Journal of Educational Psychology

ISSN: 0022-0663

2023, Vol. 115, No. 6, 798–812
https://doi.org/10.1037/edu0000802

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

Sarah Shi Hui Wong, Kagen Y. L. Lim, and Stephen Wee Hun Lim
Department of Psychology, National University of Singapore

Asking good questions is vital for scientific learning and discovery, but improving this complex skill is a formidable challenge. Here, we show in two experiments ($N = 152$) that teaching others—*learning-by-teaching*—enhances one's ability to generate higher-order research questions that create new knowledge, relative to two other well-established generative learning techniques: retrieval practice and concept-mapping. Learners who taught scientific expository texts across natural and social sciences topics by delivering video-recorded lectures outperformed their peers who practiced retrieval or constructed concept maps when tested on their ability to generate *create*-level research questions based on the texts (Experiment 1). This advantage held reliably even on a delayed test 48 hr later, and when all learners similarly received and responded to poststudy questions on the material (Experiment 2). Moreover, across both immediate and delayed tests, learning-by-teaching produced a recall benefit that rivaled that of the potent technique of retrieval practice. In contrast, despite recalling more than twice the study content that the concept-mapping group did, learners who practiced retrieval were unable to generate more *create*-level research questions based on that content. Three supplemental experiments ($N = 168$) further showed that retrieval practice consistently did not improve higher-order question generation over restudying, despite yielding superior long-term retention. Altogether, these findings reveal that simply possessing a wealth of factual knowledge is insufficient for generating higher-order research questions that create new knowledge. Rather, teaching others is a powerful strategy for producing deep and durable learning that enables research question generation. To ask better questions, teach.

Educational Impact and Implications Statement

Scientific discovery often begins with the art of asking good questions. Here, we show that teaching others enhances students' ability to generate higher-order research questions that create new knowledge. Across immediate and delayed tests, students who taught scientific material by delivering a video-recorded lecture successfully generated more *create*-level research questions based on the material, as compared to their peers who used well-established learning methods such as retrieval practice and concept-mapping. While we teach, we learn to ask better research questions.

Keywords: question generation, generative learning, learning by teaching, retrieval practice, concept-mapping

Supplemental materials: https://doi.org/10.1037/edu0000802.supp

This article was published Online First May 25, 2023.
Sarah Shi Hui Wong https://orcid.org/0000-0003-4243-212X
Kagen Y. L. Lim https://orcid.org/0000-0003-4868-1127
Stephen Wee Hun Lim https://orcid.org/0000-0003-3636-7587

We are grateful to the following individuals for their invaluable assistance with data collection and/or scoring: Jia Jie Chua, Ryan Qi Xian Chua, Gabriel Rongyang Lau, Jeremy Yang Jing Tan, and Winston Wen Jie Tan.

This research was supported by a National University of Singapore Educational Research grant (C-581-000-222-091) awarded to Sarah Shi Hui Wong, alongside a Fulbright Singapore Researcher Fellowship and a National University of Singapore Faculty of Arts and Social Sciences Heads and Deanery Research Support Scheme grant (R-581-000-150-133) awarded to Stephen Wee Hun Lim.

All data and experimental materials are available in the online supplemental materials.

Sarah Shi Hui Wong served as lead for data curation, formal analysis, writing–original draft, and writing–review and editing and contributed equally to conceptualization, methodology, and resources. Kagen Y. L. Lim served as lead for investigation and served in a supporting role for methodology and data curation. Stephen Wee Hun Lim served as lead for conceptualization, methodology, resources, and supervision and served in a supporting role for writing–review and editing.

Correspondence concerning this article should be addressed to Stephen Wee Hun Lim, Department of Psychology, Faculty of Arts & Social Sciences, National University of Singapore, Block AS4, 9 Arts Link, Singapore 117570, Singapore. Email: psylimwh@nus.edu.sg

> It is easier to judge the mind of a man by his questions rather than his answers.
>
> —Pierre-Marc-Gaston de Lévis, *Maximes et Réflexions sur Différents Sujets de Morale et de Politique*

To equip learners to tackle wicked real-world problems that are complex, ill-structured, and dynamic (Rittel & Webber, 1973), educators have applied inquiry- or problem-based approaches that guide learners' knowledge construction when reasoning about a problem in successive iterations (e.g., Barrows & Tamblyn, 1980; Hmelo-Silver, 2004; Pedaste et al., 2015). This inquiry process is critically driven and catalyzed by the questions that learners ask when making sense of the problem (e.g., in articulating the problem space and constraints, identifying assumptions and knowledge gaps to be resolved), toward generating hypotheses and exploring solutions (for discussions, see Agee, 2009; Alvesson & Sandberg, 2011; Tawfik et al., 2020). Clearly, the ability to ask good research questions is crucial for knowledge construction that, in turn, triggers further scientific inquiry.

As set forth in the National Research Council's (2013) *Next Generation Science Standards*, proficiency in asking research questions is integral to scientific literacy. Besides stimulating inquiry, questioning is a form of meaningful learning that piques students' curiosity and interest in the subject matter, while diagnosing their conceptual understanding and higher-order thinking (Chin & Brown, 2002). Thus, improving students' ability to formulate good research questions has been of keen interest to researchers and educators across the natural and social sciences (Chin & Osborne, 2008; White, 2017). Here, we investigated the extent that *teaching* others is a powerful way to enhance this valued educational outcome.

What Are Good Research Questions?

Questioning taxonomies and hierarchies have often categorized research questions as the peak of students' question types since they involve complex thinking skills such as integrating knowledge in new ways, relative to less sophisticated questions that are factual or have readily available answers (Keeling et al., 2009; Marbach-Ad & Sokolove, 2000). For instance, questions that specify contingencies or cause-and-effect relations among various phenomena, as opposed to simply describing their properties or comparing them, have been classified as higher-order questions that characterize expert-like reasoning (Tawfik et al., 2020), representing the kind of knowledge that scientific inquiry ultimately aspires toward (Dillon, 1984; see also Allison & Shrigley, 1986; Cuccio-Schirripa & Steiner, 2000; Hartford & Good, 1982).

Indeed, good research questions are not only grounded in extant knowledge, but are further aimed at *creating* new knowledge, thereby contributing to theoretical and practical innovation. In educational contexts, Bloom's classic taxonomy (Anderson et al., 2001; Bloom, 1956) has often been used to differentiate such higher-order questions from lower-order ones (Agarwal, 2019; Renaud & Murray, 2007), with students' question quality correlating positively with measures of their academic achievement (Graesser & Person, 1994; Harper et al., 2003; Person et al., 1994). Specifically, Bloom's taxonomy outlines six distinct categories of cognitive processes in increasing complexity, ranging from the *remember* and *understand* categories that are associated with "lower-order" learning requiring memory and comprehension, to the *apply*, *analyze*, *evaluate*, and *create* categories that constitute "higher-order" learning (see Table 1). Notably, good research questions can be viewed as questions reflecting the *create* category—the pinnacle or "holy grail" of Bloom's taxonomy, whereby one combines or reorganizes elements to form a new structure by generating hypotheses (Anderson et al., 2001).

Whereas enhancing students' research question generation is a vital educational goal, attaining it remains a formidable challenge. Notwithstanding low rates of student questioning in classrooms (Dillon, 1988; Graesser & Person, 1994; Newman & Goldin, 1990), even when students do ask questions, the majority of these tend to be lower-order ones (Dillon, 1988; Keeling et al., 2009). How can we boost students' ability to generate higher-order research questions? In view that formulating such questions requires building connections among various elements of to-be-learned material and integrating them to create new knowledge, learning strategies that promote such generative processes may offer a promising solution.

Generative Learning

Grounded in the constructivist view of learning (Steffe & Gale, 1995), generative learning involves actively constructing meaning by integrating incoming information with one's prior knowledge and experiences (Osborne & Wittrock, 1983; Wittrock, 1974; see also Chi, 2009). For instance, Mayer's (1984, 1996, 2014) select–organize–integrate model posits that meaningful learning draws on three cognitive processes: selecting relevant information, organizing the selected information into a coherent mental representation, and integrating the newly constructed representation with existing knowledge structures. Accordingly, generative learning strategies are those that encourage learners to meaningfully make sense of to-be-learned information through engaging in such cognitive processes (Fiorella & Mayer, 2015, 2016). Of particular interest, one such generative strategy is *learning-by-teaching*.

Learning-by-Teaching

A growing body of research has shown that teaching others enhances one's own learning of the taught material (e.g., Bargh & Schul, 1980; Duran & Topping, 2017; Fiorella & Mayer, 2013; Roscoe & Chi, 2007; for recent meta-analyses, see Kobayashi, 2019; Lachner et al., 2021; Ribosa & Duran, 2022). Learning-by-teaching has often been implemented via peer tutoring in classrooms or synchronous online learning environments (e.g., Roscoe & Chi, 2007, 2008; for meta-analyses, see Bowman-Perrott et al., 2013; Cohen et al., 1982; Leung, 2019), or via computer-based teachable agents in educational software (e.g., Biswas et al., 2005; Chin et al., 2010). However, teaching also benefits the tutor's learning when their audience is imaginary rather than physically present or remote (Lachner et al., 2022), as when delivering video-recorded lectures to fictitious others (e.g., Fiorella & Mayer, 2013; Hoogerheide et al., 2014, 2016, 2019) or even writing a verbatim teaching script (Lim et al., 2021).

Why is teaching beneficial for the tutor's own learning? To date, three main nonmutually exclusive accounts have been established: (a) the retrieval hypothesis, (b) the generative hypothesis, and (c) the social presence hypothesis (for a review, see Lachner et al., 2022). According to the *retrieval hypothesis*, teaching from memory involves substantive retrieval of the material, thereby inducing

Table 1
Question Levels Based on Bloom's Taxonomy

Level	Category	Associated cognitive processes	Sample action prompts
1	Remember	Answer requires *recall/remembering* of terminology, specific facts, definitions, and basic concepts covered in the text	Identify, recognize, indicate, list, name specific events, locations, people, dates, sources of information (e.g., Who? What? Where? When? Which?)
2	Understand	Answer requires *basic understanding* (i.e., descriptions, explanations, examples) of concepts in the text	Describe, explain, give examples of, summarize, generalize
3	Apply	Answer requires *using/applying* acquired knowledge, facts, and concepts in a new situation or in a different way	Predict, give other examples in other contexts, seek exceptions
4	Analyze	Answer requires *examining and breaking down information* into constituent parts by identifying motives/causes, making inferences and finding evidence to support generalizations, or seeking causes and/or consequences	Compare, contrast, differentiate, organize, deconstruct
5	Evaluate	Answer requires *making judgments* about information, validity of ideas, or quality of work based on a set of criteria	Appraise, assess how effective/optimal or which is most important/valuable, check for discrepancies/inconsistencies in information
6	Create	Answer requires *creating new knowledge, ideas, or perspectives* by compiling information in a different way, combining elements in a new pattern, or proposing alternative solutions	Adapt, produce alternative hypotheses or solutions

testing effects that improve the tutor's learning (Koh et al., 2018). Second, the *generative hypothesis* suggests that teaching encourages generative processes that boost the tutor's learning when selecting, organizing, and integrating new information with one's knowledge structures (Fiorella & Mayer, 2016; Roscoe & Chi, 2008), while monitoring one's own understanding (Lachner et al., 2020; Muis et al., 2016). Third, the *social presence hypothesis* posits that teaching an audience induces social presence—an awareness of others and viewing them as "real" (Gunawardena, 1995; Short et al., 1976)—that triggers greater generative processing for better learning (Hoogerheide et al., 2016, 2019; Jacob et al., 2020; Lachner et al., 2021). In principle, a combination of any of these teaching-related processes may promote deeper knowledge inquiry and (re) construction.

The teaching process entails three stages that each uniquely contributes to the learning benefits of teaching: expecting to teach, actually teaching, and responding to tutee questions (Fiorella & Mayer, 2013, 2014, 2015; Kobayashi, 2019; Nestojko et al., 2014; Roscoe & Chi, 2007, 2008). For instance, expecting to teach may motivate students to select relevant information and organize it in a coherent mental structure during their teaching preparation in anticipation of their tutees' needs, thereby facilitating deep learning (Bargh & Schul, 1980; Benware & Deci, 1984). Subsequently, actually teaching—for example, by explaining the material to a target audience via a video-recorded lecture—encourages the tutor's reflective knowledge-building when generating inferences during their explanations, thus stimulating the construction and integration of new ideas with their prior knowledge (Roscoe & Chi, 2007, 2008). Responding to tutee questions may then prompt the tutor's self-monitoring of their comprehension as they detect and remedy any gaps in their understanding, while promoting further knowledge-building and creation of previously unconceived connections when elaborating in more detail or clarifying content in different ways (Kobayashi, 2018; Roscoe, 2014; Roscoe & Chi, 2008; see also Lachner et al., 2020).

Together, these generative processes may enable the tutor to build rich mental models of the taught information that aid meaningful learning (Coleman et al., 1997). According to Kintsch's (1988, 1994) construction–integration model, textual information can be processed at the *textbase* level (i.e., mentally representing propositional content as explicitly stated in the text) and *situation model* level (i.e., a global representation of the text's meaning by creating a mental model of the implicit causal relations between propositions). Whereas textbase level processing may suffice for recalling a text, it is often insufficient for deeper understanding. Rather, learners must construct a situation model of the text by elaborating on and integrating it with their relevant prior knowledge, such as by going beyond the text to make inferences about its implicit relations. By prompting the tutor to select, organize, and integrate information, learning-by-teaching may thus facilitate the construction of a cohesive global representation of the text as integrated with one's prior knowledge, in turn benefiting higher-order learning that demands an elaborate situation model (Coleman et al., 1997; Guerrero & Wiley, 2021).

Indeed, learning-by-teaching has proven helpful for a variety of educational outcomes across immediate and delayed tests (Kobayashi, 2019), attesting to the efficacy of this technique for producing learning that is both meaningful and durable. For instance, teaching has been found to enhance the tutor's memory for and comprehension of scientific expository texts, as assessed via recall tests and inference questions that require making connections among ideas in the text (Fiorella & Mayer, 2013, 2014; Guerrero & Wiley, 2021; Nestojko et al., 2014). In addition, learning-by-teaching benefits higher-order transfer in applying the studied information to new problems (Coleman et al., 1997; Hoogerheide et al., 2014, 2016).

To date, however, the effects of learning-by-teaching on the complex educational outcome of research question generation have not been explored. Yet, the deep learning that teaching promotes may boost the tutor's ability to formulate higher-order research questions that create new knowledge based on the taught material. The present research investigated this possibility.

The Present Study

Here, we tested the benefits of learning-by-teaching for research question generation and further examined whether these benefits

persist over a delay after the initial learning session. The "gold standard" of educational innovation is to compare a novel intervention against existing practice (Roediger & Pyc, 2012). Thus, we compared learning-by-teaching against two other well-established generative learning techniques in contemporary education: *retrieval practice* and *concept-mapping*.

Over the past decades, an explosion of research on retrieval practice—the act of testing oneself from memory—has robustly shown that it is a potent technique for enhancing durable and meaningful learning (Roediger & Karpicke, 2006a, 2006b; for recent reviews, see Adesope et al., 2017; Agarwal et al., 2021; Carpenter et al., 2022; Karpicke, 2017; Yang et al., 2021). Besides reducing mind-wandering (Wong & Lim, 2022a), the generative learning strategy of retrieval practice prompts learners to selectively activate and retrieve relevant knowledge, organize it by strengthening connections among learned ideas, and integrate the learned information with their prior knowledge by building new connections (Fiorella & Mayer, 2015, 2016). For instance, the elaborative retrieval account (Carpenter, 2009, 2011) posits that retrieval activates cue-related semantic information, which may become bound with the target information to yield a more elaborated memory trace that aids future recall. Although some studies have questioned the benefits of retrieval practice for some complex learning outcomes such as inferencing (McDaniel et al., 2009; Nguyen & McDaniel, 2016) and integrative argumentation (Wong & Lim, 2019a), other studies have reported that retrieval practice improves not only recall (see Rowland, 2014 for a meta-analysis) but also transfer of learning (e.g., Butler, 2010; Wong et al., 2019; for reviews, see Adesope et al., 2017; Carpenter, 2012; Pan & Rickard, 2018). Indeed, retrieval practice has been hailed as a high-utility learning technique with broad applicability across diverse learning materials, outcome measures, retention intervals, and learner characteristics, relative to other techniques that students commonly adopt such as highlighting, summarizing, and rereading (Dunlosky et al., 2013).

Likewise, concept-mapping is a generative learning strategy that involves actively making sense of incoming information (Fiorella & Mayer, 2015, 2016; Karpicke & Blunt, 2011b). In concept-mapping, learners graphically organize to-be-learned material by selecting key concepts to be represented as nodes, while organizing them into a coherent structure using links that represent their relations, and integrating the information with prior knowledge by determining the overall hierarchical arrangement of concepts (Novak & Gowin, 1984). This technique is widely used in diverse educational settings across science, technology, engineering, and mathematics (STEM) and non-STEM subjects alike. Moreover, concept-mapping has been found to be effective for knowledge retention and transfer, relative to other instructional conditions such as participating in lectures or discussions, and constructing lists or outlines (Chularut & DeBacker, 2004; Nesbit & Adesope, 2006; Schroeder et al., 2018). Thus, both retrieval practice and concept-mapping served as strong contenders against learning-by-teaching in enhancing the higher-order outcome of generating research questions.[1]

In two experiments, all learners were first instructed on generating *create*-level research questions based on the same training procedures. After which, they received a scientific expository text on either a natural or social science topic ("food allergies" or "intelligence quotient") and were randomly assigned to study it using one of three learning methods: either (a) constructing a concept map that graphically organized the text's ideas, (b) practicing retrieval of the text via a free recall procedure, or (c) teaching the text by preparing teaching notes, delivering a video-recorded lecture to a fictitious audience with reference to one's notes, and responding to "tutee" questions. All learners were then similarly tested on their ability to generate *create*-level research questions based on the studied text, as well as their recall of the text content. In Experiment 1, the final tests were administered immediately after the initial study session. Experiment 2 aimed to replicate Experiment 1's effects, while probing whether they held durably on a delayed test 48 hr later. Furthermore, to ascertain that any learning benefits of teaching did not stem from answering "tutee" questions per se, learners in Experiment 2's concept-mapping and retrieval practice conditions similarly received and answered questions about the study material.

Experiment 1

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study, and we follow the APA *Journal Article Reporting Standards*. All data and experimental materials are available in the online supplemental materials. Data were analyzed using SPSS version 26. This study's design and analyses were not preregistered.

Participants

Seventy-eight undergraduate students (50 were female) aged between 19 and 25 ($M = 20.63$, $SD = 1.62$) from the National University of Singapore participated in this study. There were 26 participants each in the concept-mapping, retrieval practice, and learning-by-teaching groups. The three learning groups did not significantly differ in their mean age or proportion of men and women. Previous studies that directly compared retrieval practice versus concept-mapping (Karpicke & Blunt, 2011b; O'Day & Karpicke, 2021) reported effect sizes ranging from $d = 0.85$ to 1.54. Based on the most conservative effect size, a power analysis (G*Power; Faul et al., 2007) indicated that at least 23 participants per condition were required to observe a retrieval-based learning effect for two-tailed between-subjects pairwise comparisons at 80% power and $\alpha = .05$. The present sample size also afforded sufficient sensitivity to detect effects of $d \geq 0.79$ for two-tailed between-subjects pairwise comparisons at 80% power and $\alpha = .05$, similar to the median effect size of learning-by-teaching ($d = 0.77$) reported by Fiorella and Mayer (2015). All experiments were conducted with ethics approval from our university's Institutional Review Board. Participants received either course credit or cash remuneration for their participation and provided their informed consent.

[1] We did not include a nongenerative learning control condition (e.g., restudying) because the primary aim of the present study was to compare learning-by-teaching against prevailing "best practice" strategies that are known to be beneficial, rather than "business-as-usual" strategies that are known to be less effective. Indeed, much prior research has robustly shown that generative learning techniques such as learning-by-teaching and retrieval practice produce greater gains than mere restudying (e.g., Adesope et al., 2017; Ribosa & Duran, 2022; Rowland, 2014).

Design

Experiment 1 used a between-subjects design with learning strategy as the key independent variable, whereby participants were randomly assigned to either the *concept-mapping*, *retrieval practice*, or *learning-by-teaching* condition. To ascertain that any effects of the learning strategies generalized across knowledge domains in the natural versus social sciences, we included study text as a second independent variable for control purposes, whereby participants were randomly assigned to study a text on either "food allergies" or "intelligence quotient."

The two main outcomes of interest were: (a) learners' research question generation performance, as assessed via the number of questions they posed that fulfilled the *create* level of Bloom's taxonomy, and (b) learners' recall performance, as assessed via the number of idea units from the study texts that they correctly recalled on an immediate test.

Materials

Question Generation Training. To ensure that all participants understood what was required of them in generating research questions (i.e., *create* questions), they were trained on all question levels corresponding to Bloom's taxonomy. This procedure was intended to facilitate learners' holistic understanding of the various question types and to guide them in differentiating among questions that constituted research questions versus those that did not. Participants received a printed handout (see Table 1) that introduced and explained the features of *remember*, *understand*, *apply*, *analyze*, *evaluate*, versus *create* questions, alongside sample action prompts associated with each question level. Participants were also given a 199-word practice text on "enzymes" (adapted from Meyer, 1975; available in the online supplemental materials) from which they practiced generating questions. The "enzymes" practice text did not relate to either of the critical study texts.

Study Texts. The critical study texts were two scientific expository texts on "food allergies" and "intelligence quotient" (adapted from Griffin et al., 2019; available in the online supplemental materials), with Flesch-Kincaid grade levels of 12.4 and 12.3, respectively. Both texts contained four paragraphs with 20 sentences and 310 words each. For scoring purposes, we identified 40 idea units in each text. A sample idea unit in the "food allergies" text was: "The allergic reaction to food particles can be manifested as skin rashes," whereas a sample idea unit in the "intelligence quotient" text was: "Even identical twins differ in IQ."

Prelearning Questionnaire. As prior knowledge has been associated with the generation of higher-quality questions (Harper et al., 2003; Taboada & Guthrie, 2006), we ascertained that the learning groups did not differ in their prior knowledge of the study texts. Before reading the texts, learners reported how much prior knowledge they had about their respective topic (1 = *not at all*; 5 = *a lot*), and indicated whether or not they possessed prior knowledge of eight specific content items related to each topic on a *yes*/*no* scale (adapted from Fiorella & Mayer, 2013, 2014; available in the online supplemental materials). Sample content items include: "I know what lactase is" for the "food allergies" topic, and "I know what cognitive processes are" for the "intelligence quotient" topic. A prior knowledge score was computed for each learner 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. As opposed to presenting actual test questions, this relatively more indirect measure of prior knowledge was used to avoid inducing any retrieval-based learning effects in the concept-mapping and learning-by-teaching conditions. In addition, participants made a judgment of learning (JOL) by predicting how well they thought they would perform on a test on their respective study topic (1 = *very poorly*; 5 = *very well*).

Standard Questions. A pool of 32 standard questions—16 questions for each critical study text—was constructed to be presented as "tutee" questions in the learning-by-teaching condition. Each standard question was based directly on the content from one of the four paragraphs in the critical study text, and took on one of four forms corresponding to the *understand* and *apply* levels of Bloom's taxonomy, either: (a) "Could you give an example of X?" or (b) "Knowing X, how might we apply this to Y?" or (c) "Could you summarize and explain X?" or (d) "How does X relate to Y?" A sample question for the "food allergies" text was: "Could you summarize and explain why the bacteria-killing properties of antibiotics could be bad?" whereas a sample question for the "intelligence quotient" text was: "Could you summarize and explain why genetic differences between races are usually superficial?" All standard questions, as well as the frequency at which each question was asked, are listed in the online supplemental materials.

Procedure

Training Phase. Upon arriving at the research laboratory, all learners were first trained on question generation, during which they received a printed handout that described and explained the various question levels based on Bloom's taxonomy. Learners were then presented with the "enzymes" practice text and were given 5 min to practice generating one *apply*, one *analyze*, and one *evaluate* question based on the text. Then, learners were given 4 min to practice generating two *create* questions (i.e., research questions) based on the "enzymes" text. For each question level, learners received concise verbal feedback on the questions that they generated. Specifically, learners were advised within a sentence whether their question(s) correctly reflected the intended question level and, if not, how they could modify their question(s) appropriately.

Study Phase. After completing the question generation training, participants responded to the prelearning questionnaire. Then, they were given a handout of the critical study text and a blank sheet of paper to write their responses, and were instructed on the learning strategy that they had been randomly assigned to use. Participants completed the study phase individually, which spanned 18 min. Thus, the total study duration was exactly matched across all learning conditions.

In the *concept-mapping* condition, learners were instructed on the characteristics of concept maps such as the use of labeled nodes denoting key concepts and links denoting the relations among these concepts. Learners were also shown examples of good concept maps for illustration (adapted from Novak, 2005). To further ascertain that learners fully understood what was required of them, they were asked to practice drawing a concept map of a brief 42-word text on "muscle tissue" (adapted from Karpicke & Blunt, 2011b) that was not related to either of the critical study texts. Then, learners were given 18 min to construct a concept map of their respective critical study text on either "food allergies" or "intelligence quotient."

In the *retrieval practice* condition, the 18-min study period comprised four consecutive 4.5-min blocks, during which participants alternated between studying their respective critical study text and practicing retrieval (i.e., study–retrieve–study–retrieve; e.g., Karpicke & Blunt, 2011b; Karpicke & Roediger, 2007). Specifically, learners first studied the text for 4.5 min, then engaged in retrieval for 4.5 min by writing down as much information from the text as they could remember. Following this, learners restudied the text for 4.5 min and recalled it again for 4.5 min. This free recall procedure is a commonly adopted and effective form of retrieval-based learning (Bae et al., 2019; Karpicke & Blunt, 2011b; Roediger & Karpicke, 2006a).

In the *learning-by-teaching* condition, participants were first given 10 min to prepare teaching notes of the study text, before teaching the material for 3 min with reference to their notes while being filmed. So that participants taught with a target audience in mind for an authentic teaching experience, they were informed that their video-recorded lecture would subsequently be viewed by an audience for educational and research purposes (Fiorella & Mayer, 2013, 2014). Then, to simulate teacher–student interactions that typically occur in real-world classroom settings (e.g., Fiorella & Mayer, 2015; Roscoe & Chi, 2007, 2008), the experimenter posed as a "tutee" and asked the participant questions about their lesson (e.g., Bargh & Schul, 1980). All questions were drawn from the pool of standard questions that had been prepared prior to the experiment. The questioning segment lasted for 5 min or when eight questions had been asked, whichever limit was reached first. Participants were asked only questions that directly related to the content that they had taught—this ensured that the questions posed were relevant to their lesson and that no "new" content was inadvertently introduced. For instance, if participants had mentioned ideas drawn from a particular paragraph of the study text during their teaching, they were asked a corresponding standard question that was associated with that paragraph. As when teaching, participants were allowed to refer to their self-made notes when responding to their "tutee's" questions—this ensured that there was no need or reason for participants to engage in retrieval (Koh et al., 2018). Participants were not provided with any feedback on their responses to the questions, similar to how concept-mapping and retrieval practice participants did not receive any feedback on their study responses.

Test Phase. All participants then completed a 10-min research question generation test without reference to the study text, in which they generated and wrote down as many *create*-level research questions as they could based on the text content that they had earlier studied. Next, participants completed a 5-min recall test in which they wrote down as much information as they could remember from the study text. Participants were allowed to paraphrase the information in their own words and write down their responses in point form. Finally, participants were debriefed and thanked.

Results

Scoring

Participants' research question generation performance was scored as the number of questions they generated that reflected the *create* level of Bloom's taxonomy. Specifically, a question would be considered a *create* question if its answer required creating new knowledge, ideas, or perspectives by compiling information from the study texts in a different way, combining elements in a new pattern, or proposing alternative solutions or hypotheses that had not been mentioned in the study texts (see Table 1). Questions that did not fulfill these criteria (e.g., those that corresponded to the other levels of Bloom's taxonomy: *remember*, *understand*, *apply*, *analyze*, or *evaluate*) did not receive any points as *create* questions but were scored as non-*create* questions. For instance, sample *create*-level research questions for the "food allergies" versus "intelligence quotient" topics were: "How can the immune system be made to 'forget' a tagged irritant so that food allergies can be cured?" and "How can we make an IQ test fair to every person regardless of their social conditions?" Conversely, sample non-*create* questions for the "food allergies" versus "intelligence quotient" topics were: "Besides skin rashes and inflammation, what other food allergic reactions occur?" and "Do genes or environmental factors play a larger role in shaping intelligence?"

In addition, participants' recall performance was scored as the number of idea units from the critical study text that they correctly recalled on the recall test, with a maximum score of 40. Both verbatim restatements and paraphrases that preserved the meaning of the text content were considered correct.

Two raters independently scored 16 of the 78 (20%) scripts blind to experimental condition. Interrater reliability was high for the classification of participants' generated questions as *create* versus non-*create* questions, Cohen's kappa (κ) = .92. There was also high interrater reliability when scoring the total number of *create*-level research questions generated and the number of idea units that each participant recalled at test, absolute agreement intraclass correlation (ICC) = .97 and .99, 95% CI [.92, .99] and [.98, .99], respectively, based on a two-way random-effects model. Discrepancies were reviewed and resolved through discussion to reach 100% agreement. Given the high interrater reliability, the remaining scripts were scored by one rater.

Preliminary Analyses

We ascertained that the three learning groups did not significantly differ in their self-reported prior knowledge of the study texts, $F(2, 75) = 0.86$, $p = .43$, $\eta_p^2 = .02$. Out of a total possible score of 13, participants reported relatively low prior knowledge across the concept-mapping ($M = 4.62$, $SD = 2.04$), retrieval practice ($M = 4.46$, $SD = 2.18$), and learning-by-teaching ($M = 5.35$, $SD = 3.38$) conditions. In addition, participants' JOL predictions of their test performance did not differ across the concept-mapping ($M = 1.77$, $SD = 0.91$), retrieval practice ($M = 1.85$, $SD = 0.93$), and learning-by-teaching ($M = 2.08$, $SD = 1.06$) conditions, $F(2, 75) = 0.72$, $p = .49$, $\eta_p^2 = .02$.

Main Analyses

Research Question Generation Performance. A 3 (learning strategy) × 2 (study text) between-subjects analysis of variance (ANOVA) revealed that the three learning groups significantly differed in their research question generation performance, $F(2, 72) = 9.03$, $p < .001$, $\eta_p^2 = .20$. As expected, learners who had taught ($M = 4.19$, $SD = 2.32$) generated more *create*-level research questions than those who had constructed concept maps ($M = 2.31$, $SD = 1.59$) or practiced retrieval ($M = 2.42$, $SD = 1.65$), $p < .001$

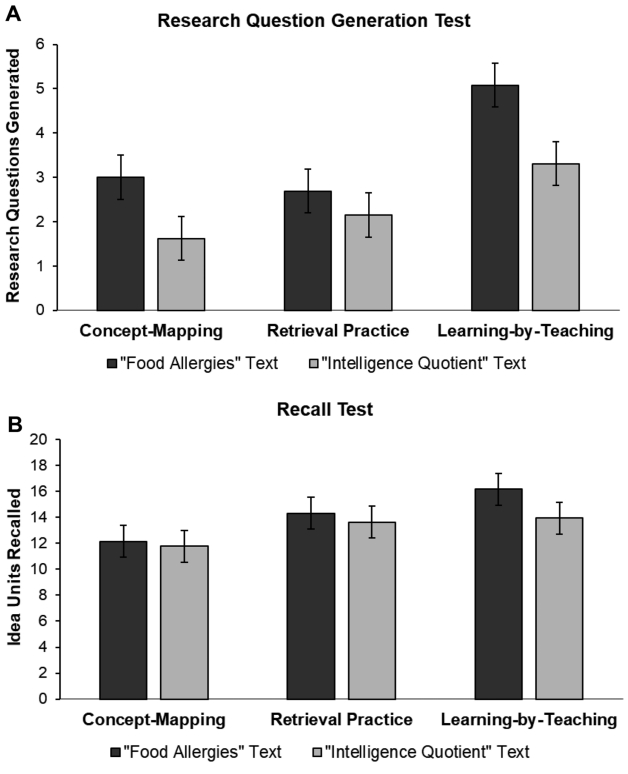
and $p = .001$, $d = 0.95$ and 0.88, respectively; the latter two groups did not differ, $p = .82$. There was also a significant main effect of study text, $F(1, 72) = 9.20$, $p = .003$, $\eta_p^2 = .11$, whereby participants generated more research questions for the "food allergies" ($M = 3.59$, $SD = 2.28$) than "intelligence quotient" ($M = 2.36$, $SD = 1.60$) text on overall. Importantly, however, there was no Learning Strategy × Study Text interaction, $F(2, 72) = 0.80$, $p = .45$, $\eta_p^2 = .02$, indicating that the advantage of learning-by-teaching generalized across study topics from both the natural and social sciences (Figure 1A).

Nonresearch Question Generation Performance. Whereas participants were explicitly instructed to generate *create*-level research questions at test, they inadvertently generated some non-*create* questions too (i.e., questions that did not fulfill the criteria for the *create* level of Bloom's taxonomy). Although non-*create* questions were not the main focus of this study, we report the data here for completeness. A 3 (learning strategy) × 2 (study text) between-subjects ANOVA indicated no significant difference in the number of non-*create* questions that the concept-mapping ($M = 7.15$, $SD = 3.32$), retrieval practice ($M = 6.92$, $SD = 2.93$), and learning-by-teaching ($M = 8.12$, $SD = 4.01$) groups generated, $F(2, 72) = 0.87$, $p = .42$, $\eta_p^2 = .02$. The number of non-*create* questions that learners generated also did not significantly differ across the "food allergies" ($M = 7.03$, $SD = 3.46$) and "intelligence quotient" ($M = 7.77$, $SD = 3.43$) texts, $F(1, 72) = 0.90$, $p = .35$, $\eta_p^2 = .01$. Neither was there a Learning Strategy × Study Text interaction, $F(2, 72) = 0.79$, $p = .46$, $\eta_p^2 = .02$. Thus, the learning-by-teaching advantage for question generation performance was specific to *create*-level research questions that all participants had been instructed to formulate at test, rather than any type of question in general.

Recall Test Performance. Analyzing participants' recall test performance, a 3 (learning strategy) × 2 (study text) between-subjects ANOVA revealed that the three learning groups significantly differed, $F(2, 72) = 3.29$, $p = .043$, $\eta_p^2 = .08$. Learning-by-teaching participants ($M = 15.04$, $SD = 4.80$) recalled more idea units from the study text than concept-mapping participants ($M = 11.96$, $SD = 4.32$), $p = .014$, $d = 0.67$, whereas retrieval practice participants ($M = 13.96$, $SD = 3.91$) did not significantly differ from either concept-mapping or learning-by-teaching participants in their recall, $p = .11$ and .38, respectively. Learners' recall performance did not significantly differ across the "food allergies" ($M = 14.21$, $SD = 4.49$) and "intelligence quotient" ($M = 13.10$, $SD = 4.48$) texts, $F(1, 72) = 1.23$, $p = .27$, $\eta_p^2 = .02$. There was also no Learning Strategy × Study Text interaction, $F(2, 72) = 0.33$, $p = .72$, $\eta_p^2 = .01$ (Figure 1B).

Figure 1

Immediate Final Test Performance Across Learning Conditions and Study Topics (Experiment 1)



Note. A and B show the mean research question generation and recall test scores, respectively. Error bars indicate standard errors.

Discussion

Experiment 1 provided initial evidence for a learning-by-teaching benefit on the higher-order educational outcome of research question generation. As compared to their peers who practiced retrieval or constructed concept maps, learners who taught were subsequently more successful in generating more *create*-level research questions based on the study material. In addition, learning-by-teaching improved participants' retention of the material over concept-mapping and did not significantly differ from retrieval practice.

Notwithstanding retrieval practice as a well-established strategy for enhancing knowledge retention (Dunlosky et al., 2013; Karpicke, 2017), it did not produce better recall than concept-mapping. While somewhat surprising, this finding is consistent with the lack of a retrieval practice benefit in some other studies that similarly administered the final test shortly after the learning phase, rather than after a longer lag (Roediger & Karpicke, 2006a, 2006b). For instance, Roediger and Karpicke (2006a) observed that repeated studying outperformed retrieval practice when the final recall test was given after 5 min, but that the opposite pattern occurred on delayed recall tests given either 2 days or 1 week later, whereby retrieval practice improved long-term retention more than repeated studying. Thus, it is possible that the lack of a testing effect in Experiment 1 was due to the absence of a delay between the studying phase and the immediate final test. Hence, Experiment 2 was conducted to foreclose this issue using delayed final tests.

Experiment 2

Experiment 2 aimed to replicate Experiment 1's findings and extend them in two ways. First, we sought to determine the effectiveness of learning-by-teaching for durable learning and research question generation. Whether learning-by-teaching effects obtain on delayed tests is particularly important because education ultimately aims to promote long-term learning rather than transient gains. Moreover, techniques that enhance short-term performance may not necessarily yield the same advantage for long-term learning (Soderstrom & Bjork, 2015).

Accordingly, the final tests in Experiment 2 were administered after a 48-hr delay following the study phase, rather than immediately after it. This simultaneously ensured favorable conditions for the retrieval practice group to be maximally well-poised for greater success; although retrieval practice can benefit retention at shorter intervals of less than 1 day, testing effects tend to be larger with longer retention intervals of at least 1 day (Adesope et al., 2017; Rowland, 2014). Thus, the delayed tests in Experiment 2 provided a more stringent assessment of learning-by-teaching's efficacy over retrieval practice and concept-mapping.

Second, although Experiment 1 faithfully administered all three learning methods in the way that they have typically been used in research and educational settings, as well as within the exact same learning duration, it is possible that some defining elements that are unique to learning-by-teaching may have placed it at an advantage. In particular, learning-by-teaching participants received and responded to standard "tutee" questions about their teaching, whereas concept-mapping and retrieval practice participants did not since responding to an audience's questions does not ordinarily comprise a defining feature of these methods. Hence, the learning-by-teaching advantage in Experiment 1 may have been driven by exposure to or engagement with more questions during study, even if these were "lower-order" questions that did not constitute research questions. To rule out this alternative account, all participants in Experiment 2—including those in the concept-mapping and retrieval practice groups—were similarly presented with standard questions to answer during the study phase.

Method

Participants

The participants were 78 undergraduate students (51 were female) aged between 18 and 28 ($M = 21.77$, $SD = 2.13$) from the National University of Singapore. Outcomes reported below are based on data from 74 participants—two participants failed to return for the delayed final test, and two participants who did not follow the experimental instructions were excluded from subsequent analyses. The final sample comprised 26 participants in the learning-by-teaching group and 24 participants each in the concept-mapping and retrieval practice groups. The three learning groups did not significantly differ in their mean age or proportion of men and women.

Design, Materials, and Procedure

The design, materials, and procedure were identical to those in Experiment 1 but with two crucial differences. First, the standard questions on the critical study texts were presented to all three learning groups during a 5-min questioning segment at the end of the study phase. Accordingly, the study phase duration was extended from 18 to 23 min across all learning conditions—after completing the question generation training and prelearning questionnaire, participants studied their respective critical study text (either "food allergies" or "intelligence quotient") for 18 min using the learning method that they had been randomly assigned, followed by a 5-min questioning segment. Specifically, as in Experiment 1: *Concept-mapping* participants were instructed and trained on the basic characteristics of concept maps before being given 18 min to construct a concept map of the text; *retrieval practice* participants alternated between studying and practicing retrieval of the text over four 4.5-min periods for a total of 18 min; *learning-by-teaching* participants prepared teaching notes of the text for 13 min, then taught the material with reference to their self-made notes whilst being filmed for 5 min. After the 18-min period, all participants underwent a questioning segment in which they were presented with standard questions on the text, one at a time and in a randomized order, and responded to these questions for a duration of 5 min or up to eight questions, whichever limit was reached first. All participants were allowed to refer to the study text or their self-made teaching notes when answering the standard questions. Across all learning conditions, no feedback on participants' responses to the standard questions was provided.

To control the frequency at which each standard question was presented across learning conditions, we computed the presentation frequency of each of the standard questions as a percentage of all questions that had been asked during the questioning segment in Experiment 1's learning-by-teaching condition. We expected that the presentation frequency of the standard questions in Experiment 1 would provide reasonable estimates of how often each of these same questions would eventually be posed to learning-by-teaching participants in Experiment 2. Accordingly, we closely matched the relative frequency at which each standard question was asked in Experiment 1 to how often it was presented to Experiment 2's concept-mapping and retrieval practice groups (see the online supplemental materials).

Second, a crucial procedural departure from Experiment 1 was that the final tests were administered after a 48-hr retention interval. Two days after all participants had studied their respective critical study text, they returned to the lab for a final research question generation test and recall test.

Results

Scoring

Two independent raters who were blind to experimental condition scored 15 of the 74 (20%) scripts in the same way as in Experiment 1. Interrater reliability was high for the classification of participants' generated questions as *create* versus non-*create* questions, Cohen's $\kappa = .91$. There was also high interrater reliability when scoring the total number of *create*-level research questions generated and the number of idea units that each participant recalled at test, absolute agreement ICC = .96 and .96, 95% CI [.87, .99] and [.85, .99], respectively, based on a two-way random-effects model. Discrepancies were reviewed and resolved through discussion to reach 100% agreement. Given the high interrater reliability, the remaining scripts were scored by one rater.

Preliminary Analyses

As in Experiment 1, we first ascertained that participants did not significantly differ in their self-reported prior knowledge of the study texts, $F(2, 71) = 1.24$, $p = .30$, $\eta_p^2 = .03$. On overall, participants reported relatively low prior knowledge of the material across the concept-mapping ($M = 4.21$, $SD = 2.00$), retrieval practice ($M = 4.29$, $SD = 2.33$), and learning-by-teaching ($M = 5.19$, $SD = 2.93$) conditions, out of a total possible score of 13. In addition, the concept-mapping ($M = 2.08$, $SD = 1.06$), retrieval practice ($M = 1.87$, $SD = 0.85$), and learning-by-teaching ($M = 2.23$, $SD = 1.07$) groups did not significantly differ in their

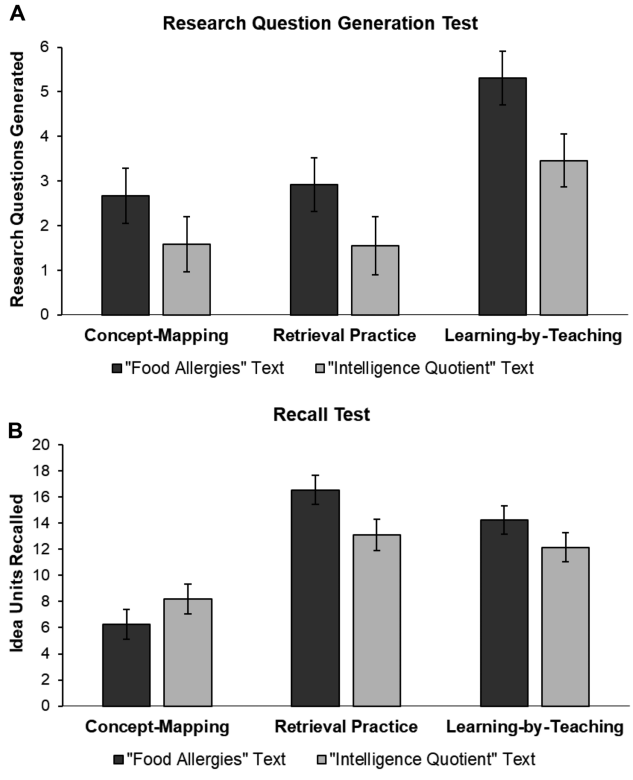
JOLs when predicting how well they would perform when tested, $F(2, 71) = 0.79$, $p = .46$, $\eta_p^2 = .02$.

Main Analyses

Research Question Generation Performance. Replicating Experiment 1's findings, a 3 (learning strategy) × 2 (study text) between-subjects ANOVA revealed that the three learning groups significantly differed in the number of *create*-level research questions that they generated at test, $F(2, 68) = 8.85$, $p < .001$, $\eta_p^2 = .21$. Strikingly, participants who had taught the material ($M = 4.38$, $SD = 2.74$) generated twice the number of research questions of their peers who had constructed concept maps ($M = 2.13$, $SD = 1.94$) or practiced retrieval ($M = 2.29$, $SD = 1.88$), $p < .001$ and $p = .001$, $d = 0.95$ and 0.89, respectively, whereas the latter two groups did not differ, $p = .86$. There was also a significant main effect of study text, whereby participants generated more research questions based on the "food allergies" ($M = 3.66$, $SD = 2.75$) than "intelligence quotient" ($M = 2.25$, $SD = 1.84$) text on overall, $F(1, 68) = 8.19$, $p = .01$, $\eta_p^2 = .11$. Crucially, however, there was no Learning Strategy × Study Text interaction, $F(2, 68) = 0.20$, $p = .82$, $\eta_p^2 = .01$, indicating that the learning-by-teaching group outperformed their peers regardless of study topic across the natural and social sciences (Figure 2A).

Figure 2

Delayed (48-hr) Final Test Performance Across Learning Conditions and Study Topics (Experiment 2)



Note. A and B show the mean research question generation and recall test scores, respectively. Error bars indicate standard errors.

Nonresearch Question Generation Performance. We further examined the number of non-*create* questions that participants inadvertently generated on the research question generation test (i.e., questions that did not fulfill the criteria for the *create* level of Bloom's taxonomy). As in Experiment 1, a 3 (learning strategy) × 2 (study text) between-subjects ANOVA indicated no significant difference across the concept-mapping ($M = 9.04$, $SD = 5.90$), retrieval practice ($M = 6.33$, $SD = 3.49$), and learning-by-teaching ($M = 6.77$, $SD = 3.01$) groups in the number of non-*create* questions generated, $F(2, 68) = 2.94$, $p = .06$, $\eta_p^2 = .08$. Thus, the benefit of learning-by-teaching for question generation performance was specific to *create*-level research questions that all participants had been instructed to formulate at test, rather than any type of question in general. On overall, participants generated more non-*create* questions based on the "intelligence quotient" ($M = 8.67$, $SD = 5.23$) than "food allergies" ($M = 6.13$, $SD = 3.00$) text, $F(1, 68) = 6.67$, $p = .01$, $\eta_p^2 = .09$. Nevertheless, the Learning Strategy × Study Text interaction was nonsignificant, $F(2, 68) = 1.00$, $p = .38$, $\eta_p^2 = .03$.

Recall Test Performance. A 3 (learning strategy) × 2 (study text) between-subjects ANOVA indicated that the three learning groups differed in their recall performance, $F(2, 68) = 25.08$, $p < .001$, $\eta_p^2 = .42$. Replicating the classic testing effect (e.g., Karpicke & Blunt, 2011b; O'Day & Karpicke, 2021), retrieval practice ($M = 14.96$, $SD = 4.07$) dramatically improved long-term retention on the delayed recall test, with learners recalling more than twice the amount of content that concept-mapping participants did ($M = 7.21$, $SD = 3.70$), $p < .001$, $d = 1.99$. Notably, learning-by-teaching ($M = 13.19$, $SD = 4.35$) also outperformed concept-mapping, $p < .001$, $d = 1.48$, and did not significantly differ from the potent technique of retrieval practice, $p = .15$. The number of idea units that learners recalled did not significantly differ across the "food allergies" ($M = 12.50$, $SD = 6.19$) and "intelligence quotient" ($M = 11.11$, $SD = 3.82$) texts on overall, $F(1, 68) = 1.73$, $p = .19$, $\eta_p^2 = .03$. There was also no Learning Strategy × Study Text interaction, $F(2, 68) = 3.03$, $p = .06$, $\eta_p^2 = .08$ (Figure 2B).

Discussion

Extending Experiment 1's key finding, Experiment 2 showed that the learning-by-teaching advantage for research question generation held even on a delayed test 48 hr later. Indeed, participants who had taught the study material generated twice the number of *create*-level research questions of their peers who practiced retrieval or constructed concept maps. Furthermore, learning-by-teaching produced superior retention than concept-mapping on the delayed test, and rivaled retrieval practice that is well-established as one of the most powerful techniques for improving long-term memory (Dunlosky et al., 2013). Clearly, the learning-by-teaching effect on the delayed recall and research question generation tests is not simply a result of greater exposure to or engagement with tutee questions, since all groups similarly received and responded to standard questions on the study material. We consider other candidate accounts in the General Discussion section.

Whereas retrieval practice did not improve immediate recall relative to concept-mapping in Experiment 1, it produced a robust benefit for long-term retention in Experiment 2. This pattern of results echoes those reported by other studies in the testing-effect literature (Karpicke & Blunt, 2011b; Roediger & Karpicke, 2006a, 2006b). Yet, the strong recall advantage that retrieval practice participants

enjoyed did not boost their research question generation performance. Replicating Experiment 1's findings, retrieval practice failed to confer any benefits over concept-mapping on the delayed research question generation test.

General Discussion

The ability to generate good research questions sets the stage for scientific inquiry and discovery. Across two experiments, we found that learning-by-teaching is an effective strategy to enhance this ability among students across topics in the natural and social sciences. Specifically, students who taught a scientific text outperformed their peers who practiced retrieval or constructed concept maps when tested on their ability to generate *create*-level research questions based on the text (Experiment 1). This advantage persisted durably after a 48-hr delay, and even when the retrieval practice and concept-mapping groups engaged with poststudy questions as did the learning-by-teaching group (Experiment 2).

In addition, learning-by-teaching improved recall more than concept-mapping across both experiments, and in fact rivaled retrieval practice in substantially enhancing long-term retention on a delayed test in Experiment 2. This finding is noteworthy given the abundant research demonstrating the potency of retrieval practice in producing large gains for long-term retention (Karpicke, 2017; Rowland, 2014). Comparatively, the benefits of learning-by-teaching for durable retention have arguably received less attention to date and thus call for further validation.

Yet, our data also reveal that simply possessing a wealth of factual knowledge is insufficient to improve higher-order research question generation. Despite being able to recall more than twice the study content that concept-mapping participants did after two days, retrieval practice participants failed to generate more research questions based on that content (Experiment 2). These findings align with those of a growing number of studies suggesting that retrieval practice alone may not benefit some complex, higher-order learning outcomes such as inferencing (McDaniel et al., 2009; Nguyen & McDaniel, 2016) and integrative argumentation (Wong & Lim, 2019a). Indeed, the present study found no evidence that retrieval practice improves *create*-level research question generation.

Of note, the lack of a retrieval practice advantage for question generation performance appears to extend to other higher-order question levels. In a concurrent, related line of work in our lab comprising three experiments ($N = 168$), we tested the extent that retrieval practice more generally boosts learners' ability to ask good questions (see Experiments 3a–3c in the online supplemental materials for full details). Learners studied a scientific expository text either by alternating between study and retrieval, or by studying it repeatedly. A week later, they returned to receive training on question generation, then generated as many lower-order (*remember* and *understand*) and higher-order (*apply*, *analyze*, *evaluate*, and *create*) questions as they could based on the text that they had earlier studied, and were also tested on their recall of the text content. Although learners who practiced retrieval displayed better long-term retention and lower-order question generation performance than their peers who restudied, we consistently found that they did not generate more higher-order questions (Experiment 3a), even when explicitly instructed to focus solely on generating such questions at test (Experiment 3b). Despite further bolstering retrieval practice with a metacomprehension monitoring intervention—judgments of higher-order learning (JOL+; Wong & Lim, 2019a)—that oriented learners' attention toward the kinds of processing needed for effective higher-order question generation, learners' performance remained at bay (Experiment 3c). Altogether, retrieval practice is a potent technique for enhancing durable retention but does not suffice for improving higher-order question generation.

Theoretical Explanations for the Learning-by-Teaching Effect

Why does learning-by-teaching surpass concept-mapping and retrieval practice in enhancing *create*-level research question generation? Although the extant data can yet fully explain the learning benefits of teaching, some explanations are more plausible than others within the given experimental parameters. Foremost, because the present study was specifically intended to dissociate the effects of teaching versus retrieval practice (Koh et al., 2018), learning-by-teaching participants were given access to their self-generated notes whilst teaching, ensuring that there was no need or grounds for them to retrieve the material from memory. Thus, the learning benefits of teaching observed here are unlikely to be due to retrieval practice.

Second, although all three learning methods in the present study have been considered generative learning strategies (Fiorella & Mayer, 2015, 2016), it is possible that teaching induced higher levels or fundamentally different types of generative processing that enabled the tutor to create new ideas and research questions about the material, relative to retrieval practice and concept-mapping. Which stage(s) of the teaching process could have contributed to this? Since all learners in Experiment 2 responded to standard questions about the study material, it is unlikely that the subsequent learning-by-teaching advantage arose from the stage of responding to tutee questions. Rather, the tutor's learning gains in the present study more likely stemmed from the stages of expecting to teach and/or actually teaching. For instance, expecting to teach rather than be tested has been found to enhance learners' organization of the material (Nestojko et al., 2014) and intrinsic motivation to learn (Benware & Deci, 1984; Guerrero & Wiley, 2021). Through taking on the role of a teacher, learners may enact behaviors that they perceive to be defined by this role when preparing to teach, such as organizing or restructuring the content when considering the relations among ideas in the text (Bargh & Schul, 1980). Furthermore, when actually teaching, the tutor may generate elaborations and inferences to ensure that their explanations are coherent and understandable by their intended audience, while monitoring their own learning to remedy any knowledge gaps (Lachner et al., 2020; Muis et al., 2016; Roscoe & Chi, 2008). In turn, such processes may enable the tutor to build a richer situation model of the material (Kintsch, 1988) for deep and durable learning that facilitates their research question generation.

Finally, teaching may elicit a sense of "productive agency" (Schwartz, 1999) or even feelings of power whilst viewing oneself as a teacher who gives advice that could influence others' actions (Schaerer et al., 2018). In anticipating their tutees' learning needs, the tutor may then engage in further adaptation processes (Clark & Brennan, 1991), such as tailoring their explanations to include more elaborations for a less knowledgeable audience (Nickerson, 1999; Wittwer et al., 2010). In contrast, generating egocentric

content for one's own learning may not prompt such processes or do so to a lesser extent.

Indeed, retrieval practice and concept-mapping may have been less successful in triggering such useful mechanisms that promote deep learning and inspire high-quality research questions. Whereas retrieval practice and learning-by-teaching both yielded strong gains for long-term retention (Experiment 2), observing similar behavioral performance in any two conditions does not always mean that similar mechanisms apply. For instance, although retrieval practice also offers metacognitive monitoring advantages in helping learners diagnose what they do not know or remember to guide their subsequent restudy (Little & McDaniel, 2015), it may not necessarily alert learners to deficits in their global-level situation model of the text (Nguyen & McDaniel, 2016; Wong & Lim, 2019a). In particular, learners tend to adopt a knowledge-telling bias in summarizing or "telling" what they know, rather than engaging in reflective knowledge-building when (re)constructing knowledge in meaningful ways that aid their deep learning (Roscoe & Chi, 2008). Thus, when taking on the role of a "learner" and testing themselves from memory during the study phase, retrieval practice participants could have relied mainly on knowledge-telling and textbase processing, as compared to knowledge-building or situation-model processing. Consequently, retrieval practice produced little benefit on the final research question generation test despite improving recall (Experiments 1 and 2). Notably, this lack of benefit could not be overcome even when retrieval practice had been supplemented with JOL+ questions as a metacomprehension monitoring intervention to guide learners' restudy toward the higher-order processes needed for research question generation (see Experiment 3c in the online supplemental materials).

Likewise, concept-mapping participants took on the role of a "learner" in preparing to be tested. Although concept-mapping fundamentally involves elaborating on the to-be-learned material by organizing and constructing links among various concepts, it is possible that its benefits may be more pronounced when learners are provided with more extensive training to construct higher-quality maps (Fiorella & Mayer, 2016). However, there seems to be little—if any—evidence from randomized controlled experiments that extensive training is necessary for concept-mapping to be effective (for discussions, see Karpicke & Blunt, 2011a; Wong & Lim, 2022b). Moreover, tedious and time-consuming training may lead students to lose interest in this technique (Fiorella & Mayer, 2016). Thus, from a practical standpoint, teaching is a relatively more efficient learning technique since students did not require any extensive training to reap its benefits for knowledge retention and research question generation within the same study duration. We are hopeful that future experimental work will distill the processes underlying learning-by-teaching effects. In turn, this knowledge would guide efforts on implementing learning-by-teaching to stimulate knowledge discovery and innovation.

Educational Implications and Future Directions

Not all generative learning strategies are equal—depending on the learning context and desired pedagogical outcomes, the most effective strategy should be appropriately applied (Fiorella & Mayer, 2016). For instance, our findings suggest that if the learning goal were to increase one's durable memory for the studied material, then practicing retrieval or engaging in teaching activities would be more effective than concept-mapping. Conversely, if the learning goal were to generate novel research questions based on the studied material, then students would profit more from learning-by-teaching than retrieval practice or concept-mapping. In sum, our findings support the idea that teaching-based learning activities are powerful for enhancing not only long-term basic retention but also in pushing students toward achieving the highest level of Bloom's taxonomy—*creating* new knowledge through generating novel research questions.

Although the present study was specifically designed to dissociate the effects of teaching versus retrieval practice, it should be noted that both techniques can be applied in combination in education. Basic knowledge about the target topic or material at hand, which lower-ability students may tend to lack relative to higher-ability students, is crucial as the foundation on which to generate associated deeper questions. Notably, retrieval practice accompanied by feedback has been shown to benefit lower-ability students during learning (Agarwal et al., 2017). Thus, retrieval practice may be judiciously introduced *prior to* learning-by-teaching (e.g., Roelle et al., 2022) in leveling the playing field for these weaker students, so that no one is left behind in the learning process.

Indeed, examining how retrieval practice can be astutely integrated with learning-by-teaching offers a promising prospect for future work. Although retrieval practice imposes a hurdle in demanding that students successfully recall the material before they can go on to organize and elaborate on the retrieved content, it may not necessarily harm the quantity and quality of students' generative responses despite reducing the amount of content covered during study (Roelle & Nückles, 2019; Waldeyer et al., 2020). However, because the reduced coverage in students' generative responses can impair their subsequent learning performance (Roelle & Nückles, 2019), remedying this issue is critical. For instance, one viable solution may be to incorporate retrieval practice in generative activities via a closed/open switch style rather than a purely closed-book style (Waldeyer et al., 2020). That is, students could engage in organization and elaboration without referring to the material as much as possible but are permitted to access the material on demand when in doubt. In line with this notion, some evidence suggests that allowing students to flexibly switch between closed- and open-book styles during generative activities may optimize learning benefits (Waldeyer et al., 2020). Extending this area of work, it would be interesting for future research to consider how learning-by-teaching can be implemented in tandem with other effective learning techniques such as *distributed practice* (for reviews, see Carpenter et al., 2022; Cepeda et al., 2006; Dunlosky et al., 2013), *interleaving* (e.g., Brunmair & Richter, 2019; Firth et al., 2021; Kornell & Bjork, 2008; Wong et al., 2020, 2021), and *learning from errors* (Metcalfe, 2017; Wong & Lim, 2019b) that are induced before instruction (Kapur, 2008; Kapur & Bielaczyc, 2012) or even deliberately committed and corrected during study (Wong, 2023; Wong & Lim, 2022b, 2022c).

How much of the benefits of learning-by-teaching for research question generation are attributable to preparing to teach versus actually teaching? Although comparisons of individual teaching stages would necessitate "stripped down" versions of the full teaching process, this could reveal important theoretical insights on the specific contributions of each stage. For instance, expecting to teach in itself has been found to boost memory and comprehension on both immediate and delayed tests (Guerrero & Wiley, 2021; cf. Fiorella &

Mayer, 2013, 2014), with greater benefits when coupled with actually teaching (Kobayashi, 2019). It remains to be explored whether such trends extend to more complex, higher-order learning outcomes such as research question generation.

Furthermore, a growing body of research suggests that higher-order thinking and learning require epistemic cognition—one's conceptions of knowledge and knowing, which influence the way that one constructs, critically evaluates, and uses knowledge (e.g., Cartiff et al., 2021; Greene & Yu, 2016; Sinatra et al., 2014). By logical extension, positioning students to enact effective epistemic cognition may moderate the effects of teaching others on their research question generation performance. Future work ought to test this prediction directly.

Conclusion

Traditionally, learning has often been assessed by having students answer teacher-prescribed test questions. The present research offers two ways of re-thinking educational processes and outcomes. First, meaningful learning can be assessed through having students devise their own higher-order research questions aimed at creating and discovering new knowledge, beyond deriving answers to their teachers' questions. Crucially, students do this best by becoming the teacher themselves. As compared to practicing retrieval or constructing concept maps, teaching others was more effective in enhancing students' research question generation across both immediate and delayed tests, while benefiting their memory retention. Indeed, researchers and educators ought to collaboratively think of how best our student might *be* the teacher, so that as the leaders of tomorrow, they may become bearers and bestowers of the best questions and ideas about the complex world that we live in.

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

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, *111*(2), 189–209. https://doi.org/10.1037/edu0000282

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory*, *25*(6), 764–771. https://doi.org/10.1080/09658211.2016.1220579

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*(4), 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

Agee, J. (2009). Developing qualitative research questions: A reflective process. *International Journal of Qualitative Studies in Education*, *22*(4), 431–447. https://doi.org/10.1080/09518390902736512

Allison, A. W., & Shrigley, R. L. (1986). Teaching children to ask operational questions in science. *Science Education*, *70*(1), 73–80. https://doi.org/10.1002/sce.3730700109

Alvesson, M., & Sandberg, J. (2011). Generating research questions through problematization. *Academy of Management Review*, *36*(2), 247–271. https://doi.org/10.5465/amr.2009.0188

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (abridged ed.). Addison Wesley Longman.

Bae, C. L., Therriault, D. J., & Redifer, J. L. (2019). Investigating the testing effect: Retrieval as a characteristic of effective study strategies. *Learning and Instruction*, *60*, 206–214. https://doi.org/10.1016/j.learninstruc.2017.12.008

Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology*, *72*(5), 593–604. https://doi.org/10.1037/0022-0663.72.5.593

Barrows, H. S., & Tamblyn, R. M. (1980). *Problem-based learning: An approach to medical education*. Springer.

Benware, C. A., & Deci, E. L. (1984). Quality of learning with an active versus passive motivational set. *American Educational Research Journal*, *21*(4), 755–765. https://doi.org/10.3102/00028312021004755

Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & The Teachable Agents Group at Vanderbilt. (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, *19*(3–4), 363–392. https://doi.org/10.1080/08839510590910200

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. McKay.

Bowman-Perrott, L., Davis, H., Vannest, K., Williams, L., Greenwood, C., & Parker, R. (2013). Academic benefits of peer tutoring: A meta-analytic review of single-case research. *School Psychology Review*, *42*(1), 39–55. https://doi.org/10.1080/02796015.2013.12087490

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. https://doi.org/10.1037/bul0000209

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133. https://doi.org/10.1037/a0019902

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1547–1552. https://doi.org/10.1037/a0024140

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279–283. https://doi.org/10.1177/0963721412452728

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, *1*(9), 496–511. https://doi.org/10.1038/s44159-022-00089-1

Cartiff, B. M., Duke, R. F., & Greene, J. A. (2021). The effect of epistemic cognition interventions on academic achievement: A meta-analysis. *Journal of Educational Psychology*, *113*(3), 477–498. https://doi.org/10.1037/edu0000490

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

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

Chin, C., & Brown, D. E. (2002). Student-generated questions: A meaningful aspect of learning in science. *International Journal of Science Education*, *24*(5), 521–549. https://doi.org/10.1080/09500690110095249

Chin, C., & Osborne, J. (2008). Students' questions: A potential resource for teaching and learning science. *Studies in Science Education*, *44*(1), 1–39. https://doi.org/10.1080/03057260701828101

Chin, D. B., Dohmen, I. M., Cheng, B. H., Oppezzo, M. A., Chase, C. C., & Schwartz, D. L. (2010). Preparing students for future learning with

teachable agents. *Educational Technology Research and Development*, *58*(6), 649–669. https://doi.org/10.1007/s11423-010-9154-5

Chularut, P., & DeBacker, T. K. (2004). The influence of concept mapping on achievement, self-regulation, and self-efficacy in students of English as a second language. *Contemporary Educational Psychology*, *29*(3), 248–263. https://doi.org/10.1016/j.cedpsych.2003.09.001

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association. https://doi.org/10.1037/10096-006

Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*(2), 237–248. https://doi.org/10.3102/00028312019002237

Coleman, E. B., Brown, A. L., & Rivkin, I. D. (1997). The effect of instructional explanations on learning from scientific texts. *The Journal of the Learning Sciences*, *6*(4), 347–365. https://doi.org/10.1207/s15327809jls0604_1

Cuccio-Schirripa, S., & Steiner, H. E. (2000). Enhancement and analysis of science question level for middle school students. *Journal of Research in Science Teaching*, *37*(2), 210–224. https://doi.org/10.1002/(SICI)1098-2736(200002)37:2<210::AID-TEA7>3.0.CO;2-I

Dillon, J. T. (1984). The classification of research questions. *Review of Educational Research*, *54*(3), 327–361. https://doi.org/10.3102/00346543054003327

Dillon, J. T. (1988). The remedial status of student questioning. *Journal of Curriculum Studies*, *20*(3), 197–210. https://doi.org/10.1080/0022027880200301

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Duran, D., & Topping, K. J. (2017). *Learning by teaching: Evidence-based strategies to enhance learning in the classroom*. Routledge.

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

Fiorella, L., & Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology*, *38*(4), 281–288. https://doi.org/10.1016/j.cedpsych.2013.06.001

Fiorella, L., & Mayer, R. E. (2014). Role of expectations and explanations in learning by teaching. *Contemporary Educational Psychology*, *39*(2), 75–85. https://doi.org/10.1016/j.cedpsych.2014.01.001

Fiorella, L., & Mayer, R. E. (2015). *Learning as a generative activity: Eight learning strategies that promote understanding*. Cambridge University Press. https://doi.org/10.1017/CBO9781107707085

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

Firth, J., Rivers, I., & Boyle, J. (2021). A systematic review of interleaving as a concept learning strategy. *Review of Education*, *9*(2), 642–684. https://doi.org/10.1002/rev3.3266

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*(1), 104–137. https://doi.org/10.3102/00028312031001104

Greene, J. A., & Yu, S. B. (2016). Educating critical thinkers: The role of epistemic cognition. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 45–53. https://doi.org/10.1177/2372732215622223

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

Guerrero, T. A., & Wiley, J. (2021). Expecting to teach affects learning during study of expository texts. *Journal of Educational Psychology*, *113*(7), 1281–1303. https://doi.org/10.1037/edu0000657

Gunawardena, C. N. (1995). Social presence theory and implications for interaction and collaborative learning in computer conferences. *International Journal of Educational Telecommunications*, *1*(2/3), 147–166.

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

Hartford, F., & Good, R. (1982). Training chemistry students to ask research questions. *Journal of Research in Science Teaching*, *19*(7), 559–570. https://doi.org/10.1002/tea.3660190705

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review*, *16*(3), 235–266. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

Hoogerheide, V., Deijkers, L., Loyens, S. M. M., Heijltjes, A., & van Gog, T. (2016). Gaining from explaining: Learning improves from explaining to fictitious others on video, not from writing to them. *Contemporary Educational Psychology*, *44–45*, 95–106. https://doi.org/10.1016/j.cedpsych.2016.02.005

Hoogerheide, V., Loyens, S. M. M., & van Gog, T. (2014). Effects of creating video-based modeling examples on learning and transfer. *Learning and Instruction*, *33*, 108–119. https://doi.org/10.1016/j.learninstruc.2014.04.005

Hoogerheide, V., Visee, J., Lachner, A., & van Gog, T. (2019). Generating an instructional video as homework activity is both effective and enjoyable. *Learning and Instruction*, *64*, Article 101226. https://doi.org/10.1016/j.learninstruc.2019.101226

Jacob, L., Lachner, A., & Scheiter, K. (2020). Learning by explaining orally or in written form? Text complexity matters. *Learning and Instruction*, *68*, Article 101344. https://doi.org/10.1016/j.learninstruc.2020.101344

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, *26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, *21*(1), 45–83. https://doi.org/10.1080/10508406.2011.591717

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. (2011a). Response to comment on "Retrieval practice produces more learning than elaborative studying with concept mapping." *Science*, *334*(6055), Article 453. https://doi.org/10.1126/science.1204035

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

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Keeling, E. L., Polacek, K. M., & Ingram, E. L. (2009). A statistical analysis of student questions in a cell biology laboratory. *CBE—Life Sciences Education*, *8*(2), 131–139. https://doi.org/10.1187/cbe.08-09-0054

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

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*(4), 294–303. https://doi.org/10.1037/0003-066X.49.4.294

Kobayashi, K. (2018). Interactivity: A potential determinant of learning by preparing to teach and teaching. *Frontiers in Psychology*, *9*, Article 2755. https://doi.org/10.3389/fpsyg.2018.02755

Kobayashi, K. (2019). Learning by preparing-to-teach and teaching: A meta-analysis. *Japanese Psychological Research*, *61*(3), 192–203. https://doi.org/10.1111/jpr.12221

Koh, A. W. L., Lee, S. C., & Lim, S. W. H. (2018). The learning benefits of teaching: A retrieval practice hypothesis. *Applied Cognitive Psychology*, *32*(3), 401–410. https://doi.org/10.1002/acp.3410

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

Lachner, A., Backfisch, I., Hoogerheide, V., van Gog, T., & Renkl, A. (2020). Timing matters! Explaining between study phases enhances students' learning. *Journal of Educational Psychology*, *112*(4), 841–853. https://doi.org/10.1037/edu0000396

Lachner, A., Hoogerheide, V., van Gog, T., & Renkl, A. (2022). Learning-by-teaching without audience presence or interaction: When and why does it work? *Educational Psychology Review*, *34*(2), 575–607. https://doi.org/10.1007/s10648-021-09643-4

Lachner, A., Jacob, L., & Hoogerheide, V. (2021). Learning by writing explanations: Is explaining to a fictitious student more effective than self-explaining? *Learning and Instruction*, *74*, Article 101438. https://doi.org/10.1016/j.learninstruc.2020.101438

Leung, K. C. (2019). An updated meta-analysis on the effect of peer tutoring on tutors' achievement. *School Psychology International*, *40*(2), 200–214. https://doi.org/10.1177/0143034318808832

Lim, K. Y. L., Wong, S. S. H., & Lim, S. W. H. (2021). The "silent teacher": Learning by teaching via writing a verbatim teaching script. *Applied Cognitive Psychology*, *35*(6), 1492–1501. https://doi.org/10.1002/acp.3881

Little, J. L., & McDaniel, M. A. (2015). Metamemory monitoring and control following retrieval practice for text. *Memory & Cognition*, *43*(1), 85–98. https://doi.org/10.3758/s13421-014-0453-7

Marbach-Ad, G., & Sokolove, P. G. (2000). Can undergraduate biology students learn to ask higher level questions? *Journal of Research in Science Teaching*, *37*(8), 854–870. https://doi.org/10.1002/1098-2736(200010)37:8<854::AID-TEA6>3.0.CO;2-5

Mayer, R. E. (1984). Aids to text comprehension. *Educational Psychologist*, *19*(1), 30–42. https://doi.org/10.1080/00461528409529279

Mayer, R. E. (1996). Learning strategies for making sense out of expository text: The SOI model for guiding three cognitive processes in knowledge construction. *Educational Psychology Review*, *8*(4), 357–371. https://doi.org/10.1007/BF01463939

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 43–71). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.005

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

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

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

Muis, K. R., Psaradellis, C., Chevrier, M., Di Leo, I., & Lajoie, S. P. (2016). Learning by preparing to teach: Fostering self-regulatory processes and achievement during complex mathematics problem solving. *Journal of Educational Psychology*, *108*(4), 474–492. https://doi.org/10.1037/edu0000071

National Research Council. (2013). *Next generation science standards: For states, by states*. The National Academies Press. https://doi.org/10.17226/18290

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, *76*(3), 413–448. https://doi.org/10.3102/00346543076003413

Nestojko, J. F., Bui, D. C., Kornell, N., & Bjork, E. L. (2014). Expecting to teach enhances learning and organization of knowledge in free recall of text passages. *Memory & Cognition*, *42*(7), 1038–1048. https://doi.org/10.3758/s13421-014-0416-z

Newman, R. S., & Goldin, L. (1990). Children's reluctance to seek help with schoolwork. *Journal of Educational Psychology*, *82*(1), 92–100. https://doi.org/10.1037/0022-0663.82.1.92

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

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*(6), 737–759. https://doi.org/10.1037/0033-2909.125.6.737

Novak, J. D. (2005). Results and implications of a 12-year longitudinal study of science concept learning. *Research in Science Education*, *35*(1), 23–40. https://doi.org/10.1007/s11165-004-3431-4

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173469

O'Day, G. M., & Karpicke, J. D. (2021). Comparing and combining retrieval practice and concept mapping. *Journal of Educational Psychology*, *113*(5), 986–997. https://doi.org/10.1037/edu0000486

Osborne, R. J., & Wittrock, M. C. (1983). Learning science: A generative process. *Science Education*, *67*(4), 489–508. https://doi.org/10.1002/sce.3730670406

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Person, N. K., Graesser, A. C., Magliano, J. P., & Kreuz, R. J. (1994). Inferring what the student knows in one-to-one tutoring: The role of student questions and answers. *Learning and Individual Differences*, *6*(2), 205–229. https://doi.org/10.1016/1041-6080(94)90010-8

Renaud, R. D., & Murray, H. G. (2007). The validity of higher-order questions as a process indicator of educational quality. *Research in Higher Education*, *48*(3), 319–351. https://doi.org/10.1007/s11162-006-9028-1

Ribosa, J., & Duran, D. (2022). Do students learn what they teach when generating teaching materials for others? A meta-analysis through the lens of learning by teaching. *Educational Research Review*, *37*, Article 100475. https://doi.org/10.1016/j.edurev.2022.100475

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169. https://doi.org/10.1007/BF01405730

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

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

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition*, *1*(4), 242–248. https://doi.org/10.1016/j.jarmac.2012.09.002

Roelle, J., Froese, L., Krebs, R., Obergassel, N., & Waldeyer, J. (2022). Sequence matters! Retrieval practice before generative learning is more

effective than the reverse order. *Learning and Instruction*, *80*, Article 101634. https://doi.org/10.1016/j.learninstruc.2022.101634

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

Roscoe, R. D. (2014). Self-monitoring and knowledge-building in learning by teaching. *Instructional Science*, *42*(3), 327–351. https://doi.org/10.1007/s11251-013-9283-4

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, *77*(4), 534–574. https://doi.org/10.3102/0034654307309920

Roscoe, R. D., & Chi, M. T. H. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science*, *36*(4), 321–350. https://doi.org/10.1007/s11251-007-9034-5

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

Schaerer, M., Tost, L. P., Huang, L., Gino, F., & Larrick, R. (2018). Advice giving: A subtle pathway to power. *Personality and Social Psychology Bulletin*, *44*(5), 746–761. https://doi.org/10.1177/0146167217746341

Schroeder, N. L., Nesbit, J. C., Anguiano, C. J., & Adesope, O. O. (2018). Studying and constructing concept maps: A meta-analysis. *Educational Psychology Review*, *30*(2), 431–455. https://doi.org/10.1007/s10648-017-9403-9

Schwartz, D. L. (1999). The productive agency that drives collaborative learning. In P. Dillenbourg (Ed.), *Collaborative learning: Cognitive and computational approaches* (pp. 197–218). Pergamon.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. Wiley.

Sinatra, G. M., Kienhues, D., & Hofer, B. K. (2014). Addressing challenges to public understanding of science: Epistemic cognition, motivated reasoning, and conceptual change. *Educational Psychologist*, *49*(2), 123–138. https://doi.org/10.1080/00461520.2014.916216

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Steffe, L. P., & Gale, J. (Eds.). (1995). *Constructivism in education*. Lawrence Erlbaum Associates

Taboada, A., & Guthrie, J. T. (2006). Contributions of student questioning and prior knowledge to construction of knowledge from reading information text. *Journal of Literacy Research*, *38*(1), 1–35. https://doi.org/10.1207/s15548430jlr3801_1

Tawfik, A. A., Graesser, A., Gatewood, J., & Gishbaugher, J. (2020). Role of questions in inquiry-based instruction: Towards a design taxonomy for question-asking and implications for design. *Educational Technology Research and Development*, *68*(2), 653–678. https://doi.org/10.1007/s11423-020-09738-9

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

White, P. (2017). *Developing research questions* (2nd ed.). Palgrave.

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist*, *11*(2), 87–95. https://doi.org/10.1080/00461527409529129

Wittwer, J., Nückles, M., Landmann, N., & Renkl, A. (2010). Can tutors be supported in giving effective explanations? *Journal of Educational Psychology*, *102*(1), 74–89. https://doi.org/10.1037/a0016727

Wong, S. S. H. (2023). Deliberate erring improves far transfer of learning more than errorless elaboration and spotting and correcting others' errors. *Educational Psychology Review*, *35*(1), Article 16. https://doi.org/10.1007/s10648-023-09739-z

Wong, S. S. H., Chen, S., & Lim, S. W. H. (2021). Learning melodic musical intervals: To block or to interleave? *Psychology of Music*, *49*(4), 1027–1046. https://doi.org/10.1177/0305735620922595

Wong, S. S. H., & Lim, S. W. H. (2019a). 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

Wong, S. S. H., & Lim, S. W. H. (2019b). Prevention–permission–promotion: A review of approaches to errors in learning. *Educational Psychologist*, *54*(1), 1–19. https://doi.org/10.1080/00461520.2018.1501693

Wong, S. S. H., & Lim, S. W. H. (2022a). A mind-wandering account of the testing effect: Does context variation matter? *Psychonomic Bulletin & Review*, *29*(1), 220–229. https://doi.org/10.3758/s13423-021-01989-8

Wong, S. S. H., & Lim, S. W. H. (2022b). Deliberate errors promote meaningful learning. *Journal of Educational Psychology*, *114*(8), 1817–1831. https://doi.org/10.1037/edu0000720

Wong, S. S. H., & Lim, S. W. H. (2022c). The derring effect: Deliberate errors enhance learning. *Journal of Experimental Psychology: General*, *151*(1), 25–40. https://doi.org/10.1037/xge0001072

Wong, S. S. H., Low, A. C. M., Kang, S. H. K., & Lim, S. W. H. (2020). Learning music composers' styles: To block or to interleave? *Journal of Research in Music Education*, *68*(2), 156–174. https://doi.org/10.1177/0022429420908312

Wong, S. S. H., Ng, G. J. P., Tempel, T., & Lim, S. W. H. (2019). Retrieval practice enhances analogical problem solving. *Journal of Experimental Education*, *87*(1), 128–138. https://doi.org/10.1080/00220973.2017.1409185

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309

Received November 3, 2022
Revision received December 28, 2022
Accepted February 7, 2023 ■