly during instruction (e.g., Godwin et al., 2016; Karweit, 1983; Sterling-Turner et al., 2001). However, in the Retrieve condition, even if children exhibit lapses in attention and in turn encode partial information it could be enough activation for children to retrieve the correct response. High levels of successful retrieval are thought to be a prerequisite to obtaining learning benefits from retrieval practice (Karpicke et al., 2014). The present experimental design increases the likelihood of successful retrieval since children are providing their response *immediately* after being given the diagnostic information. In contrast, with the Generating Predictions condition, children are less likely to provide a correct initial response as children are making a prediction before learning the diagnostic information.

In line with the ACT-R predictions described above, retrieval should yield activation that decays more slowly resulting in higher performance at posttest; however, if children in the generate condition are attending at higher levels (i.e., looking at the screen more) then we would expect high levels of initial activation that would yield strong learning performance despite faster decay. Thus, we might expect generally better posttest performance in the Retrieve condition compared to the Generating Predictions condition, but a reversal in the effect if a certain threshold of attention is surpassed. The precise threshold of attention required to elicit a reversal effect is an open question that we return to in the Discussion. To investigate this question, in addition to including a pre and posttest measure, we also included an online measure of attention to the task (eyetracking) and a performance-based measure of selective sustained attention (Track-It, Fisher et al., 2013). Unfortunately, the analysis and interpretation of the Track-It data is limited as only a subset of the children had Track-it data that could be analyzed (See Method section for details on the missing data and exclusion criteria). The analyses are presented in the Appendix, for completeness.

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions, and all measures in the study, and we follow JARS (Journal Article Reporting Standards; Kazak, 2018). All data and analyses code are available at OSF (https://osf.io/qb8ac/). Data were analyzed using R, Version 4.1.1 (R Core Team, 2017) and the following packages: ggplot2, version 3.3.5 (Wickham, 2016), and psych version 2.1.9 (Revelle, 2021), effsize version 0.8.1 (Torchiano, 2020), emmeans (Lenth, 2023), sjPlot (Lüdecke, 2023), and plotrix version 3.8.2 (Lemon, 2006). This study's design and its analyses were not pre-registered.

Participants

We chose an initial sample size target of n = 100 based on sample sizes of previous similar work. However, due to the Covid-19 pandemic, we could not collect data beyond 90 participants. We conducted a sensitivity analysis that suggests we were well-powered to detect the effect of interest (see Appendix for details). A total of 90 (45 in each condition) children participated in the present study. Two children were excluded from analyses due to their performance on the pretest catch trials (described below). Thus, the final sample included 88 (44 in each condition) children (M_{age} = 6.72 years, Range = [6:9], *SD* = .68 years, Female = 46, Male = 42) including 46 kindergarten and 42 primary school children (37 first graders and 5 second-graders). An additional 17 children (7 from the Generating Predictions condition and 10 from the Retrieve condition) were excluded from analyses of the eye tracking data either because they did not have eye tracking data (n = 8) or due to issues with the quality of the eye tracking data (n = 9; described below). Thus, for analyses including eyetracking measures the sample included 71 children (37, or 82% of the sample, in the Generating Predictions condition and 34, or 76% of the sample, in the Retrieve condition).

Participants were from a medium-sized city in the Midwest. Parents reported children's race and ethnicity information as follows: 2% Asian/Pacific Islander, 5% African American, 74%

Caucasian, 3% selected the category other, and 6% identified as two or more categories, while 10% elected not to disclose this information. These demographics largely reflect the demographics of the county where the data was collected according to the US Census Bureau. Participants were tested individually by trained research assistants in a space adjacent to their classroom or in a research laboratory.

Materials and Procedure

Participants listened to a computer-based lesson about insects. Participants were presented with animals and animal pairs (e.g., ant | pillbug) and asked to identify which animal was the insect. Participants were also shown negative examples (e.g. spider) and asked if the animal was an insect. Participants were randomly assigned to conditions: Generating Predictions or Retrieve. Based on condition assignment, children either made predictions *prior* to being told the diagnostic feature (Generating Predictions condition) or *after* being told the diagnostic feature (Retrieve condition). Eye-tracking technology was utilized to measure participants' attention to the lesson. A pretest and posttest were used to assess learning gains. An independent performance-based measure of attention, Track-It (Fisher et al., 2013), was also administered. All procedures were approved by the relevant institutional review board.

Science Lesson

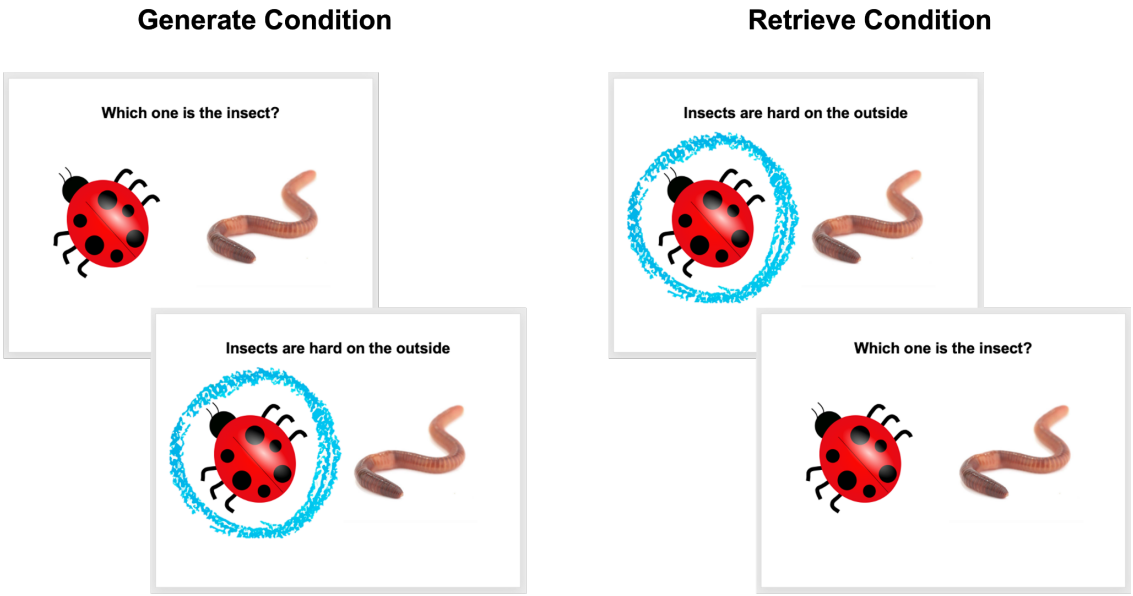
The science lesson was a short computer-based lesson that introduced children to insects and the diagnostic features of insects. The lesson content was based partly on the children's book, Bugs are Insects: Let's Read and Find Out Science, by Anne Rockwell. Both conditions started with content that motivated learning about whether bugs are insects or not. After the motivation pages, children were presented with classification pages in which they were asked to select the insect (n = 5) or determine whether an animal was an insect (n = 4). Regardless of condition, for the initial classification page, children were shown an array of animals and asked to identify which ones were insects. For this initial item, all of the presented images were insects.

No feedback was provided on this initial classification page. It served merely as a motivational device and children were subsequently told they would find out as they learned more about insects. For the 8 critical classification pages children were presented with animal pairs (n=4) or an example animal (negative example; n=4) and asked to make predictions either prior to (Generating Predictions condition) or after (Retrieve condition) being told the diagnostic features of insects (see Figure 2). For example, children were presented with a ladybug and a worm and asked to identify the insect (*which one is the insect?*). In the Generating Predictions condition, children made their prediction prior to being told a relevant diagnostic feature (e.g., *insects are hard on the outside*) and before receiving visual feedback (i.e., the correct answer, the insect, encircled); see Figure 2. In contrast, in the Retrieve condition, the presentation order was reversed and children were first told the diagnostic feature (e.g., *insects are hard on the outside*) and given visual feedback (i.e., they saw the correct answer, the insect, encircled). Then, children were asked to identify the insect. For the negative example items, children in the Generating Predictions condition were shown a negative example (e.g., Slug) and asked to indicate if it was an insect. Children made a prediction prior to receiving the diagnostic information (e.g., *slugs are <u>not</u> hard on the outside*). After receiving the diagnostic information, they were then given feedback (e.g., children were told: Slugs are not insects and the slug was crossed out with a large “X”). In the Retrieve condition, children were first given the diagnostic information (e.g., *slugs are <u>not</u> hard on the outside)* and then asked if the animal is an insect. The children were then given feedback (e.g., the children were told: Slugs are not insects and the slug was crossed out with a large “X”). The pairings and selected animals served to highlight the diagnostic features of insects (e.g., 3 body parts, 3 pairs of legs). Children’s response to each classification page was largely self-paced. However, if children did not respond, the experimenter would administer a prompt inviting children to make their best guess. The presentation order of the lesson questions was fixed and identical across conditions. After the initial classification item, the presentation order of the remaining 8 lesson questions was alternated between animal pairs and negative examples. The

lesson ended with content that again motivated learning about whether a bug is an insect or not. All lesson content was read aloud to the children by the experimenter.

Figure 2

Schematic depiction of the science lesson content



Note. Children learn about a diagnostic feature of insects either before or after participants make a prediction as to which animal is an insect (Retrieve and Generating PredictionsGenerate conditions respectively). Images obtained from Google Images. All lesson content was read aloud by the experimenter.

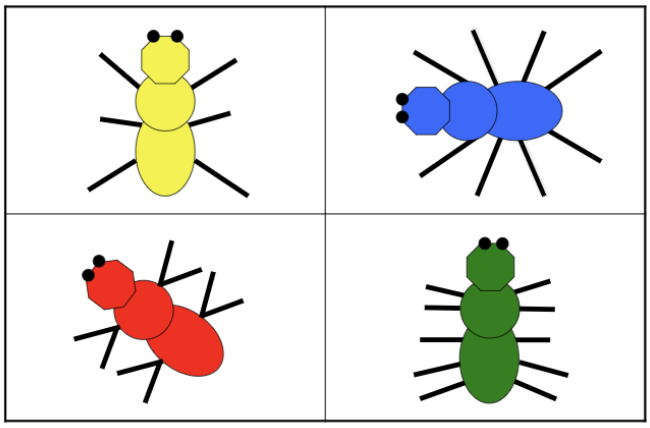
Learning Assessments

Children completed a pretest to assess their prior knowledge for the topic and ensure the content was generally novel to children of this age group. A posttest was administered to assess what children learned from the lesson and to calculate any learning gains. The learning assessments included 12 multiple choice test items (7 critical trials and 5 catch trials to help ensure participants understood the task structure). The critical trials focused on the features of

insects (e.g., point to the animal with 3 body parts), identifying a negative example (Which animal is not an insect?; Target = spider; Lures = butterfly, mosquito, bee) or identifying insects. For example, children were shown abstract depictions of bugs and asked to identify which one was the insect based on the diagnostic features of insects covered in the lesson (e.g.,3 pairs of legs; See Figure 3 for an example assessment item). The questions were read aloud to the children. The children selected their answer (verbally or by pointing) to one of four pictorial response options. Participants who had 50% or less accuracy on the catch trials were excluded from the analyses (n=2, 1 participant from each condition). Two presentation orders were created. For Order 1 the item order was randomized and for Order 2 the item order was reversed. The posttest was identical to the pretest; however, the item order was reversed. Accuracy by question type and a total score was calculated. The learning assessments were largely self-paced; however, if children did not respond the experimenter would prompt children to make their best guess.

Figure 3

Schematic depiction of a learning assessment question



Note. For this assessment item, children were asked to “point to the picture of the insect.” The correct answer is the yellow bug (top left) as it depicts an insect with 3 pairs of legs, a diagnostic feature of insects discussed in the lesson.

Attention: Eye tracking

Global attention to the lesson was measured online (during the task) using a Tobii X3-120 eye tracker and the Tobii Studio software. The eye tracker is a mobile unit and does not require a head rest, instead, the eye tracker connects directly to the bottom of the computer screen. Each participant completed a 5-point calibration at medium speed in which participants were asked to track a red ball as it moved across a gray screen. As noted above, 17 children were ultimately excluded from analyses of the eye tracking data either because they did not have eye tracking data (n = 8) or due to issues with the quality of the eye tracking data (n = 9 participants with valid gaze samples at or below 50%). In Tobii Studio we created lesson slide AOIs (AOIs encompassing the entire slide) in order to record participants’ fixation duration to the lesson. For each participant, the proportion of time fixating to the lesson was then calculated (Total fixation to lesson slide AOIs/Total lesson time). We used the proportion of time fixating on the lesson as our measure of attention because it reflects learners' global attention to the task. Our hypothesis, described above, does not make differential predictions depending on what lesson content was being fixated or differential fixation times for different information, although this could be a fruitful direction for future research. Instead, our main hypothesis is that, the benefits of having students generate predictions instead of retrieving a response will depend on initial activation, which we are operationalizing as global looking time during the lesson.

Attention: Track-It

An independent, performance-based, measure of attention, Track-It (Fisher et al., 2013), was administered. In this computer task, children are asked to visually track a target shape (e.g., an

orange triangle) as it moves around the computer screen amongst a group of heterogeneous distractors. At the end of each trial, the shapes disappear and the child's task is to report the last location the target visited prior to disappearing. After each trial, a memory check is administered in which children are asked to report which shape they had been tracking during the preceding trial. The memory check is designed to help dissociate the contribution of working memory and attention on task performance. Accordingly, for each child, a corrected accuracy score was calculated in which children's tracking accuracy was calculated *only* for those trials in which they had successfully identified the target shape on the memory check.

Unfortunately, the analysis and interpretation of the Track-It data is limited as only a subset of the children (n= 47) had Track-it data that could be analyzed, due to several reasons including missing data (n=12), task completion issues (n=10), administration error in task parameters (n=1), and/or a delay between administration of the primary task and Track-It as a subset of participants (n=21) were part of a larger parent study. Therefore, we will not include the analyses in the Results sections but include them in Appendix, for completeness.

Results

Pretest performance

Children in the Generating Predictions group correctly answered on average 8 out of 12 questions at pretest (M = 0.64, 95% CI = [0.62, 0.66]). Similarly, children in the Retrieve group also correctly answered on average 8 out of 12 questions at pretest (M = 0.64, 95% CI = [0.62, 0.66]). To investigate these differences, we used a two-sample t-test for null hypothesis significance testing (NHST), and an equivalence test, via two one-sided tests (TOST) with alpha-level of .05. These tested the null hypotheses that true mean difference is equal to 0 (NHST), and true mean difference is more extreme than -0.1667 and 0.1667, equivalent to one question difference in the pretest, on average, between the two conditions (TOST).

The NHST was not statistically significant, *t* (86) = -0.05, *p* = .961, *d* = -0.010, *95% CI* = [-0.43,0.41], but statistically equivalent, t(86) = -12, p <.01, mean difference = -0.0013, 90% C.I.[-0.048, 0.045]; Hedges's g(av) = -0.010 90% C.I.[-0.36, 0.34]. Therefore, we can conclude that there are no differences between the groups in how they perform at pretest.

Across both groups, children were also highly and equally accurate on the catch trials at pretest for both the Generating Predictions, *M* = 0.96, 95% *CI* = [0.95, 0.98] and Retrieve groups, *M*= 0.96, 95% *CI* = [0.95, 0.98]. To test this equivalence we used a true mean difference between -0.2 and 0.2 (equivalent to one question accuracy difference between the groups). We found that performance was statistically equivalent between the two groups, t(86) = -12, p <.01, mean difference = 0, 90% C.I.[-0.028, 0.028]; Hedges's g(av) = 0 90% C.I.[-0.359, 0.359]. Thus, both groups performed equally well on the pretest and understood the task. Overall, these results suggest that there were no *a priori* differences between the two groups on how much they knew about the target content or understood the task. In subsequent analyses, we use only the posttest scores to compare the groups' performance.

Lesson performance

Recall that the number of learning opportunities was the same across conditions; however, the conditions differ in the number of expected opportunities children have to encode the *correct* information on each trial due to inherent differences in the expected accuracy with which participants are expected to generate predictions vs. retrieve a *correct* answer. Consistent with this prediction, children were more accurate on the lesson classification pages in the Retrieve group (*M* = 0.81, *95% CI* = [0.79, 0.82]) than in the Generating Predictions group (*M* = 0.67, *95% CI* = [0.64, 0.70]), *t* (86) = -3.93, *p* = .0002, *d* = 0.84, *95% CI* = [0.40,1.28], suggesting that as anticipated generating a response is more error-prone than immediate retrieval. Although not a main goal of the research, we repeated the analyses including age as a main effect. Performance in the lessons was overall better for children in primary school (*M* = 0.791, *SD* = 0.169) compared to kindergarten children (*M* = 0.696, *SD* = 0.165), *beta* = 0.79, *t*

(84) = 2.99, p = .004, but there was no interaction between age and condition, that is, children's performance during the lesson was better in the Retrieve than in the Generate condition regardless of their age. This pattern of results was similar when we included age as a factor in all the other analyses reported: there was no significant interaction between age and condition on pre or posttest.

Looking time during the lesson

As a measure of initial activation/encoding during the lesson, we calculated for each child the proportion of time fixating on the lesson materials by dividing the fixation time by the total time in the lesson. Using this measure, we found that children in the Generating Predictions group spent on average a larger proportion of time fixating the lesson materials (M = 0.60, *95% CI* = [0.58, 0.62]) than children in the Retrieve group (M = 0.54, *95% CI* = [0.50, 0.57]), t (69) = 8.08, p <.0001, d = 1.92, *95% CI* = [1.35,2.49]. Thus, generating a prediction before instruction led to greater attention to the materials than retrieving a response after instruction.

Moreover, looking time was more varied in the Retrieve group (IQR = 0.223) than in the Generate Predictions group (IQR = 0.110) and the proportion of time fixating during the lesson was positively correlated with posttest performance for the Generating Predictions group, r^2 = .43, p = .0007, *95% CI* = [.13, .66], but not for the Retrieve group, r^2 = .08, p = .637, *95% CI* = [-.26, .41]. We investigate this relation further when analyzing the group differences in posttest performance.

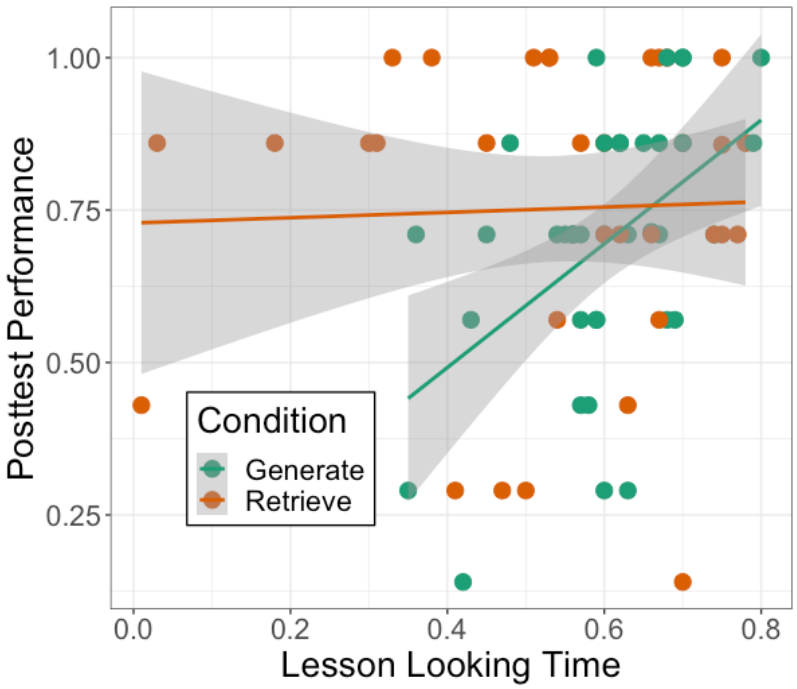
Posttest performance

In line with our hypothesis and predictions that the benefit of generating predictions vs. retrieval for learning will depend on attention control and initial encoding, we analyzed the independent and combined effects that lesson format (Generating Predictions vs. Retrieve) and encoding during the lesson have on posttest performance using a regression model predicting posttest performance in the critical trials with condition (dummy coded, Generate = 0, Retrieve =

1) and proportion of time fixating the study materials (as a measure of initial encoding and attention), as well as their interaction, as predictors.

Figure 4

Relation between the proportion of time fixating during the lesson and posttest performance for each condition



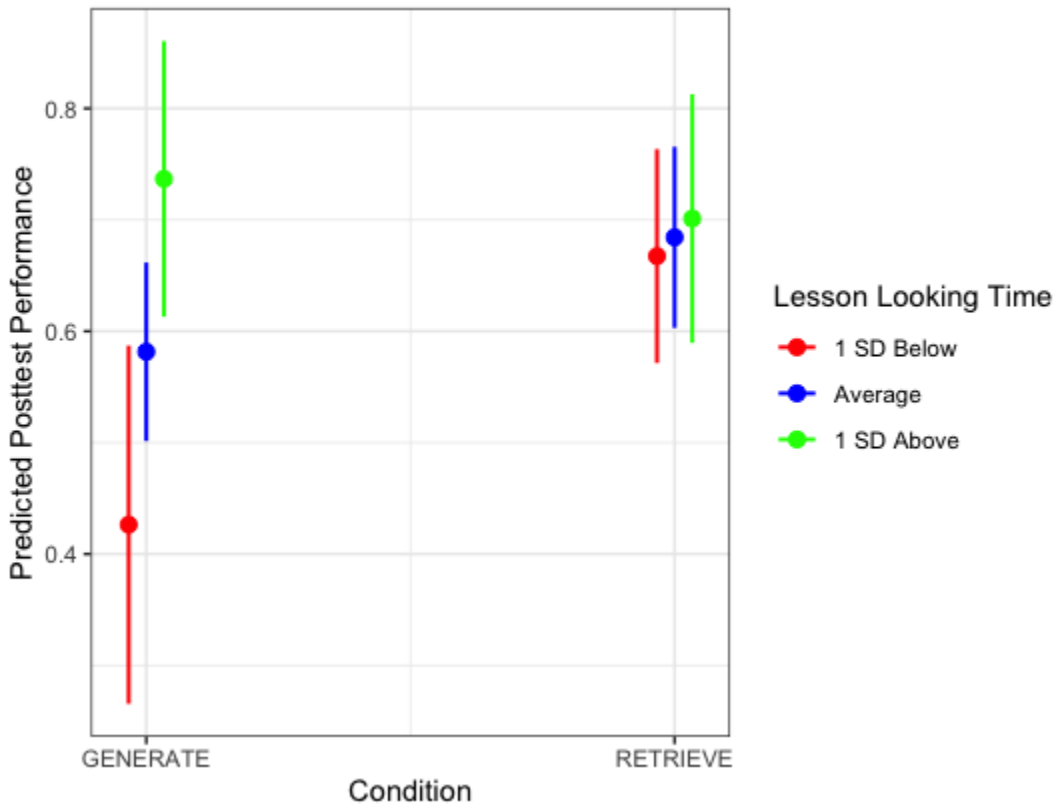
Note. Each dot represents average data from an individual participant. Posttest performance includes only critical trials. Lines represent best-fit regression lines for each condition.

Children in the Retrieve group were more accurate in the critical trials of the posttest (i.e., excluding the catch trials; M = 0.65, *95% CI* = [0.61, 0.68]) than in the Generating Predictions group (M = 0.61, *95% CI* = [0.57, 0.64]), β = 0.42, t (67) = 2.33, p = .023, *95% CI* = [-0.05,0.89]. Moreover, overall, as seen in Figure 4, children who spent a larger proportion of time fixating on the lesson materials during the lesson were more accurate at posttest, β = 0.63, t (67) = 2.60, p = .011, *95% CI* = [0.15,1.12]. Critically, the interaction effect between condition

and proportion of time fixating in the lesson on posttest performance was also statistically significant, β = -0.57, t (67) = -2.04, p = .04, *95% CI* = [-1.12, -0.01].

Figure 5

Marginal effects of regression model predicting posttest performance as a function of learning condition and proportion of looking time during the lesson.



To investigate this interaction we plotted the marginal effects of the interaction term and conducted planned contrasts. As can be seen in Figure 5, at low levels of fixation during the lesson, children in the Retrieve condition performed better in the posttest than children in the Generating Predictions condition, p = .012. However, this is not the case at higher levels of

fixation during the lesson. When children spend more time looking at the lesson materials, we found no evidence of a difference between the two conditions, $p = .6712$.

Discussion

The main question of this work was whether generating predictions or retrieval practice would yield better learning outcomes for young children learning new science facts. Previous research has suggested both active learning approaches are beneficial (e.g., Karpicke et al., 2016; Marsh et al., 2012; Pashler et al., 2007; Rittle-Johnson & Kmicikewycz, 2008; Roediger & Karpicke, 2006a; Slamecka & Graf, 1978). We hypothesized that the key difference between the two approaches is the attentional and encoding demands. Whereas generating predictions produces greater attention to the materials during encoding and in turn boosts performance, retrieval practice delays forgetting of information presented, such that even less well attended and encoded information will be remembered and at levels perhaps sufficient for good performance in delayed tests. Importantly, because children's attentional control system is still developing (Fisher & Kloos, 2016; Ruff & Rothbart, 2001), we expected that children with more developed attentional skills would be able to focus their attention on the information presented while generating predictions and thus benefit the most from it. Consistent with these predictions, and the results of our simulations using a simple version of a memory mechanism, we found that (a) children attended more to the materials when learning through generating predictions, but (b) performed better at posttest when learning through retrieval. Importantly, attention to the materials varied less for children who were learning through generating predictions, and, in that condition only, those who attended to the instructional materials the most, benefited the most.

Taken as a whole, the results presented here are consistent with previous evidence suggesting that generating predictions can improve learning, compared to retrieval, by increasing attention and encoding of critical information (e.g., Brod, 2021; McNamara & Healy, 2000). Our findings add an important caveat: such increases in attention and encoding are not equally distributed across children and might critically depend on children's a priori attentional

regulation abilities. Taking into account children's attention regulation abilities may be particularly important for primary school children given the protracted developmental trajectory of endogenously regulated attention (e.g., Ruff & Rothbart, 2001). To date, this link is only suggestive, but this work provides an important foundation for future research in which scientists can collect independent measures of attention in order to more directly assess whether individual differences in attention regulation predict children's attention to the instructional materials which may in turn predict children's learning.

Moreover, we proposed a detailed mechanism for why and when generating predictions or retrieval would improve learning. Provided that children can successfully regulate their attention in order to attend to the instructional materials, generating predictions will improve learning through increased initial encoding, as seen by the smaller variation in looking time in the Generate predictions condition and stronger relation between looking time and posttest performance in that condition. That is, under some circumstances, the greater initial encoding provided by generating predictions can boost learning more than the hypothesized slower decay seen for retrieval practice.

It is important to note that our measure of initial encoding was how much children looked at the materials during study. Eye tracking has been widely used as a measure of attention (e.g., Blair et al., 2009; Deng & Sloutsky, 2012), but its relation to initial encoding is an indirect measure of how well the information is encoded. Future research should include additional measures of initial encoding such as a memory or change task. Furthermore, our research used only an immediate posttest. As mentioned previously, retrieval practice is hypothesized to improve learning by slowing forgetting (Roediger & Karpicke, 2006a). As such, the benefits of retrieval practice have been shown to increase with increased retention intervals (Roediger & Butler, 2011). As can be seen from the model simulations in Figure 1, if the retention interval is increased and the function remains the same, at some point retrieval practice will result in a higher probability of recall than generating predictions, even when generating predictions

increases initial encoding substantially compared to retrieval practice. Although it remains an open question at what point increases in initial encoding from generating predictions can surpass the slower forgetting of retrieval and result in improved test performance, it is worth noting that even at short retention intervals as used in this study, retrieval practice also resulted in better learning for children who did not attend to the study materials. One interesting future area of research would be to systematically vary attention during generation and retrieval intervals to understand the interaction between the two and test the limits of generating predictions as well as explore the specific threshold of attention required to observe a reversal effect.

As one would expect, children were substantially more accurate on the classification pages embedded within the lesson when retrieving the information just presented than when having to generate a prediction before information is presented. Thus, one potential issue with generating predictions is the likelihood of encoding incorrect information which could be used later during the final test resulting in worse performance (e.g, Storm & Nestojko, 2010). For that reason, we provided both groups with feedback in the form of the correct information, varying only on whether that information was presented before or after the prediction. It is unlikely that generating predictions without feedback would have produced the benefits seen here (Klahr & Nigam, 2004), as children would not have had the correct information to encode after making a prediction. Success during the retrieval task itself has also been shown to influence the benefit of retrieval practice (Kornell et al., 2011), and it is possible that children would have benefited more from retrieval practice than what we observed here if we had also provided feedback after each response (Pyc & Rawson, 2011). However, given children's overall high accuracy during the lesson in the retrieve condition, this is unlikely to be a great concern in the present study.

The current results have clear theoretical implications; however, they also provide a foundation for future translational research to begin exploring the potential educational implications of this line of work. For example, the present study may ultimately provide insights

for educational practice. When educators are designing instruction, it is important to utilize techniques that are aligned with children's developing cognitive systems. As discussed previously, attention regulation is important for learning, but the ability to effectively regulate attention is a cognitive skill that develops slowly. Educators should be mindful that certain instructional techniques may have higher attentional demands and thus may be less effective for both younger children, whose attentional system is still developing, and children with weaker attention regulation skills. Although suggestive, the present study indicates that generating predictions may be one technique that requires stronger attention regulation (assuming attention is not primarily exogenously regulated during this task), and as a result, its benefits may be somewhat curtailed when this instructional technique is utilized with younger children or children who have attention difficulties. Future research can explore whether providing primary school children with appropriate attentional scaffolding, could enable even young children to benefit from generating predictions as potentially an effective and more motivating (Brod, 2021; Brod et al., 2018) instructional approach than retrieval. Future research should test this hypothesis by collecting independent measures of attention in order to evaluate more fully the role of individual differences in attention on learning via generating predictions. The prior literature points to sustained benefits of retrieval practice and the present results may further encourage educators to leverage retrieval practice as an instructional technique to benefit learning, particularly given that retrieval practice seems to require less attentional control than generating predictions. However, it is unclear if retrieval practice works as well for more complex learning tasks as opposed to more simplified learning tasks that require recall and application of simple facts, as we used here (i.e., if higher order generalization questions were used, would retrieval practice still be as good?).

It is also important to acknowledge that the present study was conducted with a largely Caucasian (74%) sample. Although the sample demographics were generally reflective of the

county in which the data were collected, future research should assess the extent to which the findings can be generalized to more diverse populations.

In conclusion, the findings of the present study suggest that generating predictions can be as effective as retrieval practice of facts for primary school children. However, our model simulations and previous evidence in the literature (e.g., Brod, 2020) suggest that the attentional control abilities of the students, which will increase with age, may be key to this benefit. These findings highlight the importance of considering student level factors such as students' age as well as individual differences among students when deciding the best learning strategies to implement.

References

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A Theory of Higher Level Cognition and Its Relation to Visual Attention. *Hum.-Comput. Interact.*, *12*(4), 439–462. https://doi.org/10/bmhrqt

Ashman, G., Kalyuga, S., & Sweller, J. (2020). Problem-solving or explicit instruction: Which should go first when element interactivity is high?. Educational Psychology Review, 32, 229-247.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210. https://doi.org/10.3758/BF03193441

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1196–1206.

Bornstein, M. H. (1990). Attention in infancy and the prediction of cognitive capacities in childhood. In J. T. Enns (Ed.), *Development of attention: Research and theory* (pp. 3–19). Elsevier. https://www.sciencedirect.com/science/article/abs/pii/S0166411508604483

Brod, G. (2020). Generative Learning: Which Strategies for What Age? *Educational Psychology Review*. https://doi.org/10/ghgjrw

Brod, G. (2021). Predicting as a learning strategy. *Psychonomic Bulletin & Review*. https://doi.org/10/gjj3bd

Brod, G., & Breitwieser, J. (2019). Lighting the wick in the candle of learning: Generating a prediction stimulates curiosity. *Npj Science of Learning*, *4*(1), 17. https://doi.org/10/gjqgmn

Brod, G., Hasselhorn, M., & Bunge, S. A. (2018). When generating a prediction boosts learning: The element of surprise. *Learning and Instruction*, *55*, 22–31.

https://doi.org/10.1016/j.learninstruc.2018.01.013

Carvalho, P. F., Manke, K., & Koedinger, K. (2018). Not all Active Learning is Equal: Predicting and Explaining Improves Transfer Relative to Answering Practice Questions. In CogSci.

Chi, M. T. H. (2009). Active-Constructive-Interactive: A Conceptual Framework for Differentiating Learning Activities. *Topics in Cognitive Science*, *1*(1), 73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

Chi, M. T. H. (2021). Translating a Theory of Active Learning: An Attempt to Close the Research-Practice Gap in Education. *Topics in Cognitive Science*, *13*(3), 441–463. https://doi.org/10/gj3qm7

Deng, W., & Sloutsky, V. M. (2012). *The Role of Linguistic Labels in Infants' Categorization: An Eye Tracking Study*.

Diamond, A. (2006). The Early Development of Executive Functions. In *Lifespan Cognition: Mechanisms of Change* (pp. 70–95). https://doi.org/10.1093/acprof:oso/9780195169539.003.0006

Erickson, L. C., Thiessen, E. D., Godwin, K. E., Dickerson, J. P., & Fisher, A. V. (2015). Endogenously and exogenously driven selective sustained attention: Contributions to learning in kindergarten children. *Journal of Experimental Child Psychology*, *138*, 126–134. https://doi.org/10.1016/j.jecp.2015.04.011

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

Fisher, A. V., Godwin, K. E., & Seltman, H. (2014). Visual Environment, Attention Allocation, and Learning in Young Children. *Psychological Science*, *25*(7), 1362–1370. https://doi.org/10.1177/0956797614533801

Fisher, A. V., & Kloos, H. (2016). Development of selective sustained attention: The role of executive functions. In *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research* (pp. 215–237). American

Psychological Association. https://doi.org/10.1037/14797-010

Fisher, A. V., Thiessen, E., Godwin, K., Kloos, H., & Dickerson, J. (2013). Assessing selective sustained attention in 3- to 5-year-old children: Evidence from a new paradigm. *Journal of Experimental Child Psychology*, *114*(2), 275–294. https://doi.org/10.1016/j.jecp.2012.07.006

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2014). An alternative time for telling: When conceptual instruction prior to problem solving improves mathematical knowledge. British journal of educational psychology, 84(3), 502-519.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., Dirks, C., & Wenderoth, M. P. (2007). Prescribed Active Learning Increases Performance in Introductory Biology. *CBE—Life Sciences Education*, *6*(2), 132–139. https://doi.org/10/fk8xq2

Godwin, K. E., Almeda, Ma. V., Seltman, H., Kai, S., Skerbetz, M. D., Baker, R. S., & Fisher, A. V. (2016). Off-task behavior in elementary school children. *Learning and Instruction*, *44*, 128–143. https://doi.org/10.1016/j.learninstruc.2016.04.003

Godwin, K. E., Leroux, A. J., Scupelli, P., & Fisher, A. V. (2022). Classroom Design and Children's Attention Allocation: Beyond the Laboratory and into the Classroom. *Mind, Brain, and Education*, *16*(3), 239–251. https://doi.org/10.1111/mbe.12319

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased Structure and Active Learning Reduce the Achievement Gap in Introductory Biology. *Science*, *332*(6034), 1213–1216. https://doi.org/10/bcmj3b

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*. 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., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, *62*(3), 227–239.

Karweit, N. L. (1983). Time on Task: A Research Reveiw. *Report No. 332*.

Klahr, D., & Nigam, M. (2004). The Equivalence of Learning Paths in Early Science Instruction: Effects of Direct Instruction and Discovery Learning. *Psychological Science*, *15*(10), 661–667. https://doi.org/10.1111/j.0956-7976.2004.00737.x

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Lovett, M. C., Daily, L. Z., & Reder, L. M. (2000). A source activation theory of working memory: Cross-task prediction of performance in ACT-R. *Cognitive Systems Research, 1*(2), 99-118.

Luna, B. (2009). DEVELOPMENTAL CHANGES IN COGNITIVE CONTROL THROUGH ADOLESCENCE. *Advances in Child Development and Behavior*, *37*, 233–278.

Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. *Memory*, *20*(8), 899–906. https://doi.org/10.1080/09658211.2012.708757

Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on children's acquisition of experimentation skills: Is it all in the timing? *Instructional Science*, *41*(3), 621–634. https://doi.org/10.1007/s11251-012-9248-z

McFarland, C. E., Duncan, E., & Bruno, J. M. (1983). Developmental aspects of the generation effect. *Journal of Experimental Child Psychology*, *36*(3), 413–428.

https://doi.org/10.1016/0022-0965(83)90043-7

McNamara, D. S., & Healy, A. F. (2000). A Procedural Explanation of the Generation Effect for Simple and Difficult Multiplication Problems and Answers. *Journal of Memory and Language*, *43*(4), 652–679. https://doi.org/10.1006/jmla.2000.2720

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, *14*(2), 187–193.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101–117. https://doi.org/10.1037/1076-898X.14.2.101

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. https://doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2011). Costs and benefits of dropout schedules of test–restudy practice: Implications for student learning. *Applied Cognitive Psychology*, *25*(1), 87–95. https://doi.org/10.1002/acp.1646

Ritchie, S. J., Sala, S. D., & McIntosh, R. D. (2013). Retrieval Practice, with or without Mind Mapping, Boosts Fact Learning in Primary School Children. *PLOS ONE*, *8*(11), e78976. https://doi.org/10.1371/journal.pone.0078976

Rittle-Johnson, B., & Kmicikewycz, A. O. (2008). When generating answers benefits arithmetic skill: The importance of prior knowledge. *Journal of Experimental Child Psychology*, *101*(1), 75–81. https://doi.org/10.1016/j.jecp.2008.03.001

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. (2006a). Test-enhanced learning—Taking memory tests

improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Roediger, H. L., & Karpicke, J. D. (2006b). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

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

Ruff, H. A., & Rothbart, M. K. (2001). *Attention in Early Development: Themes and Variations*. Oxford University Press.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Sterling-Turner, H., Robinson, S., & Wilczynski, S. (2001). Functional Assessment of Distracting and Disruptive Behaviors in the School Setting. *School Psychology Review*, *30*. https://doi.org/10.1080/02796015.2001.12086110

Storm, B. C., & Nestojko, J. F. (2010). Successful inhibition, unsuccessful retrieval: Manipulating time and success during retrieval practice. *Memory*, *18*(2), 99–114. https://doi.org/10.1080/09658210903107853

Sutterer, D. W., & Awh, E. (2016). Retrieval practice enhances the accessibility but not the quality of memory. *Psychonomic Bulletin & Review*, *23*(3), 831–841. https://doi.org/10.3758/s13423-015-0937-x

Varga, N. L., & Bauer, P. J. (2013). Effects of delays on 6-year-old children's self-generation and retention of knowledge through integration. *Journal of Experimental Child Psychology*, *115*(2), 326–341. https://doi.org/10.1016/j.jecp.2013.01.008

Winstanley, P. A. de, Bjork, E. L., & Bjork, R. A. (1996). Generation Effects and the Lack Thereof: The Role of Transfer-appropriate Processing. *Memory*, *4*(1), 31–48.

https://doi.org/10.1080/741940667

Yannier, N., Hudson, S. E., Koedinger, K. R., Hirsh-Pasek, K., Golinkoff, R. M., Munakata, Y., Doebel, S., Schwartz, D. L., Deslauriers, L., McCarty, L., Callaghan, K., Theobald, E. J., Freeman, S., Cooper, K. M., & Brownell, S. E. (2021). Active learning: “Hands-on” meets “minds-on.” *Science*, *374*(6563), 26–30. https://doi.org/10.1126/science.abj9957

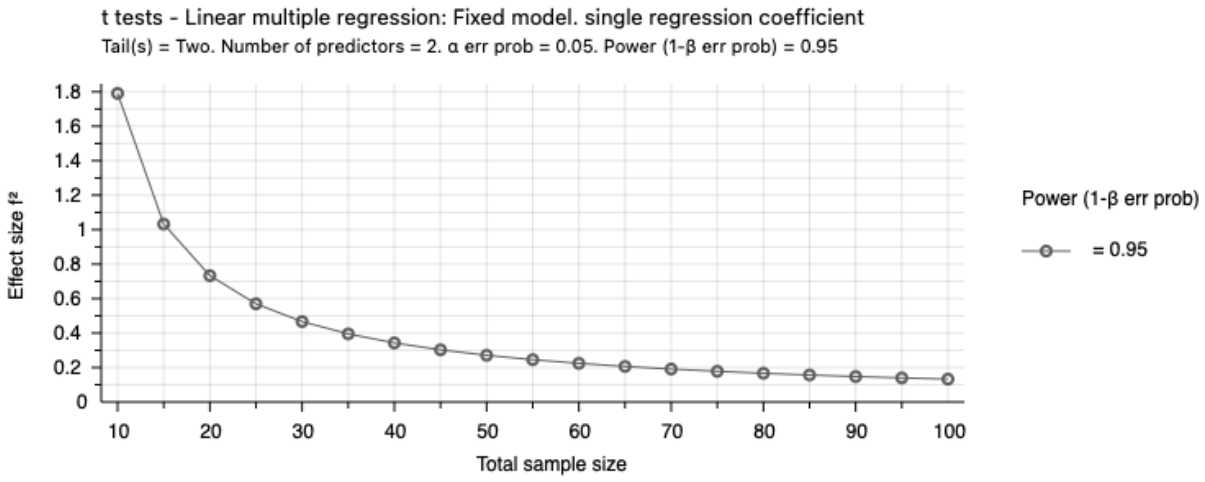
Funding Statement

This work was supported in part by the National Science Foundation Grant BCS-1824257 to PFC.

Appendix

Sensitivity analyses

Because we did not conduct an *a priori* power analysis, instead using prior sample sizes as a guide for our target sample size, we ran a sensitivity analysis using g*power (Faul et al., 2009) for the main effect of interest, i.e., the difference between the two conditions in attention. As can be seen in the figure below, this sensitivity analysis suggests that at a sample size of 71 for the eye-tracking analyses, we had the power to detect an effect as small as f^2 = 0.188 (d = 0.376). We found an effect of d = 1.92.



Attention: Track-It.

We administered an independent, performance-based, measure of attention, Track-It (Fisher et al., 2013). Track-It is a computer task, in which children visually track a target shape (e.g., an orange triangle) as it moves around the computer screen amongst heterogeneous distractors. The shapes disappear at the end of each trial, and the child is asked to report the last location the target visited prior to disappearing. A memory check is also administered in which children are asked to report which shape they had been tracking during the preceding

trial. The memory check is designed to help dissociate the contribution of working memory and attention on task performance. For example, if a child succeeds on the memory check, it suggests they have sufficiently encoded the target. Under these circumstances, a failure to identify the target's hiding location suggests a failure of attention vs. a working memory limitation. On the other hand, if a child fails to correctly identify the target shape on the memory check the contribution of working memory and attention to the child's tracking accuracy is difficult to disentangle. Accordingly, for each child, a corrected accuracy score was calculated in which children's tracking accuracy was calculated only for those trials in which they had successfully identified the target shape on the memory check.

As noted previously the analysis and interpretation of the Track-It data is unfortunately limited as only a subset of the children (n= 47) had Track-it data that could be analyzed, due to the following reasons: missing data (n=12), task completion issues (n=10), administration error in task parameters (n=1), and/or a delay between administration of the primary task and Track-It as a subset of participants (n=21) were part of a larger parent study. The mean delay for these children was 129 (SD = 46) days. As a result, the Track-It data was not a focus of the present paper; however, we elected to include the analysis here, for completeness.

Due to the issues with missing data described above, the analysis presented below is based on a small subset of children (N = 47). Mean accuracy on the Track-It task was 85% [0.5:1]. Overall performance in the posttest and Track-it were correlated (r (45) = 0.34, p = .019). These results tentatively suggest that children who exhibited better attention regulation (as indexed by a performance-based measure of attention, Track-It) tended to perform better on the learning assessment. The correlation between Track-It and the learning assessments was positive for both conditions, and of greater magnitude for the generating predictions condition (r (22)2 = 0.38, p = .071) than the retrieve condition (r2 (21) = 0.32, p = .135); however, neither correlation was statistically significant. It is important to view the Track-It data cautiously due to

the large amount of missing data. Thus, collecting a performance-based measure of attention remains an important step for future research.