**Supplemental Materials for**

**To Ask Better Questions, Teach: Learning-by-Teaching Enhances Research Question Generation More Than Retrieval Practice and Concept-Mapping**

**Table of Contents**

|  |  |
| --- | --- |
| **Content** | **Page** |
| Practice Text on “Enzymes” | 2 |
| Study Text on “Food Allergies” | 3 |
| Idea Units in Study Text on “Food Allergies”  | 4 |
| Study Text on “Intelligence Quotient” | 6 |
| Idea Units in Study Text on “Intelligence Quotient” | 7 |
| Table S1: Pre-Learning Questionnaire Items (Experiments 1 and 2) | 9 |
| Table S2: Standard Questions (Experiments 1 and 2) | 11 |
| Table S3: Standard Questions Presentation Frequency (Experiments 1 and 2) | 14 |
| Supplementary Experiment 3a | 16 |
| Supplementary Experiment 3b | 22 |
| Supplementary Experiment 3c | 26 |
| Supplementary References | 33 |

**Practice Text on “Enzymes”**

**(adapted from Meyer, 1975)**

**ENZYMES**

In enzymatic reactions, molecules called substrates are converted into different molecules called products. Before a reaction can take place, the activation energy must be overcome. Activation energy can be thought of as the minimum energy required to start a chemical reaction. An enzyme catalyzes a reaction by lowering the activation energy needed to allow the reaction to take place. For example, hydrogen peroxide breaks down into water and oxygen gas on its own, but at an incredibly slow rate. The enzyme catalase lowers the activation energy of the reaction and the reaction happens very quickly.

Catalytic activity is greatly affected by temperature. Since enzymes are proteins, they lose their structure at high temperatures. This not only eliminates the catalytic properties, but essentially destroys the protein. Different enzymes have different specific temperatures. The activity of animal catalase peaks at about normal body temperature, or 35-40°C. Once the temperature increases beyond this temperature range, the catalase proteins die.

Increasing temperature will also increase the amount of free energy because all the molecules will be moving faster. This increased speed and temperature results in an increased rate of collision between the enzyme and substrate molecules, which leads to a faster reaction time.

**Study Text on “Food Allergies”**

**(adapted from Griffin et al., 2019)**

**FOOD ALLERGIES**

Antibiotics are drugs used to kill bacteria. Since their discovery in the 1930s, antibiotics have cured many diseases. Diseases like pneumonia, tuberculosis, and meningitis are caused by bacteria. Yet, some bacteria that naturally exist in our body help protect us from diseases. For example, bacteria in our digestive tract manufacture B-vitamins and folic acid. They also provide the enzyme lactase for food digestion.

One of the most important functions of bacteria in our digestive tract is to keep in check the growth of parasites, fungi, and yeasts. In a healthy situation, the small intestine epithelium maintains tight cell junctions. This contributes to the physical barrier in intestinal absorption.

The situation changes when people take antibiotics, and consequently, there are not enough bacteria to control the growth of yeasts such as Candida. Candida produces aldehyde secretions that make small intestine epithelial cells shrink, and the cell junctions loosen. This allows intestinal toxins to infiltrate through the epithelium and pass into the blood.

When the intestinal barrier has been compromised by antibiotic use, intestinal toxins are not the only pathogens to be absorbed. The barrier, in a healthy state, selectively allows nutrients to enter the small intestine only once they are fully digested. However, when the barrier is compromised, nutrients can be absorbed before they are fully digested. Just like when intestinal toxins cross the barrier, this prompts an immune response. The body’s immune system will use specific antigen-antibody markers to tag these foods as foreign irritants. Once these foods are tagged as irritants, an immune response will be mounted every time in the future that the particular food touches the epithelia. What started as a Candida irritation with shrinking of the epithelial cells has now been complicated with active inflammation every time a particular food is eaten. The allergic reaction to food particles can be manifested as skin rashes and inflammations.

**Idea Units in Study Text on “Food Allergies”**

1. Antibiotics are drugs / used to kill bacteria.
2. Since their discovery in the 1930s, / antibiotics have cured many diseases.
3. Diseases like pneumonia, / tuberculosis, / and meningitis are caused by bacteria.
4. Yet, some bacteria that naturally exist in our body / help protect us from diseases.
5. For example, bacteria in our digestive tract manufacture B-vitamins / and folic acid.
6. They also provide the enzyme lactase / for food digestion.
7. One of the most important functions of bacteria in our digestive tract is to keep in check the growth of parasites, / fungi, / and yeasts.
8. In a healthy situation, the small intestine epithelium maintains tight cell junctions.
9. This contributes to the physical barrier in intestinal absorption.
10. The situation changes when people take antibiotics, and consequently, there are not enough bacteria to control the growth of yeasts / such as Candida.
11. Candida produces aldehyde secretions / that make small intestine epithelial cells shrink, / and the cell junctions loosen.
12. This allows intestinal toxins to infiltrate through the epithelium / and pass into the blood.
13. When the intestinal barrier has been compromised by antibiotic use, intestinal toxins are not the only pathogens to be absorbed.
14. The barrier, in a healthy state, selectively allows nutrients to enter the small intestine / only once they are fully digested.
15. However, when the barrier is compromised, nutrients can be absorbed before they are fully digested.
16. Just like when intestinal toxins cross the barrier, this prompts an immune response.
17. The body’s immune system will use specific antigen-antibody markers / to tag these foods as foreign irritants.
18. Once these foods are tagged as irritants, an immune response will be mounted / every time in the future / that the particular food touches the epithelia.
19. What started as a Candida irritation with shrinking of the epithelial cells / has now been complicated with active inflammation / every time a particular food is eaten.
20. The allergic reaction to food particles can be manifested as skin rashes / and inflammations.

*Note*. Each idea unit is demarcated by a dash (‘/’). Each idea unit that participants recalled verbatim or paraphrased appropriately was awarded one point. There are 310 words, 4 paragraphs, 20 sentences, and 40 idea units in this passage. It has a Flesch-Kincaid grade level of 12.4.

**Study Text on “Intelligence Quotient”**

**(adapted from Griffin et al., 2019)**

**INTELLIGENCE QUOTIENT**

*(Note: This text was designed to argue for a particular viewpoint rather than reflect current scientific consensus.)*

The notion that ethnic groups might differ in intelligence lacks evidence. This is logically implausible, as a mere difference in IQ scores does not provide evidence of real differences in any kind of general mental ability.

A fundamental problem is that there is no such thing as a single “intelligence” that measures a given person’s cognitive ability. To behave intelligently means to adapt to one’s environment and make appropriate choices. Different situations require different types of adaptations, and therefore different “intelligences.” Professionals from very different jobs use different mental skills to perform their jobs well. IQ tests reduce many different abilities and aspects of thinking down to a single score.

The differences on IQ test scores are also not caused by biological differences between races. Ethnic groups differ genetically in only very superficial ways. Usually, limited genetic differences could be involved. This results in differing gene expressions, manifesting as skin color, height, and other differences in outward physical appearance. Solving complex mental problems, however, involves many aspects of our cognitive abilities. These mental abilities are the result of complex gene-environment interactions. Even identical twins differ in IQ, and siblings are often more different from each other than to people they are not related to.

The social conditions of the child’s upbringing play a large role in how the child performs on IQ tests. On average, most minority groups are lower in economic and social status than the majority racial groups. Quality of education and pedagogical factors like class size would be influenced by a given child’s social conditions. Those studies that show ethnic differences in IQ often look at students who had already completed their primary school education. These students had already been affected by poor quality education. In addition, parents and grandparents who are more highly educated will be able to better facilitate their child’s or grandchild’s educational attainment.

**Idea Units in Study Text on “Intelligence Quotient”**

1. The notion that ethnic groups might differ in intelligence / lacks evidence.
2. This is logically implausible, as a mere difference in IQ scores / does not provide evidence of real differences in any kind of general mental ability.
3. A fundamental problem is that there is no such thing as a single “intelligence” / that measures a given person’s cognitive ability.
4. To behave intelligently means to adapt to one’s environment and / make appropriate choices.
5. Different situations require different types of adaptations, and / therefore different “intelligences.”
6. Professionals from very different jobs use different mental skills to / perform their jobs well.
7. IQ tests reduce many different abilities / and aspects of thinking / down to a single score.
8. The differences on IQ test scores are also not caused by biological differences / between races.
9. Ethnic groups differ genetically in only very superficial ways.
10. Usually, limited genetic differences could be involved.
11. This results in differing gene expressions, / manifesting as skin color, / height, / and other differences in outward physical appearance.
12. Solving complex mental problems, however, involves many aspects of our cognitive abilities.
13. These mental abilities are the result of complex gene-environment interactions.
14. Even identical twins differ in IQ, / and siblings are often more different from each other than to people they are not related to.
15. The social conditions of the child’s upbringing play a large role in / how the child performs on IQ tests.
16. On average, most minority groups are lower in economic and social status / than the majority racial groups.
17. Quality of education / and pedagogical factors like class size / would be influenced by a given child’s social conditions.
18. Those studies that show ethnic differences in IQ / often look at students who had already completed their primary school education.
19. These students had already been affected by / poor quality education.
20. In addition, parents and grandparents who are more highly educated / will be able to better facilitate their child’s or grandchild’s educational attainment.

*Note*. Each idea unit is demarcated by a dash (‘/’). Each idea unit that participants recalled verbatim or paraphrased appropriately was awarded one point. There are 310 words, 4 paragraphs, 20 sentences, and 40 idea units in this passage. It has a Flesch-Kincaid grade level of 12.3.

**Table S1**

*Pre-Learning Questionnaire Items (Experiments 1 and 2)*

|  |  |  |
| --- | --- | --- |
| **Variable** | **“Food Allergies”** | **“Intelligence Quotient”** |
| Prior Knowledge Rating(1 = *not at all*; 5 = *a lot*) | How much prior knowledge do you have about the biological basis of Food Allergies? | How much prior knowledge do you have about Intelligence Quotient?  |
| Prior Knowledge Content Items (*Yes* / *No*) | I have an academic background in Biology (i.e., university course/GCE O- or A-Levels/Diploma). | I have an academic background in Psychology (i.e., university course/GCE O- or A-Levels/Diploma). |
| I know what lactase is. | I know what cognitive processes are. |
| I know that there are intestinal toxins. | I know that every individual’s Intelligence Quotient could differ, from person to person. |
| I know about yeasts like Candida. | I know that different contexts might require different intelligences. |
| I understand what folic acid is. | I understand what gene-environment interactions are. |
| I understand what is meant by ‘antigen-antibody markers’. | I understand what is meant by ‘gene expression’. |
| I can define what the small intestine epithelial cells are. | I can define the term ‘pedagogy’. |
| I can describe how inflammations occur. | I can name a test that can measure Intelligence Quotient. |
| Judgment of Learning (JOL)(1 = *very poorly*; 5 = *very well*) | How well do you think you will do on a test on Food Allergies? | How well do you think you will do on a test on Intelligence Quotient? |

*Note.* Participants’ prior knowledge was measured by summing their prior knowledge rating (out of 5) and the number of content items that they reported having prior knowledge of (out of 8), with a maximum possible score of 13.

**Table S2**

*Standard Questions (Experiments 1 and 2)*

|  |  |
| --- | --- |
| **“Food Allergies”** | **“Intelligence Quotient”** |
| **Code** | **Question** | **Code** | **Question** |
| **FA-A1** | Could you illustrate how killing bacteria can be good and bad, using examples? Please try to explain your examples in your own words. | **IQ-E1** | Could you give examples of how different ethnic groups do not differ in terms of intelligence? Please try to explain your examples in your own words. |
| **FA-A2** | Could you illustrate how the epithelium can be compromised, using examples? Please try to elaborate on these examples in your own words. | **IQ-E2** | Could you give examples of how different situations or jobs might need different kinds of ‘intelligence’? Please try to explain your examples in your own words. |
| **FA-A3** | Could you illustrate how the growth of yeasts can compromise the epithelium? Please try to elaborate on this in your own words. | **IQ-E3** | Could you give a reason why gene-environment interactions leading to mental abilities are complex? Please try to explain this using your own words. |
| **FA-A4** | Could you give an example for how someone might develop food allergies because of an immune response? Please try to put this in your own words. | **IQ-E4** | Could you give an example of how one’s social conditions could affect how they perform on IQ tests? Please try to explain your example in your own words. |
| **FA-B1** | How might doctors or pharmacists learn from the positive and negative roles that bacteria can play? | **IQ-F1** | Different ethnic groups do not differ in intelligence, but their IQ scores might differ. Let’s say that we find a difference in IQ scores between North and South Americans. What can we conclude? |
| **FA-B2** | We know that food allergies might develop because of yeast overgrowth. How might we treat food allergies, knowing this? | **IQ-F2** | How would you apply the idea that different professions require different kinds of ‘intelligence’? |
| **FA-B3** | We know that food allergies might develop because of how the epithelial junction has been compromised. How might we treat food allergies, knowing this? | **IQ-F3** | How should governments and policymakers apply the fact that different races do not differ in intellectual or mental abilities? |
| **FA-B4** | We know that food allergies might develop because of how immune responses happen. How might we treat food allergies, knowing this? | **IQ-F4** | A child’s social conditions can affect their educational or IQ test performance. How should this change how we use IQ tests? |
| **FA-C1** | Could you summarize and explain why the bacteria-killing properties of antibiotics could be bad? | **IQ-G1** | Could you summarize and explain how different ethnic groups might not differ in intelligence? |
| **FA-C2** | Could you summarize and explain the role of the small intestine epithelium? | **IQ-G2** | Could you summarize and explain why there is no single ‘intelligence’ that measures someone’s cognitive ability? |
| **FA-C3** | Could you summarize and explain how the growth of yeasts like Candida happens? | **IQ-G3** | Could you summarize and explain why genetic differences between races are usually superficial? |
| **FA-C4** | Could you summarize and explain how the immune response is triggered? | **IQ-G4** | Could you summarize and explain why social conditions would affect a child’s upbringing and intellectual development? |
| **FA-D1** | How do the bacteria-killing properties of antibiotics link to the development of food allergies? | **IQ-H1** | Different ethnic groups do not differ in intelligence, but their IQ scores might differ. What does this mean? |
| **FA-D2** | How does the functioning of the epithelium relate to the development of food allergies? | **IQ-H2** | Why do we need more ways to measure ‘intelligence’, above and beyond traditional IQ scores? |
| **FA-D3** | How does the growth of yeasts like Candida contribute to the development of food allergies? | **IQ-H3** | Given that differences in IQ test scores aren’t caused by biological differences, how does this relate to the overall argument of the text? |
| **FA-D4** | How does the triggering of the immune response once contribute to the development of food allergies? | **IQ-H4** | How does children’s upbringing lead to their intellectual development? |

**Table S3**

*Standard Questions Presentation Frequency (Experiments 1 and 2)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Question | Frequency in Learning-by-Teaching Condition (Experiment 1) | Frequency in Concept-Mapping Condition (Experiment 2) | Frequency in Retrieval Practice Condition (Experiment 2) | Frequency in Learning-by-Teaching Condition (Experiment 2) |
| **FA-A1** | 4.82% | 5.68% | 4.21% | 4.51% |
| **FA-A2\*** | 0.60% | - | - | - |
| **FA-A3** | 2.41% | 2.27% | 0.00% | 0.75% |
| **FA-A4** | 1.20% | 2.27% | 2.11% | 0.00% |
| **FA-B1** | 8.43% | 6.82% | 7.37% | 7.52% |
| **FA-B2** | 7.23% | 4.55% | 5.26% | 8.27% |
| **FA-B3** | 4.82% | 3.41% | 8.42% | 7.52% |
| **FA-B4** | 1.81% | 0.00% | 1.05% | 0.75% |
| **FA-C1** | 6.63% | 6.82% | 8.42% | 6.02% |
| **FA-C2** | 6.02% | 3.41% | 2.11% | 2.26% |
| **FA-C3** | 3.01% | 3.41% | 3.16% | 4.51% |
| **FA-C4** | 1.20% | 1.14% | 1.05% | 3.76% |
| **FA-D1** | 2.41% | 3.41% | 2.11% | 3.01% |
| **FA-D2** | 3.61% | 5.68% | 4.21% | 1.50% |
| **FA-D3** | 1.81% | 1.14% | 1.05% | 2.26% |
| **FA-D4\*** | 0.00% | - | - | - |
| **IQ-E1** | 3.01% | 2.27% | 4.21% | 0.75% |
| **IQ-E2** | 4.82% | 10.23% | 6.32% | 7.52% |
| **IQ-E3\*** | 0.60% | - | - | - |
| **IQ-E4** | 3.01% | 3.41% | 5.26% | 4.51% |
| **IQ-F1\*** | 0.00% | - | - | - |
| **IQ-F2** | 7.23% | 7.95% | 5.26% | 3.01% |
| **IQ-F3** | 1.20% | 2.27% | 1.05% | 0.75% |
| **IQ-F4** | 7.23% | 6.82% | 6.32% | 5.26% |
| **IQ-G1** | 4.82% | 4.55% | 5.26% | 4.51% |
| **IQ-G2** | 2.41% | 3.41% | 3.16% | 6.02% |
| **IQ-G3** | 3.01% | 3.41% | 5.26% | 6.77% |
| **IQ-G4** | 1.20% | 1.14% | 2.11% | 1.50% |
| **IQ-H1\*** | 0.00% | - | - | - |
| **IQ-H2** | 3.01% | 3.41% | 4.21% | 6.02% |
| **IQ-H3** | 2.41% | 1.14% | 1.05% | 0.00% |
| **IQ-H4\*** | 0.00% | - | - | - |

*Note*. Each standard question’s presentation frequency was computed as a percentage of the total standard questions that were presented in each learning condition. Six questions marked with an asterisk (\*) were not administered in Experiment 2 due to the relatively low frequency (< 1% of all standard questions) at which they were presented in Experiment 1.

**Supplementary Experiment 3a**

Experiment 3a investigated whether retrieval practice confers any benefits for generating questions that correspond more broadly to the “lower-order” (*remember* and *understand*) and “higher-order” (*apply*, *analyze*, *evaluate*, and *create*) levels of Bloom’s taxonomy. We compared retrieval practice against restudy, which is an especially common control condition in extant studies on the testing effect (Karpicke, 2017; Rowland, 2014). Specifically, learners studied a scientific expository text either by alternating between study and retrieval (Study–Retrieve–Study–Retrieve; SRSR) or by repeatedly studying it (Study–Study–Study–Study; SSSS). One week after their initial study, all learners returned to be trained on question generation, then completed a test in which they generated as many lower- and higher-order questions as they could based on the text that they had earlier studied, but without access to it. In addition, learners completed a free recall test in which they wrote down as much as they could remember from the study text.

**Method**

***Participants***

The participants were 70 undergraduates (46 were female) between the ages of 19 to 25 (*M* = 21.65, *SD* = 1.59). Outcomes reported here are based on data from 69 participants—one participant who reported having encountered the study materials before was excluded from analyses. Based on the effect size of practice testing with retention intervals of 1 to 6 days (*g* = 0.82) reported in Adesope et al.’s (2017) meta-analysis, a power analysis (G\*Power; Faul et al., 2007) indicated that at least 25 participants per condition were required to observe a retrieval-based learning effect for two-tailed between-subjects pairwise comparisons at 80% power and α = .05. In all supplementary experiments reported here, all participants provided their informed consent and received either course credit or monetary remuneration for their participation.

***Design***

Experiment 3a employed a between-subjects design with learning strategy as the primary independent variable, whereby participants were randomly assigned to either the *retrieval practice* or *restudy* condition. To ascertain that any effects of the learning strategies generalized across study topics, we also included study text as a second independent variable for control purposes, whereby participants were randomly assigned to study a text on either “The Human Ear” or “Tropisms”.

The three dependent variables of interest were assessed on a delayed test administered one week after initial study: (a) learners’ higher-order question generation performance, as assessed via the number of questions they posed that fulfilled the *apply*, *analyze*, *evaluate*, or *create* levels of Bloom’s taxonomy, (b) learners’ lower-order question generation performance, as assessed via the number of questions they posed that fulfilled the *remember* or *understand* levels of Bloom’s taxonomy, and (c) learners’ recall performance, as assessed via the number of idea units from the study texts that they correctly recalled.

***Materials***

**Study Texts.** The study materials comprised of two scientific expository texts on “The Human Ear” and “Tropisms” (adapted from Cook & Mayer, 1988; Karpicke & Blunt, 2011). The texts contained 259 and 263 words, 28 and 24 idea units, had Flesch Reading Ease scores of 68.8 and 62.4, and had Flesch-Kincaid grade levels of 7.6 and 8.3, respectively.

**Prior Knowledge Measure.** As a measure of their prior knowledge of the study materials, learners reported how much information in the text they knew prior to reading it, and how well they knew the subject matter covered in the text prior to reading it. Both ratings were performed on a 7-point scale (1 = *not very much*; 7 = *very much*). Prior knowledge was computed as learners’ mean rating across both items.

***Procedure***

The experiment consisted of two phases: study and test. In the study phase, learners were randomly assigned to study one of the two texts—either “The Human Ear” or “Tropisms”—via retrieval practice or restudy. In the *retrieval practice* (SRSR) condition, learners studied the text for 6 min, practiced retrieval for 6 min by writing down as much information from the text as they could recall without reference to it, then restudied the text for 6 min, and practiced retrieval again for 6 min (e.g., Karpicke & Blunt, 2011). In the *restudy* (SSSS) condition, learners (re)studied the text for four consecutive 6-min blocks. Thus, the total study duration was exactly 24 min across both the retrieval practice and restudy conditions. At the end of the 24-min study period, all learners completed the prior knowledge measure.

One week later, all learners returned for the test phase, during which they were first trained on question generation, as in Experiments 1 and 2. Specifically, all learners received a printed handout that described and explained the various question levels based on Bloom’s taxonomy (see Table 1), and were instructed on *factual questions* versus *thinking questions* (e.g., King, 1989). Whereas factual questions require simple recall or comprehension of facts and information that had been presented in the text (i.e., lower-order questions corresponding to the *remember* and *understand* levels of Bloom’s taxonomy), thinking questions require drawing connections among different ideas, such as asking for applications, comparisons, analyses, inferences, predictions, evaluation, or creation of information (i.e., higher-order questions corresponding to the *apply*, *analyze*, *evaluate*, and *create* levels of Bloom’s taxonomy). All learners were then provided with a practice text on “Enzymes” (adapted from Meyer, 1975) to practice generating both factual and thinking questions over 10 min.

After completing the question generation training, all learners were then given a question generation test, in which they were asked to generate as many factual and thinking questions as they could over 10 min based on the text that they had studied a week earlier. Learners did not have access to the study text during the question generation test, but were allowed to refer to the question generation training handout as a guide for the kinds of questions that were expected of them. They were further informed that both factual and thinking questions were equally valued on the test. Finally, learners completed a free recall test in which they wrote down as much as they could remember from the study text.

**Results**

***Scoring***

**Question Generation Test.** Learners generated a total of 879 questions, which were classified as lower- or higher-order based on the level of cognitive processing that they demanded corresponding to Bloom’s taxonomy. For instance, questions with answers that simply entailed facts or explanations of concepts presented in the text (i.e., questions that merely required one to *remember* or *understand* previously studied information) were classified as lower-order questions. In contrast, questions that could only be answered by drawing connections between different ideas in the text or by seeking information beyond the text (i.e., questions that required one to *apply*, *analyze*, *evaluate*, or *create* information) were classified as higher-order questions. Questions containing misconceptions in their formulation or that had no relation to the study text were excluded from analyses since they could not be definitively classified as lower- or higher-order.

As in previous question generation studies (e.g., Costa et al., 2000; Harper et al., 2003), two raters first jointly classified a sample of 122 questions as either lower- or higher-order to establish inter-rater agreement and consistency. Subsequently, both raters independently scored another 186 questions (approximately 21% of the total number of questions generated). Cohen’s kappa was .92, indicating high inter-rater reliability. Discrepancies between both raters’ classifications were further reviewed and resolved through discussion to reach 100% agreement. Given the high inter-rater reliability, the remaining questions were scored by a single rater.

**Recall Test.** Learners’ recall performance was scored as the proportion of idea units from the study text that they correctly recalled on the final delayed test. Two raters first jointly scored 14 of the 69 scripts to establish inter-rater agreement, then independently scored another 14 scripts (20%). As inter-rater reliability was high, intraclass correlation (ICC) = .99, 95% CI [.993, .999] based on a two-way random effects model, the remaining scripts were scored by one rater.

***Prior Knowledge***

We ascertained that learners across the retrieval practice (*M* = 2.79, *SD* = 1.52) and restudy (*M* = 2.46, *SD* = 1.28) conditions did not significantly differ in their self-reported mean prior knowledge of the study content, *t*(67) = -0.98, *p* = .33, 95% CI [-1.01, 0.35]. In addition, learners’ prior knowledge did not significantly correlate with their higher-order question generation performance, *r*(67) = -.06, *p* = .60, lower-order question generation performance, *r*(67) = .02, *p* = .84, and recall performance, *r*(67) = -.01, *p* = .93.

***Higher-Order Question Generation Performance***

 A 2 (learning strategy) × 2 (study text) between-subjects ANOVA revealed that learners who practiced retrieval (*M* = 4.91, *SD* = 3.63) did not significantly differ from their peers who restudied (*M* = 4.50, *SD* = 2.87) in the number of higher-order questions that they generated on the final test, *F*(1, 65) = 0.41, *p* = .52, *ŋ*p2 = .01. Thus, retrieval practice did not enhance higher-order question generation performance relative to a restudy control. In addition, learners’ higher-order question generation performance did not significantly differ across the “Human Ear” (*M* = 4.49, *SD* = 2.58) and “Tropisms” (*M* = 4.94, *SD* = 3.87) texts on overall, *F*(1, 65) = 0.46, *p* = .50, *ŋ*p2 = .01. Neither was there a significant interaction between learning strategy and study text, *F*(1, 65) = 0.14, *p* = .71, *ŋ*p2 = .002.

***Lower-Order Question Generation Performance***

In contrast, a 2 (learning strategy) × 2 (study text) between-subjects ANOVA indicated that learners who practiced retrieval (*M* = 7.86, *SD* = 2.86) generated significantly more lower-order questions than their peers who restudied (*M* = 5.79, *SD* = 3.99), *F*(1, 65) = 5.93, *p* = .02, *ŋ*p2 = .08. Learners’ lower-order question generation performance did not significantly differ across the “Human Ear” (*M* = 6.94, *SD* = 3.29) and “Tropisms” (*M* = 6.74, *SD* = 3.93) texts, *F*(1, 65) = 0.05, *p* = .82, *ŋ*p2 = .001. Neither was there a significant interaction between learning strategy and study text, *F*(1, 65) = 0.25, *p* = .62, *ŋ*p2 = .004.

***Final Recall Test Performance***

Analyzing the proportion of idea units that learners recalled on the final delayed test, a 2 (learning strategy) × 2 (study text) between-subjects ANOVA revealed that retrieval practice (*M* = .42, *SD* = .19) yielded significantly better recall performance than restudy (*M* = .22, *SD* = .22), *F*(1, 65) = 22.14, *p* < .001, *ŋ*p2 = .25. Thus, replicating the robust testing effect, retrieval practice benefited long-term retention more than repeated study did. There was also a main effect of study text, whereby learners recalled more idea units from the “Tropisms” text (*M* = .37, *SD* = .25) than the “Human Ear” text (*M* = .28, *SD* = .19), *F*(1, 65) = 7.61, *p* = .01, *ŋ*p2 = .11. Importantly, however, there was no significant interaction between learning strategy and study text, *F*(1, 65) = 0.55, *p* = .46, *ŋ*p2 = .01, indicating that the recall advantage conferred by retrieval practice held across both study texts.

**Discussion**

Relative to repeated study, retrieval practice improved long-term retention and lower-order question generation performance, but did not enhance learners’ ability to generate higher-order questions. Despite enabling learners to recall more information from the study text and even generate more *remember* and *understand* questions related to it, retrieval practice did not enable learners to use their greater recalled content to generate more *apply*, *analyze*, *evaluate*, and *create* questions. One potential account, though, is that learners’ attention may not have been fully oriented toward generating higher-order questions, since they were told that both factual and thinking questions were equally valued on the final test. Hence, Experiment 3b was conducted to address this possibility.

**Supplementary Experiment 3b**

 Experiment 3b was identical to Experiment 3a, except that learners were specifically instructed to generate as many higher-order questions as they could during the question generation test. Thus, the question generation test prompt in Experiment 3b explicitly focused participants’ attention solely on generating higher-order questions, rather than both lower- and higher-order ones.

**Method**

***Participants***

The participants were 51 undergraduates (32 were female) aged between 18 and 26 (*M* = 20.02, *SD* = 1.60). Outcomes below are based on data from 49 participants—two participants who failed to return for the final delayed test were excluded from analyses.

***Design, Materials, and Procedure***

Experiment 3b’s design and materials were identical to those in Experiment 3a—participants were randomly assigned to study a scientific text on “The Human Ear” or “Tropisms” using either retrieval practice or restudy, then returned after 1 week to undergo question generation training, followed by a question generation test and free recall test. The procedure was similar to that in Experiment 3a, with one important exception: Instead of being told that both factual and thinking questions were equally valued, participants were instructed to generate as many thinking questions as they could during the question generation test.

**Results**

***Scoring***

**Question Generation Test.** Participants generated a total of 388 questions in Experiment 3b. As in Experiment 3a, participants’ questions were classified either as lower- or higher-order based on the level of cognitive processing that they demanded corresponding to Bloom’s taxonomy, whereas questions containing misconceptions in their formulation or that had no relation to the study text could not be classified and were thus excluded from analyses. To establish consistency in scoring, two raters first jointly classified a sample of 21 questions, then independently classified another 125 questions (approximately 32% of the total number of questions generated). Inter-rater reliability was high, Cohen’s kappa = .88. Discrepancies between both raters’ classifications were further reviewed and resolved through discussion to reach 100% agreement. Given the high inter-rater reliability, one rater scored the remaining questions.

**Recall Test.** As in Experiment 3a, participants’ recall performance was scored as the proportion of idea units from the study text that they correctly recalled on the delayed final test. Two raters first jointly scored 8 of the 49 scripts to establish inter-rater consistency, then independently scored another 15 scripts (31%). Inter-rater reliability was excellent, ICC = 1.00 based on a two-way random effects model. Thus, the remaining scripts were scored by one rater.

***Prior Knowledge***

We ascertained that learners across the retrieval practice (*M* = 2.02, *SD* = 1.30) and restudy (*M* = 2.50, *SD* = 1.60) conditions did not significantly differ in their self-reported prior knowledge of the study content, *t*(47) = 1.15, *p* = .26, 95% CI [-0.36, 1.32]. As in Experiment 3a, learners’ prior knowledge did not significantly correlate with their higher-order question generation performance, *r*(47) = .25, *p* = .08, lower-order question generation performance, *r*(47) = .07, *p* = .62, and recall performance, *r*(47) = .10, *p* = .49.

***Higher-Order Question Generation Performance***

A 2 (learning strategy) × 2 (study text) between-subjects ANOVA revealed that learners who practiced retrieval (*M* = 4.08, *SD* = 2.52) did not significantly differ from their peers who restudied (*M* = 4.00, *SD* = 2.61) in the number of higher-order questions that they generated, *F*(1, 45) = 0.01, *p* = .91, *ŋ*p2 < .001. Thus, replicating Experiment 3a’s findings, retrieval practice did not enhance higher-order question generation relative to restudy. Learners’ higher-order question generation performance did not significantly differ across the “Human Ear” (*M* = 3.92, *SD* = 2.13) and “Tropisms” (*M* = 4.16, *SD* = 2.93) study texts on overall, *F*(1, 45) = 0.11, *p* = .74, *ŋ*p2 = .002. In addition, there was no significant interaction between learning strategy and study text, *F*(1, 45) = 0.11, *p* = .74, *ŋ*p2 = .002.

***Lower-Order Question Generation Performance***

Although participants’ lower-order question generation performance was not the main focus of Experiment 3b since they were not explicitly instructed to generate such questions, we report the data here for completeness. As in Experiment 3a, a 2 (learning strategy) × 2 (study text) between-subjects ANOVA indicated that learners who practiced retrieval (*M* = 3.42, *SD* = 2.86) generated significantly more lower-order questions than their peers who repeatedly studied (*M* = 1.76, *SD* = 1.99), *F*(1, 45) = 5.40, *p* = .03, *ŋ*p2 = .11. The number of lower-order questions that learners generated did not significantly differ across the “Human Ear” (*M* = 2.54, *SD* = 2.04) and “Tropisms” (*M* = 2.60, *SD* = 3.03) study texts, *F*(1, 45) = 0.02, *p* = .88, *ŋ*p2 < .001. In addition, there was no significant interaction between learning strategy and study text, *F*(1, 45) = 1.06, *p* = .31, *ŋ*p2 = .02.

***Final Recall Test Performance***

A 2 (learning strategy) × 2 (study text) between-subjects ANOVA revealed a significant testing effect, whereby retrieval practice (*M* = .42, *SD* = .23) produced better long-term recall than restudy (*M* = .19, *SD* = .12), as assessed by the proportion of idea units that learners recalled on the final test, *F*(1, 45) = 22.44, *p* < .001, *ŋ*p2 = .33. There was a significant main effect of study text, whereby learners recalled a greater proportion of idea units for the “Tropisms” text (*M* = .36, *SD* = .23) than “The Human Ear” text (*M* = .24, *SD* = .18), *F*(1, 45) = 6.72, *p* = .01, *ŋ*p2 = .13. Nevertheless, there was no significant learning strategy × study text interaction, *F*(1, 45) = 3.20, *p* = .08, *ŋ*p2 = .07, indicating that the recall advantage of retrieval practice persisted across both study texts. In view that study text topic consistently did not interact with learning strategy in predicting learners’ question generation and recall test performance across both Experiments 3a and 3b, we did not analyze the effects of study text any further in the subsequent experiment.

**Discussion**

 Despite now being specifically instructed to focus on generating higher-order questions, learners who practiced retrieval still fared no better than those who restudied. Thus, replicating Experiment 3a’s findings, there was again no evidence that retrieval practice confers any advantage for higher-order question generation, even if it improved learners’ long-term recall performance and their ability to generate lower-order questions based on the text.

These findings converge with those of some extant studies showing that retrieval practice alone may not boost some higher-order learning outcomes such as inferencing (McDaniel et al., 2009; Nguyen & McDaniel, 2016) and integrative argumentation (Wong & Lim, 2019), even if it enhances basic knowledge retention. In particular, whereas retrieval practice may encourage textbase processing by alerting learners to gaps in their comprehension, recalling information alone may not orient learners toward global-level situation model processing in constructing a more cohesive representation of the text, which is crucial for higher-order learning outcomes that demand more than mere recall of information (Kintsch, 1988; Nguyen & McDaniel, 2016; Wong & Lim, 2019).

Rather, the efficacy of retrieval practice for higher-order learning can potentially be boosted by supplementing it with a metacomprehension monitoring intervention that explicitly guides learners to consider the intended higher-order educational outcome at stake and whether they are adequately prepared to meet it. For instance, Nguyen and McDaniel (2016) found that augmenting retrieval practice with metacognitive judgments of inference (JOI)—having learners judge how well they can apply their knowledge—during study improved learners’ inferencing performance more than retrieval practice alone. Interestingly, this advantage was specific to supplementing retrieval practice with JOI but not metacognitive judgments of learning (JOL), which involve asking learners to judge how well they have learned the material or how well they have remembered the information. Presumably, JOI may more effectively orient learners’ attention toward the key higher-order learning outcome to stimulate greater global-level situation model processing during restudy, whereas JOL may be inadequate in pushing learners toward such processing (Nguyen & McDaniel, 2016; see also Wong & Lim, 2019). Accordingly, Experiment 3c tested whether retrieval practice may enhance higher-order question generation when it is supplemented with a metacomprehension monitoring intervention that explicitly orients learners toward the kinds of processing required for this complex learning outcome.

**Supplementary Experiment 3c**

Following previous studies (Nguyen & McDaniel, 2016; Wong & Lim, 2019), we developed and investigated two types of metacomprehension monitoring interventions in Experiment 3c: metacognitive judgments of learning (JOL) versus metacognitive judgments of higher-order learning (JOL+). Whereas the JOL questions related to how well one has learned the text in remembering or understanding it, the JOL+ questions related to the higher-order educational outcome of generating higher-order questions based on the text. All participants responded to either the JOL or JOL+ questions at the halfway mark of the study phase (i.e., Study–Study–JOL/JOL+–Study–Study in the restudy condition versus Study–Retrieve–JOL/JOL+–Study–Retrieve in the retrieval practice condition).

 We investigated the extent that supplementing retrieval practice with JOL+ would yield better higher-order question generation than a restudy condition that had similarly been boosted with JOL+. Conversely, we expected that augmenting retrieval practice with JOL would not necessarily enhance higher-order question generation relative to restudy with JOL, since making JOL does not explicitly orient learners’ attention toward the critical global-level situation model processes required for effective higher-order question generation (e.g., Nguyen & McDaniel, 2016; Wong & Lim, 2019).

**Method**

***Participants***

The participants were 50 undergraduates (40 were female) aged between 18 and 22 (*M* = 19.50, *SD* = 1.02).

***Design***

The two between-subjects factors of interest were: learning strategy (*retrieval practice* vs. *restudy*) and metacomprehension judgment (*JOL* vs. *JOL+*). The dependent variables were the same as those in the earlier supplementary experiments: (a) learners’ higher-order question generation performance, (b) lower-order question generation performance, and (c) recall performance. All tests were administered after a 1-week delay following initial study, as in the previous experiments.

***Materials***

The materials were identical to those in Experiments 3a and 3b, except that we further developed and introduced a series of JOL and JOL+ questions (adapted from Nguyen & McDaniel, 2016; Wong & Lim, 2019) that were intended to guide all learners’ metacomprehension monitoring during their study. Corresponding to the “lower-order” (i.e., *remember* and *understand*) levels of Bloom’s taxonomy, the JOL questions asked learners to estimate how well they had learned the text on a 7-point scale, including: (a) “How well do you think you understood the text?” (1 = *definitely not well*; 7 = *definitely well*), (b) “How likely are you to remember this text on a future test?” (1 = *definitely will not*; 7 = *definitely will*), and (c) “How much do you think you have learned about the topic presented in the text?” (1 = *nothing at all*; 7 = *everything that the text presented*).

In contrast, the JOL+ questions related to the higher-order learning outcome of generating questions that fulfilled the *apply*, *analyze*, *evaluate*, and *create* levels of Bloom’s taxonomy, including: (a) “How difficult do you think it is to ask conceptual questions about the text?” (1 = *definitely not difficult*; 7 = *definitely difficult*), (b) “How ready are you to ask meaningful questions about the text?” (1 = *definitely not ready*; 7 = *definitely ready*); (c) “How confident are you in generating high-quality questions which require answers that involve: (i) applying the knowledge and facts you have acquired in other contexts/in a different way (e.g., hypothesizing how the concept raised might be used differently), (ii) examining and breaking down the material you have learned into its constituent parts (e.g., identifying cause and effect), (iii) making inferences or generalizations, possibly using your pre-existing knowledge (e.g., comparing and contrasting concepts), (iv) making judgments about information, validity of ideas or quality of work based on a set of criteria (e.g., appraising whether the ideas are right or wrong), (v) creating new knowledge, ideas, or perspectives (e.g., combining new information with past experiences to propose new solutions)?” (each item rated 1 = *not at all*; 7 = *definitely*).

***Procedure***

 The procedure was the same as that in Experiment 3b, except that after their first review opportunity during the study phase, learners additionally responded to either the JOL or JOL+ questions, depending on their randomly assigned metacomprehension judgment condition. That is, participants in the restudy (SSSS) condition studied their randomly assigned text (either “The Human Ear” or “Tropisms”) for 6 min, restudied it for 6 min, responded to either the JOL or JOL+ questions, then restudied the text for two consecutive 6-min periods. Conversely, participants in the retrieval practice (SRSR) condition studied their randomly assigned text for 6 min, practiced retrieval for 6 min by writing down as much as they could remember from the text, responded to either the JOL or JOL+ questions, then restudied the text for 6 min, and practiced retrieval again for 6 min.

One week later, all learners returned to be trained on question generation and to complete a question generation test—as in Experiment 3b, they were told to generate as many thinking questions (i.e., higher-order questions) as they could. Finally, all learners completed a free recall test.

**Results**

***Scoring***

**Question Generation Test.** The scoring procedure was identical to that in Experiments 3a and 3b. Participants generated a total of 400 questions in Experiment 3c. To establish consistency in scoring, two raters first jointly classified a sample of 73 questions as either lower- or higher-order. Then, both raters independently classified another 57 questions (approximately 14% of the total number of questions generated). Cohen’s kappa was .93, indicating high inter-rater reliability. Discrepancies between the two raters’ classifications were reviewed and resolved through discussion to reach 100% agreement. Given the high inter-rater reliability, one rater scored the remaining questions.

**Recall Test.** Participants’ final recall test responses were scored in the same way as in Experiments 3a and 3b. Recall data from 14 of the 50 scripts were first jointly scored by two raters to establish inter-rater consistency. Then, both raters independently scored the recall data from a further 14 scripts (28%). Inter-rater reliability was high, ICC = .99, 95% CI [.97, .99] based on a two-way random effects model. Thus, a single rater proceeded to score the remaining free recall test responses.

***Prior Knowledge***

We ascertained that learners in the retrieval practice (*M* = 2.76, *SD* = 1.44) and restudy (*M* = 2.32, *SD* = 1.46) conditions did not significantly differ in their self-reported prior knowledge of the study content, *t*(48) = -1.08, *p* = .29, 95% CI [-1.26, 0.38]. In addition, learners’ prior knowledge ratings did not significantly correlate with their higher-order question generation performance, *r*(48) = .07, *p* = .62, and lower-order question generation performance, *r*(48) = -.02, *p* = .90, although learners with higher prior knowledge tended to recall a greater proportion of idea units from the study text on the final recall test, *r*(48) = .30, *p* = .03.

***Higher-Order Question Generation Performance***

Replicating the previous experiments’ findings, a 2 (learning strategy) × 2 (metacomprehension judgment) between-subjects ANOVA yielded no main effect of learning strategy on higher-order question generation performance, *F*(1, 46) = 0.05, *p* = .83, *ŋ*p2 = .001. Thus, learners who practiced retrieval (*M* = 6.08, *SD* = 3.08) did not significantly differ from their peers who restudied (*M* = 5.88, *SD* = 3.31) in the number of higher-order questions that they generated. There was no main effect of metacomprehension judgment, *F*(1, 46) = 1.27, *p* = .27, *ŋ*p2 = .03. Thus, learners’ higher-order question generation performance on overall did not differ whether they responded to the JOL (*M* = 5.38, *SD* = 2.73) versus JOL+ (*M* = 6.41, *SD* = 3.43) questions. Inconsistent with predictions, there was also no significant learning strategy × metacomprehension judgment interaction, *F*(1, 46) = 0.07, *p* = .80, *ŋ*p2 = .001. Thus, retrieval practice failed to confer an advantage over restudy for higher-order question generation, even when it had been supplemented with JOL+ than JOL.

***Lower-Order Question Generation Performance***

Similarly, a 2 (learning strategy) × 2 (metacomprehension judgment) between-subjects ANOVA yielded no main effect of learning strategy on learners’ lower-order question generation performance, *F*(1, 46) = 1.16, *p* = .29, *ŋ*p2 = .03. Retrieval practice participants (*M* = 2.08, *SD* = 2.58) did not significantly differ from restudy participants (*M* = 1.40, *SD* = 1.78) in the number of lower-order questions that they generated. There was no main effect of metacomprehension judgment, *F*(1, 46) = 2.07, *p* = .16, *ŋ*p2 = .04, indicating that learners’ lower-order question generation performance on overall did not differ whether they responded to the JOL (*M* = 2.29, *SD* = 2.87) versus JOL+ (*M* = 1.34, *SD* = 1.54) questions. There was also no significant learning strategy × metacomprehension judgment interaction, *F*(1, 46) = 0.15, *p* = .70, *ŋ*p2 = .003. Thus, the lack of advantage from practicing retrieval than restudying for lower-order question generation persisted across both types of metacomprehension judgments—JOL and JOL+.

***Final Recall Test Performance***

A 2 (learning strategy) × 2 (metacomprehension judgment) between-subjects ANOVA on learners’ final recall test performance yielded only a significant main effect of learning strategy, *F*(1, 46) = 22.39, *p* < .001, *ŋ*p2 = .33. Replicating the robust testing effect, learners who practiced retrieval (*M* = .55, *SD* = .16) recalled a greater proportion of idea units from the study text on a final test after a 1-week delay, relative to their peers who restudied (*M* = .32, *SD* = .18). There was no main effect of metcaomprehension judgment, *F*(1, 46) = 0.04, *p* = .85, *ŋ*p2 = .001, indicating that learners’ long-term retention did not differ across the JOL (*M* = .44, *SD* = .21) and JOL+ (*M* = .42, *SD* = .20) conditions. There was also no significant learning strategy × metacomprehension judgment interaction, *F*(1, 46) = 0.03, *p* = .88, *ŋ*p2 = .001.

**Discussion**

Altogether, our findings across Experiments 3a to 3c suggest that there is no evidence that retrieval practice enhances higher-order question generation, as compared to a restudy control. Indeed, retrieval practice did not produce better higher-order question generation performance even when learners had been explicitly instructed to focus solely on generating such questions at test (Experiments 3b and 3c), and even when retrieval practice had been further boosted with JOL+ as a metacomprehension monitoring intervention that oriented learners’ attention toward the kinds of global-level situation model processing required for effective higher-order question generation (Experiment 3c). Hence, although retrieval practice promoted learners’ durable retention of the studied information and even their generation of lower-order *remember* and *understand* questions in some instances (Experiments 3a and 3b), it did not enable learners to effectively use that information to generate more higher-order *apply*, *analyze*, *evaluate*, and *create* questions than mere restudying.

**Supplementary References**

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. <https://doi.org/10.3102/0034654316689306>

Cook, L. K., & Mayer, R. E. (1988). Teaching readers about the structure of scientific text. *Journal of Educational Psychology*, *80*(4), 448–456. <https://doi.org/10.1037/0022-0663.80.4.448>

Costa, J., Caldeira, H., Gallástegui, J. R., & Otero, J. (2000). An analysis of question asking on scientific texts explaining natural phenomena. *Journal of Research in Science Teaching*, *37*(6), 602–614. [https://doi.org/10.1002/1098-2736(200008)37:6<602::AID-TEA6>3.0.CO;2-N](https://doi.org/10.1002/1098-2736%28200008%2937%3A6%3C602%3A%3AAID-TEA6%3E3.0.CO;2-N)

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. <https://doi.org/10.3758/BF03193146>

Griffin, T. D., Wiley, J., & Thiede, K. W. (2019). The effects of comprehension-test expectancies on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(6), 1066–1092. <https://doi.org/10.1037/xlm0000634>

Harper, K. A., Etkina, E., & Lin Y. (2003). Encouraging and analyzing student questions in a large physics course: Meaningful patterns for instructors. *Journal of Research in Science Teaching*, *40*(8), 776–791. <https://doi.org/10.1002/tea.10111>

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. T. Wixted (Ed.), Cognitive psychology of memory, Vol. 2 of Learning and memory: A comprehensive reference (J. H. Byrne, Series Ed.) (pp. 487–514). Academic Press. <https://doi.org/10.1016/B978-0-12-809324-5.21055-9>

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. <https://doi.org/10.1126/science.1199327>

King, A. (1989). Effects of self-questioning training on college students’ comprehension of lectures. *Contemporary Educational Psychology*, *14*(4), 366–381. [https://doi.org/10.1016/0361-476X(89)90022-2](https://doi.org/10.1016/0361-476X%2889%2990022-2)

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. <https://doi.org/10.1037/0033-295X.95.2.163>

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

Meyer, B. J. (1975). *The organization of prose and its effects on memory*. North-Holland.

Nguyen, K., & McDaniel, M. A. (2016). The JOIs of text comprehension: Supplementing retrieval practice to enhance inference performance. *Journal of Experimental Psychology: Applied*, *22*(1), 59–71. <https://doi.org/10.1037/xap0000066>

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>

Wong, S. S. H., & Lim, S. W. H. (2019). From JOLs to JOLs+: Directing learners’ attention in retrieval practice to boost integrative argumentation. *Journal of Experimental Psychology: Applied*, *25*(4), 543–557. <https://doi.org/10.1037/xap0000225>