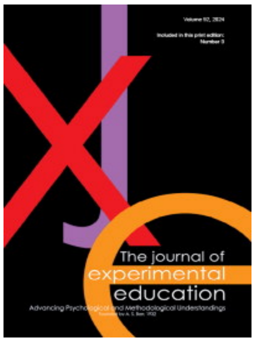

Routledge
Taylor & Francis Group

The Journal of Experimental Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/vjxe20

Durable Benefits of Learning-by-Teaching for Research Question Generation Performance: A Field Experiment

Stephen Wee Hun Lim, Sarah Shi Hui Wong & Piyawan Visessuvanapoom

THE JOURNAL OF EXPERIMENTAL EDUCATION
https://doi.org/10.1080/00220973.2024.2364625

Routledge
Taylor & Francis Group

LEARNING, INSTRUCTION, AND COGNITION

Durable Benefits of Learning-by-Teaching for Research Question Generation Performance: A Field Experiment

Stephen Wee Hun Lim[a,b], Sarah Shi Hui Wong[a,c], and Piyawan Visessuvanapoom[b]

[a]Department of Psychology, National University of Singapore, Singapore; [b]Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University, Bangkok, Thailand; [c]Division of Social Sciences, Yale-NUS College, Singapore

ABSTRACT

Generating good questions is central to scientific inquiry. How can we improve this skill in classrooms? This field experiment showed that teaching others enhances students' ability to generate higher-order research questions that create new knowledge. Whereas *learning-by-teaching* often involves delivering face-to-face or video-recorded lectures, we tested its efficient implementation via writing a verbatim (i.e., word-for-word) teaching script, exactly as how one would orate a lecture. In a research and statistical methods course, 199 undergraduates studied statistical concepts by writing verbatim teaching scripts or study notes. One month later, students' long-term learning was assessed on a high-stakes test, whereby they explained the concepts, applied them to design a study to test a given hypothesis, and generated *create*-level research questions about the concepts. Writing teaching scripts enhanced students' research question generation and concept application more than writing study notes. The teaching advantage for these higher-order outcomes held although both techniques produced comparably high basic understanding when students explained the concepts at test. Further, students displayed greater generative processing, metacognitive monitoring, and social presence when writing teaching scripts than study notes. Learning-by-teaching can be leveraged in an efficient and inexpensive way via writing verbatim teaching scripts to improve meaningful, durable learning in classrooms.

KEYWORDS

Explaining; generative learning; learning by teaching; question generation; field experiment

The formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill. To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science.

—Albert Einstein and Leopold Infeld,

The Evolution of Physics: From Early Concepts to Relativity and Quanta

The ability to ask good questions lies at the heart of scientific inquiry and meaningful learning. In inquiry- or problem-based learning, students engage in scientific discovery processes to construct knowledge (e.g., Barrows & Tamblyn, 1980; Hmelo-Silver, 2004; Pedaste et al., 2015). The inquiry cycle has been viewed as comprising five general inquiry phases: (a) *orientation*—stimulating interest and curiosity in a problem, (b) *conceptualization*—generating research questions

CONTACT Stephen Wee Hun Lim psylimwh@nus.edu.sg Department of Psychology, National University of Singapore, Block AS4, 9 Arts Link, Singapore; Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University, Bangkok, Thailand.

and hypotheses, (c) *investigation*—exploring, experimenting, and interpreting data, (d) *conclusion*—drawing conclusions from data, and (e) *discussion*—communicating conclusions and reflecting on the inquiry process (Pedaste et al., 2015). During the inquiry cycle, students are guided to make sense of a problem in successive iterations by asking questions that identify assumptions to be challenged or knowledge gaps to be filled, toward formulating hypotheses and solutions (Alvesson & Sandberg, 2011; Pedaste et al., 2015; Tawfik et al., 2020). Clearly, the questions that students ask matter—asking good questions catalyzes the inquiry processes that follow in determining what information should be sought and in learning from the newly discovered knowledge (Kedrick et al., 2023; Tawfik et al., 2020).

Unsurprisingly, then, developing students' ability to ask good research questions has been of keen interest to educators and researchers (e.g., Chin & Osborne, 2008; Marbach-Ad & Claassen, 2001), and more broadly emphasized in curricular frameworks such as the *Next Generation Science Standards* (National Research Council, 2013). In the present field experiment, we investigated how the learning technique of teaching others can be efficiently applied in a real-world classroom to enhance students' research question generation performance and long-term learning.

Good research questions: Creating new knowledge

Whereas generating questions can serve as a learning technique to improve recall and comprehension (e.g., Bugg & McDaniel, 2012; Ebersbach et al., 2020; King, 1992, 1994; Weinstein et al., 2010; for a review, see Rosenshine et al., 1996), students' questions are also diagnostic of their deep learning (Chin & Brown, 2002; Chin & Osborne, 2008; Graesser & Olde, 2003). Indeed, the quality of students' questions has been found to positively predict their academic achievement (Graesser & Person, 1994; Harper et al., 2003; Renaud & Murray, 2007).

As classified in questioning hierarchies and aligning with the pinnacle of Bloom's taxonomy (Anderson et al., 2001; Bloom, 1956; see Table A.1 in Appendix), good research questions go beyond existing knowledge to *create* new knowledge (Chin & Brown, 2002; Dillon, 1984; Tawfik et al., 2020), thereby driving scientific discovery. For instance, questions that specify contingencies or causal relations by integrating elements in new ways have been categorized as higher-order questions associated with expert-like reasoning (Dillon, 1984; Keeling et al., 2009; Marbach-Ad & Sokolove, 2000a, 2000b; Tawfik et al., 2020). In contrast, less sophisticated questions tend to be factual with readily available answers or may simply describe or compare various phenomena.

How can we enhance students' ability to ask *create*-level research questions? Given that formulating such questions demands reorganizing and integrating various elements of to-be-learned information, generative learning techniques that promote such cognitive processes may be viable.

Generative learning

Generative learning involves actively constructing meaning by building connections among different elements of incoming information, and relating that information to one's prior knowledge and experience (Osborne & Wittrock, 1983; Wittrock, 1974, 1989). For instance, according to Mayer's (1989, 2014) select–organize–integrate model of meaningful learning, learners must select relevant incoming information, organize the selected information by building structural relations among its elements, and integrate the information with their prior knowledge. Such processes also align with the constructive mode of cognitive engagement in Chi and Wylie's (2014) ICAP framework, in which learners generate additional outputs such as inferences that go beyond the given information.

Of particular interest, one such generative learning technique is *learning-by-teaching*, whereby students take on the role of a "tutor" and teach others to-be-learned material with the intention of helping them learn (for a review, see Fiorella & Mayer, 2016). When teaching, the tutor actively makes sense of the to-be-learned material by engaging in generative processes: selecting relevant information to include in their teaching explanations, organizing this information in coherent structures, and integrating it with their prior knowledge when generating elaborations and inferences (Fiorella & Mayer, 2016; see also Fiorella, 2023).

Learning-by-teaching

Research over the past decades has revealed that teaching others enhances one's own learning of the material (e.g., Bargh & Schul, 1980; Duran & Topping, 2017; for recent meta-analyses, see Kobayashi, 2019, 2024; Leung, 2019; Ribosa & Duran, 2022). As students prepare to teach others, they engage in more metacognitive processing (Muis et al., 2016), and gain further learning benefits by actually teaching the material (Fiorella & Mayer, 2013, 2014).

For instance, Ribosa and Duran's (2022) meta-analysis of 23 articles found an overall effect of 0.17 (standardized mean difference) in favor of creating teaching materials for others over "business-as-usual" or alternative control strategies. More recently, Kobayashi's (2024) meta-analysis of 35 articles examined the interactive effects of preparing to teach (i.e., studying with teaching expectancy) and actually teaching (i.e., communicating knowledge to an audience with the intention of helping them learn). Relative to merely studying without teaching expectancy, teaching after studying with or without teaching expectancy yielded a weighted mean effect of 0.27 (Hedges' g). Importantly, this benefit was moderated by teaching expectancy: Whereas the effect of preparing to teach then actually teaching was medium ($g = 0.48$), the effect of teaching without having prepared to teach did not significantly differ from zero ($g = -0.02$). Moreover, teaching after preparing to teach had a small-to-medium effect ($g = 0.38$) over preparing to teach only. Thus, preparing to teach and actually teaching both uniquely and synergistically contribute to learning-by-teaching effects.

To explain the benefits of teaching, three non-mutually exclusive theoretical accounts have been proposed: (a) the retrieval hypothesis, (b) generative hypothesis, and (c) social presence hypothesis (see Lachner et al., 2022 for a review). First, according to the *retrieval hypothesis* (Koh et al., 2018), tutors engage in substantial retrieval practice when teaching from memory, which improves their durable retention and learning of the taught material (Roediger & Karpicke, 2006). However, when tutors teach the material immediately from memory without having first prepared to teach, retrieval failure (e.g., incomplete and impoverished explanations) may limit any learning benefits that tutors experience (Sibley et al., 2022; see also Kobayashi, 2022 for a discussion).

Second, the *generative hypothesis* posits that teaching others induces sense-making processes such as actively organizing and integrating new information with one's prior knowledge, in turn benefiting the tutor's deep learning and knowledge generalization (Fiorella, 2023; Fiorella & Mayer, 2016). For instance, in reflective knowledge-building, the tutor may generate inferences in their teaching that go beyond the material to produce coherent explanations for their intended audience, while reflecting on and monitoring their own understanding (Roscoe & Chi, 2007). Together, such elaboration and metacognitive processes promote the tutor's content mastery (Roscoe, 2014).

Third and relatedly, the *social presence hypothesis* suggests that perceiving one's audience—whether actual or imagined—as "real" (Kreijns et al., 2022) evokes greater arousal and generative processing on the tutor's part (Hoogerheide et al., 2016, 2019a; Jacob et al., 2020; Lachner et al., 2018). For instance, tutors may adapt or tailor their teaching in anticipation of their intended audience's learning needs (Clark & Brennan, 1991; Nickerson, 1999), such as generating more elaborations to facilitate understanding for a less knowledgeable audience (Wittwer et al., 2010). Thus, even without interacting with their intended audience (i.e., *learning by non-interactive teaching*; see Lachner et al., 2022 for a review), the tutor could reap learning gains from the teaching process, particularly when potential extraneous processing costs (e.g., state anxiety) during teaching are minimized (Cheng et al., 2023; Wang et al., 2023). It should be noted, though, that tutors' self-reported attention to their audience can fluctuate during their teaching, and they may engage in limited knowledge-building even with increased feelings of social presence (Ribosa & Duran, 2023), such that inducing social presence alone does not necessarily improve the tutor's learning (Jacob et al., 2021).

Current empirical evidence for the benefits of learning-by-teaching has largely centered on the tutor's basic recall and comprehension (e.g., Fiorella & Mayer, 2013, 2014; Guerrero & Wiley, 2021; Hoogerheide et al., 2019b; Jacob et al., 2020; Koh et al., 2018; Lim et al., 2021; Nestojko et al., 2014) or transfer of learning to new problems (e.g., Hoogerheide et al., 2014, 2016, 2019a; Lachner et al., 2018, 2021). However, some recent evidence has revealed that learning-by-teaching can also enhance more complex educational outcomes such as the tutor's research question generation. In two lab experiments, Wong et al. (2023) found that across both immediate and 48-hr delayed tests, students who had taught scientific expository texts on natural and social sciences topics by delivering video-recorded lectures successfully generated more *create*-level research questions based on the texts, relative to their peers who had practiced retrieval or constructed concept maps. Furthermore, the learning-by-teaching advantage held even when all three groups similarly received and responded to post-study questions on the material. This suggests that teaching can benefit the tutor's learning even when controlling for engagement with audience questions (Wong et al., 2023). Taken together, extant findings unveil promising unchartered ground to explore whether learning-by-teaching sustains meaningful learning outcomes such as research question generation in real-world education.

Efficient and effective implementation of learning-by-teaching

To transform actual practice, educational research must be effectively translated and contextualized on the ground to be usable for students and teachers. To date, learning-by-teaching has commonly been implemented via peer tutoring (e.g., Roscoe & Chi, 2007, 2008) or delivering video lectures to fictitious others (e.g., Fiorella & Mayer, 2013; Hoogerheide et al., 2014, 2016, 2019a, 2019b). These formats typically involve engaging other students to take on the role of tutee (e.g., in cross- or same-year group tutoring, and cross- or same-year dyadic fixed-role or reciprocal tutoring; see Topping, 1996 for a typology of peer tutoring) and/or utilizing video-recording equipment, which may pose practical barriers for students' independent use of learning-by-teaching. Alternatively, such barriers can potentially be overcome by teaching in written modality, even without interacting with one's audience.

But in recent years, the effectiveness of written teaching relative to oral teaching has been questioned (e.g., Hoogerheide et al., 2016; Jacob et al., 2020; Lachner et al., 2018; see Lachner et al., 2022 for a review). In studies comparing both modalities, learners are typically asked to read or study the to-be-learned material, then explain it to a fictitious audience either aloud while being audio- or video-recorded (i.e., oral explaining) or in writing (i.e., written explaining). The common finding is that oral explaining tends to produce greater learning benefits for the tutor than written explaining (for meta-analyses, see Kobayashi, 2024; Lachner et al., 2021). For instance, Hoogerheide et al. (2016) had participants study a text on syllogistic reasoning, then either restudy it or explain it on video or in writing. On immediate and delayed tests, oral but not written explaining outperformed restudying, although both explaining groups did not significantly differ. Relatedly, when using a relatively more complex (i.e., high element interactivity) text on internal-combustion engines, Lachner et al. (2018) found that oral explaining produced better transfer than written explaining, although both modalities did not differ in their effects on conceptual knowledge (see also Jacob et al., 2020).

Why does the written modality tend to be less effective in learning-by-teaching? One account is that written explaining often induces less social presence and generative processing than oral explaining (e.g., Hoogerheide et al., 2016; Jacob et al., 2020; Lachner et al., 2018). When both modalities trigger similar levels of social presence and generative processing, then written explaining can in fact yield comparable learning outcomes as oral explaining when studying scientific texts (Lim et al., 2021) or scientific concepts in inquiry-based settings (Jacob et al., 2022). In particular, one promising way of effectively implementing written learning-by-teaching is to have students write a verbatim—word-for-word—teaching script, exactly as how they would orate an actual lecture (i.e., "*silent teaching*"; Lim et al., 2021).

"Silent teaching": writing verbatim teaching scripts

As opposed to written essays or expository prose, verbatim teaching scripts constitute "written teaching" in which the tutor purposefully transcribes speech that is originally meant to be communicated as an oral lecture to written text (Lim et al., 2021). In contrast, written instructional explanations are meant precisely to be communicated in written form (e.g., when writing explanations to a fictitious peer in a text editor or even a messenger chat; Jacob et al., 2021). Thus, whereas written instructional explanations tend to draw primarily on written discourse, verbatim teaching scripts are more likely to draw on spoken discourse that is typically associated with conversation or oral teaching (Lim et al., 2021; for discussions of written versus spoken discourse, see Jahandarie, 1999; Sindoni, 2013).

In particular, spoken discourse often reflects a relative focus on interpersonal involvement, whereas written discourse tends to reflect a relative focus on content (Tannen, 1983). This distinction between the relative focus on involvement versus content, rather than oral versus written modalities per se, has been proposed to produce the different features of spoken versus written language (Tannen, 1983, 1985). In other words, communication strategies aimed at building interpersonal involvement in spoken discourse have typically been associated with the oral modality, but can actually be used in discourse in the written modality too for heightened social presence, as when writing a verbatim teaching script.

Indeed, in a recent lab study, Lim et al. (2021) found that writing a verbatim teaching script or teaching orally were both more effective in enhancing the tutor's conceptual knowledge retention than a restudying control. Crucially, both teaching groups did not significantly differ in their learning performance. This suggests that writing a verbatim teaching script is a viable alternative to oral teaching in preserving its effectiveness while efficiently applying the learning-by-teaching strategy in a less resource-intensive way. Of note, silent teaching increased social presence and elaboration to comparable degrees as oral teaching (Lim et al., 2021). For example, the following is an excerpt from a student's verbatim teaching script on the Doppler effect, as reported in Lim et al. (2021):

> Today, we'll be studying the Doppler effect. You might not realize that actually, we witness this effect quite often in our lives. Say, for example, have you ever noticed that a police car with its siren on will sound higher-pitched when it's approaching us and will sound lower-pitched as it leaves us? That's the Doppler effect in action. (p. 1498)

In comparison, the following is an excerpt from the transcript of a student's oral teaching:

> Hi everyone, today we're going to talk about the Doppler effect. For those of you that don't know what the Doppler effect is, maybe you've been on the side of the road while a fire engine passes by and it goes "wewwewwew". Therefore, if you notice, there's something very specific about the sound that's happening and we'll talk about what that is in a bit. (Lim et al., 2021, p. 1498)

Without being told that these excerpts had originally been written versus spoken, respectively, one would likely be hard-pressed to distinguish between them. Thus, when students bear their target audience in mind and engage in generative processing while writing a verbatim teaching script, they could reap learning gains just as they would during oral teaching, even without actually interacting with an audience.

The present field experiment

The present study's goals were twofold. First, this study was conceived as a field experiment to investigate the extent that learning-by-teaching boosts students' research question generation in a real-world classroom. To date, most empirical studies on learning-by-teaching have been conducted in the laboratory (Lachner et al., 2022). Moreover, research on how learning-by-teaching impacts more complex outcomes such as *create*-level question generation remains in its infancy. Thus far, only one study has examined these effects, albeit in a lab setting and with learning-by-teaching implemented in the oral modality via delivering video-recorded lectures (Wong et al., 2023). Hence, much room remains to test whether learning-by-teaching effects transfer to actual educational settings and more complex, higher-order outcomes when efficiently implemented via writing verbatim teaching scripts.

Second, we sought to test whether any benefits of learning-by-teaching persist over longer durations. Study strategies that improve short-term performance do not necessarily produce long-term learning (Soderstrom & Bjork, 2015). Given that education aims to promote students' durable learning rather than mere temporary fluctuations in knowledge, it is pedagogically vital to assess whether the benefits of learning-by-teaching last over time. Although the effects of this technique have been posited to be most evident on immediate higher-order tests, they should theoretically also extend to delayed tests (Roelle & Nückles, 2019; see Kobayashi, 2019 for a meta-analysis) since generative learning activities yield richer mental representations of the material (Fiorella & Mayer, 2016). Accordingly, the present field experiment assessed students' long-term learning on a high-stakes final exam one month after they had completed the intervention. The 1-month delay enabled a relatively more stringent test of students' durable learning, beyond the delays in extant studies typically ranging from a few days (e.g., Hoogerheide et al., 2014, 2019b; Wong et al., 2023) to a week (e.g., Fiorella & Mayer, 2013, 2014; Lim et al., 2021).

Using a within-subjects design, we compared the effects of writing verbatim teaching scripts versus study notes on students' ability to formulate *create*-level questions based on the studied concepts. Note-taking is a popular study technique among students (Blasiman et al., 2017; Miyatsu et al., 2018; Morehead et al., 2019) that has been found to enhance memory and learning (e.g., Di Vesta & Gray, 1972; Kiewra, 1985; Wong & Lim, 2023). By holding constant the explaining modality (written format) across both learning methods, we sought to distill any unique learning benefits from "teaching" per se.

Undergraduate students in an introductory research and statistical methods course received study materials and participated in learning activities that were designed to suit real-world course teaching on research and statistical methods. The teaching team comprised one course instructor who taught the lectures and three graduate teaching assistants who facilitated tutorial sessions. Further to acquiring and applying statistical knowledge, students learned to drive scientific discovery by asking good research questions that create new knowledge. During a tutorial, all students were trained on *create*-level question generation, and practiced writing study notes versus verbatim teaching scripts to ensure fidelity when they later used both learning methods independently. Thereafter, the main experimental intervention took place via a graded learning activity, which was a take-home open-book assignment on writing study notes versus verbatim teaching scripts about statistical concepts. For generalizability purposes, two types of concepts were included: standalone concepts (e.g., "effect size") and pairs of juxtaposed concepts (e.g., "Type I vs. Type II errors").

One month after the learning activity, all students' long-term learning was assessed on a high-stakes final exam, in which they generated as many *create*-level research questions as possible about the concepts that they had earlier studied. Although we were primarily interested in the higher-order outcome of research question generation, for exploratory purposes, we also assessed the learning methods' effects on two other outcomes corresponding to the *understand* and *apply* levels of Bloom's (1956) taxonomy, respectively: (a) students' ability to explain the learned concepts, and (b) apply the concepts to design a study to test a given hypothesis.

We further explored potential processes underlying the learning benefits of teaching: (a) elaboration, (b) metacognitive monitoring, and (c) social presence. Based on the generative hypothesis, tutors elaborate on the material by organizing and integrating it with their prior knowledge when generating high-quality explanations (Fiorella & Mayer, 2016), as evidenced by the number of elaborations (e.g., examples, analogies, and personal experiences that go beyond the material) they produce. Tutors may also self-monitor their understanding of the material while teaching, as evidenced by the number of monitoring statements they produce (e.g., expressing ideas that they do not fully understand; Roscoe & Chi, 2007, 2008; see also Fukaya, 2013). Accordingly, we scored the number of elaborations and monitoring statements in students' teaching scripts and study notes (for similar approaches, see Fiorella & Kuhlmann, 2020; Lachner et al., 2020; Roscoe, 2014; Roscoe & Chi, 2008). We expected that students would produce more elaborations and monitoring statements (i.e., greater knowledge-building) when writing verbatim teaching scripts than study notes.

Additionally, based on the social presence hypothesis, perceiving one's audience as real may stimulate learners' arousal and engagement during teaching (Hoogerheide et al., 2016, 2019a; Jacob et al., 2020; Lachner et al., 2018). As a behavioral proxy of the degree of social presence induced by each learning method, we measured the frequency of audience-directed utterances (i.e., self-other referential terms such as "I", "me", "you", "us", "let's", "our", "we", "your", "yourself") in students' teaching scripts and study notes (for similar approaches, see Chafe, 1982; Hoogerheide et al., 2016; Lachner et al., 2018; Redeker, 1984). We expected that students would use self-other referential terms more frequently when writing verbatim teaching scripts than study notes, implicating higher social presence.

Method

Participants

The participants were 199 undergraduates (150 were female, 46 were male, 3 undisclosed) aged between 18 and 25 ($M = 20.54$, $SD = 1.46$) enrolled in an introductory course on research and statistical methods for psychology at the National University of Singapore. Students who had previously read any similar courses—gained relevant prior knowledge about the course content—were precluded by the university's course allocation system. A power analysis (G*Power; Faul et al., 2007) indicated that this sample size afforded sufficient sensitivity to detect small within-subjects effects of $d \geq 0.20$ for two-tailed pairwise comparisons at 80% power and $\alpha = .05$.

Design and materials

The independent variable was learning strategy (within-subjects), whereby each student wrote study notes about some concepts (*study notes* condition), and verbatim teaching scripts about other concepts (*silent teaching* condition). Four concepts were prescribed: "effect size", "statistical power", "Type I versus Type II errors", and "*z*-tests versus *t*-tests". The former two concepts were standalone (basic) concepts, whereas the latter two concepts were pairs of juxtaposed (advanced) concepts that require comparing and contrasting. Two equivalent concept list versions were created by counterbalancing the prescribed concepts across the lists (see Table 1), ascertaining that

Table 1. Concept List Versions.

Concept List Version	Learning Strategy			
	Study Notes		Teaching	
1	Statistical power	*z*-tests versus *t*-tests	Effect size	Type I versus Type II errors
2	Effect size	Type I versus Type II errors	Statistical power	*z*-tests versus *t*-tests

learning benefits, if any, generalized across concepts; students were each randomly assigned to study either list.

The dependent variable of main interest was students' research question generation performance on a delayed high-stakes test 1 month later, as assessed by the number of *create*-level questions they generated relating to the concepts that they had studied. Students' ability to explain and apply the prescribed concepts was also assessed on the test.

Procedure

During the course, students were introduced to foundational concepts and skills for evaluating and conducting empirical psychological research at a level appropriate for aspiring psychology majors. Specifically, they surveyed and solved research design issues and problems, and applied basic descriptive and inferential statistical techniques. Across a 13-week semester, students attended a 90-min lecture weekly, and five 90-min tutorial sessions bi-weekly.

Training on research question generation and learning methods

Throughout the semester during their lectures and tutorials, students were consistently exposed to the idea and examples of "good" versus "poor" research questions. In line with the *create* level of Bloom's (1956) taxonomy, students were instructed that "good" research questions comprise those that can potentially generate new knowledge, ideas, or perspectives (see Table A.1 in Appendix). During Tutorial 3, all students were explicitly guided by their teaching assistant in developing their own research questions. Students shared their questions individually, receiving immediate feedback on the extent that they qualified as good research questions and, if they did not, further opportunities to improve them until they did. This served as training to ensure that all students understood what was required of them in generating *create*-level research questions.

During the tutorial session, students also practiced typing study notes versus teaching scripts via the following pair of activities: (a) "Write study notes on the concept of *third variables*. In supporting your writing, you may wish to use at least one specific example."; (b) "Teach the concept of *third variables* to your tutor and your fellow course mates. In supporting your teaching preparations as well as actual teaching, you may wish to use at least one specific example." These activities ensured fidelity when students subsequently used both learning methods independently during the experimental intervention.

Writing study notes versus teaching scripts

The main experimental intervention took place via a graded written learning activity—a take-home open-book assignment relating to writing study notes versus teaching scripts about the prescribed concepts. The activity comprised two equivalent versions, with each student randomly assigned to attempt one version. Students embarked on this activity after they had attended Tutorial 3, and had a month to complete it.[1]

Across both conditions, specific instructions were prescribed so that students clearly understood the learning goals. Using no more than 300 words for each concept, each student wrote (a) study notes on "statistical power" and "*z*-tests versus *t*-tests" (Version 1) or "effect size" and "Type

I versus Type II errors" (Version 2) as how they would personally word it when engaging in independent learning (*study notes* condition), and (b) verbatim (word-for-word) teaching scripts on "effect size" and "Type I versus Type II errors" (Version 1) or "statistical power" and "*z*-tests versus *t*-tests" (Version 2) as how they would exactly teach it orally to their peers in an introductory psychology class but who have yet to take a research and statistical methods course for psychology (*silent teaching* condition). Across both conditions, students were allowed to include pictorial illustrations and/or examples of the concepts as deemed fit. The open-book nature of this take-home assignment also meant that all students were allowed to refer to their course material when writing their study notes and verbatim teaching scripts. Thus, unlike closed-book learning activities, there was little basis for students to retrieve the information from memory when completing the intervention. This allowed us to distill the unique effects of "teaching" from those of retrieval practice (e.g., when teaching from memory; Koh et al., 2018; Sibley et al., 2022).

After completing the learning activity, students rated the difficulty of each prescribed concept on a 7-point scale (1 = *very easy*; 7 = *very difficult*). Three days after all student submissions had been received, the course instructor provided feedback on the prescribed concepts at a live lecture. Students' learning performance was assessed based on how accurately they explained the prescribed concepts in their study notes and teaching scripts.

Research question generation test

One month after completing the learning activity, students' long-term learning was assessed in a 2-hr timed final exam at the end of the semester. Each student generated as many *create*-level research questions as possible that could potentially create new knowledge, ideas, or perspectives about each of the four concepts that they had earlier studied in the learning activity. Students were told that there was no need to "answer their own questions".

Concept explanation and application tests

On the final exam, students were also tested on their ability to explain and apply the prescribed concepts, respectively. Given the following hypothesis: "Learners with high hope will demonstrate higher academic achievement than learners with low hope.", students were required to first explain each of the four concepts in the specific context of the hypothesis, and then apply the four concepts in designing a study to test the hypothesis.

Scoring procedure

Two independent raters scored 50 of the 199 (25%) test scripts for students' research question generation, concept explanation, and concept application performance, as well as students' study notes and teaching scripts for elaboration, metacognitive monitoring, and perceived social presence. Discrepancies were reviewed and resolved to reach 100% agreement. As interrater reliability was high across all scoring elements, the remaining scripts were scored by one rater.

Research question generation performance

One point was awarded for each *create*-level research question that students generated at test. Specifically, a question was scored as a *create*-level question if its answer required creating new knowledge, ideas, or perspectives by compiling information in a different way, combining elements in a new pattern, or proposing alternative solutions (Wong et al., 2023; see Table A.1 in Appendix). Sample *create*-level questions were: "How can we efficiently detect a small effect?", "How can we accurately determine statistical power for radically new areas of research?", "How can we derive a set of rubrics for weighing Type I versus Type II errors?", and "Is it possible to devise a universal t-test (like the universal z-test)?". In contrast, sample non-*create* questions that did not receive points were: "Are there other ways of measuring effect size?", "How can we ensure statistical power is high enough so as to correctly reject the null hypothesis?", "Should z-tests and t-tests be used in combination?", and "How can both Type I and Type II errors be minimized at the same time?". There was high interrater reliability in scoring the number of *create*-level research questions generated, absolute agreement intraclass correlation coefficient (ICC) = .94, 95% CI [.92, .95], based on a two-way random-effects model.

In addition, both raters scored the total number of questions that students attempted to generate on the exam, including questions that actually fulfilled the *create* level and those that did not. Interrater reliability was high, ICC = .998, 95% CI [.998, .999].

Concept explanation and application performance

Students' concept explanation performance was scored by awarding one point for each concept that they correctly explained in the context of the test question's given hypothesis. Likewise, students' concept application performance was scored by awarding one point for each concept that they correctly applied when designing a study to test the given hypothesis. Hence, if a response accurately explained a concept but did not effectively apply it to test the hypothesis specified in the question, it would receive one point for the concept explanation test but no points for the concept application test. For instance, to earn both points for "effect size", a student would not only have explained this concept as the extent that high-hope versus low-hope learners differ in academic achievement, but also applied "effect size" in designing a study to test the hypothesis—proposing to observe and report the differential academic achievement as represented by Cohen's d upon using an appropriate statistical tool (e.g., an independent-samples t-test where trait hope is a categorical variable and academic achievement is a continuous variable). There was high interrater reliability when scoring students' concept explanation and application performance, ICC = .95 and .96, 95% CI [.94, .97] and [.95, .97], respectively.

Elaborations, monitoring statements, and social presence

For exploratory purposes, we also scored the number of (a) elaborations, (b) monitoring statements, and (c) self-other referential terms in students' study notes and teaching scripts. First, the number of elaborations served as a behavioral indicator of students' generative processing, operationalized as idea units that had not been covered in the study material, including examples, analogies, and personal experiences (e.g., Fiorella & Kuhlmann, 2020; Jacob et al., 2020; Lachner et al., 2018, 2020). A sample elaboration is: "Suppose eating milk chocolate increases happiness by 2.167 *SD*s while dark chocolate 12.82 *SD*s—dark chocolate has an effect six times bigger than milk chocolate!". Interrater reliability was high when scoring students' elaborations, ICC = .97, 95% CI [.96, .98].

Second, the number of monitoring statements served as an indicator of students' metacognitive processing (Roscoe, 2014). A monitoring statement was defined as an instance where students reflected on and controlled their own thought processes or extent of understanding. A sample monitoring statement is: "This is a hard concept, let me explain it again to make sure everyone follows". There was high interrater reliability when scoring students' monitoring statements, ICC = .97, 95% CI [.96, .97].

Third, as a proxy for social presence, the frequency of self-other referential terms (e.g., "I", "me", "you", "us", "let's", "our", "we", "your", "yourself") was scored as a percentage of the total number of words in students' study notes or teaching scripts (e.g., Chafe, 1982; Hoogerheide et al., 2016; Lachner et al., 2018; Redeker, 1984). Interrater reliability was excellent, ICC = 1.00.

Results

Learning outcomes

Total questions generated

Because all students were instructed to generate as many *create*-level research questions as possible on the final exam, we ascertained that the total number of questions (including questions that actually fulfilled the *create* level and those that did not) that students attempted to generate did not significantly differ for concepts for which they had written study notes ($M = 6.23$, $SD = 2.37$) versus teaching scripts ($M = 6.28$, $SD = 2.42$), $t(198) = 0.42$, $p = .68$, $d = 0.03$, 95% CI [−0.11, 0.17]. Thus, any differences in students' generation of, specifically, *create*-level questions on the exam cannot be attributed to more (or fewer) attempts in any particular learning condition.

Research question generation performance

Analyzing the number of *create*-level questions that students generated on the exam, a learning-by-teaching advantage emerged across both standalone concepts and pairs of juxtaposed concepts (Figure 1A). As predicted, students generated more *create*-level research questions on standalone concepts for which they had written teaching scripts ($M = 1.26$, $SD = 0.48$) than study notes ($M = 1.15$, $SD = 0.42$), $t(198) = 3.06$, $p = .003$, $d = 0.22$, 95% CI [0.08, 0.36]. Similarly, for pairs of juxtaposed concepts, silent teaching ($M = 1.29$, $SD = 0.47$) produced better research question generation performance than writing study notes ($M = 1.15$, $SD = 0.37$), $t(198) = 3.59$, $p < .001$, $d = 0.26$, 95% CI [0.11, 0.40].

Concept application performance

Likewise, learning-by-teaching enhanced students' application of the concepts to design a research study that tests a given hypothesis (Figure 1B). For standalone concepts, writing verbatim teaching scripts ($M = .80$, $SD = .40$) produced better application performance than writing study notes ($M = .75$, $SD = .43$), $t(198) = 2.39$, $p = .02$, $d = 0.17$, 95% CI [0.03, 0.31]. For pairs of juxtaposed concepts, silent teaching ($M = .73$, $SD = .45$) also improved application performance more than writing study notes ($M = .61$, $SD = .50$), $t(198) = 2.63$, $p = .01$, $d = 0.19$, 95% CI [0.05, 0.33].

Concept explanation performance

Interestingly, both learning methods did not differ in their effects on students' concept explanation performance. For standalone concepts, there was no significant difference in students' explanation performance across the silent teaching ($M = .98$, $SD = .14$) and study notes ($M = .97$, $SD = .16$) conditions, $t(198) = 0.45$, $p = .66$, $d = 0.03$, 95% CI [−0.11, 0.17]. Similarly, students' performance did not differ when explaining the pairs of juxtaposed concepts for which they had written verbatim teaching scripts ($M = .96$, $SD = .20$) versus study notes ($M = .95$, $SD = .22$), $t(198) = 0.50$, $p = .62$, $d = 0.04$, 95% CI [−0.10, 0.17]. Thus, students' weaker research question generation and application performance in the study notes condition was not because they failed to understand or explain the concepts per se. Indeed, students demonstrated high levels of understanding of all concepts across both learning conditions (Figure 1C).

Teaching processes

Elaborations

Across both types of concepts, silent teaching induced greater generative processing than writing study notes. For standalone concepts, students' teaching scripts ($M = 1.41$, $SD = 0.87$) contained more elaborations than their study notes ($M = 0.79$, $SD = 0.76$), $t(198) = 8.74$, $p < .001$, $d = 0.62$,

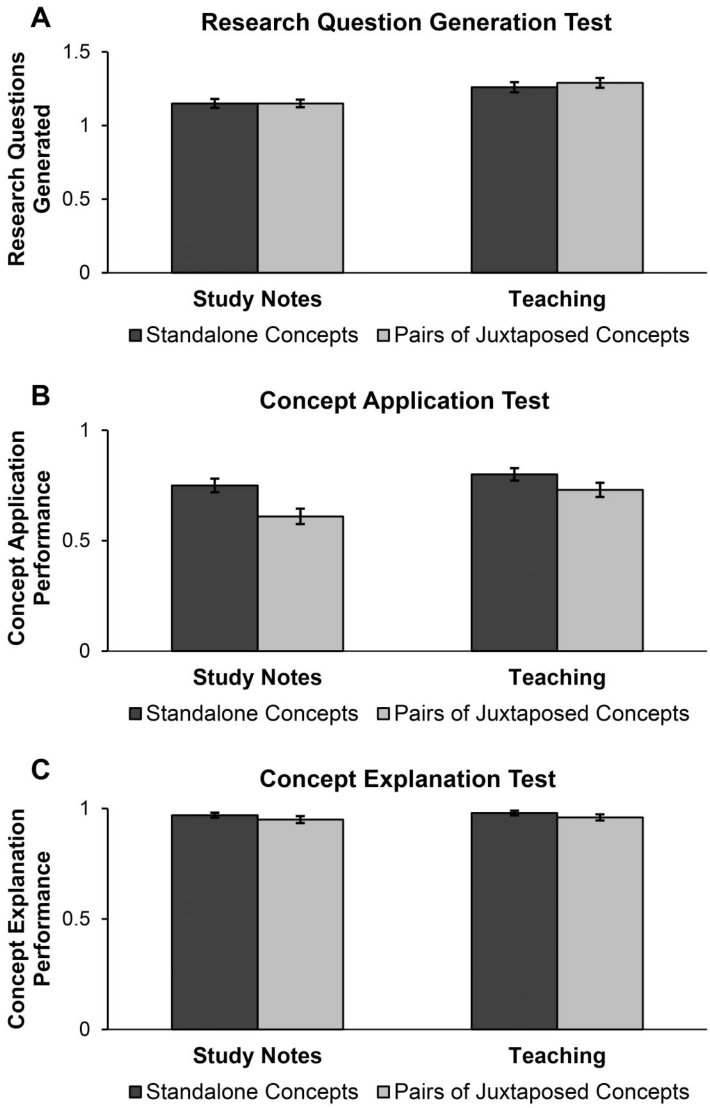

Figure 1. Delayed (1 month) research question generation, concept application, and concept explanation test performance across learning conditions and concept types.
Note. $N = 199$. (**A**), (**B**), and (**C**) show the mean scores for the research question generation test, concept application test, and concept explanation test, respectively. Error bars indicate standard errors.

95% CI [0.47, 0.77]. Similarly, for juxtaposed concepts, students generated more elaborations in their teaching scripts ($M = 1.27$, $SD = 0.96$) than study notes ($M = 0.76$, $SD = 0.88$), $t(198) = 5.86$, $p < .001$, $d = 0.42$, 95% CI [0.27, 0.56].

Metacognitive monitoring

Relative to the study notes condition, silent teaching triggered greater metacognitive processing across both types of concepts. For standalone concepts, students' teaching scripts ($M = 0.37$, $SD = 0.64$) contained more monitoring statements than their study notes ($M = 0.15$, $SD = 0.37$), $t(198) = 4.43$, $p < .001$, $d = 0.31$, 95% CI [0.17, 0.46]. Likewise, for juxtaposed concepts, students produced more monitoring statements in their teaching scripts ($M = 0.13$, $SD = 0.42$) than study notes ($M = 0.06$, $SD = 0.23$), $t(198) = 2.26$, $p = .03$, $d = 0.16$, 95% CI [0.02, 0.30].

Social presence

Silent teaching evoked significantly higher levels of social presence than writing study notes across both concept types. For standalone concepts, students' teaching scripts ($M = 2.01\%$, $SD = 1.67\%$) contained a greater percentage of self-other referential terms than their study notes ($M = 0.68\%$, $SD = 1.10\%$), $t(198) = 10.30$, $p < .001$, $d = 0.73$, 95% CI [0.57, 0.89]. Likewise, for juxtaposed concepts, students used a greater percentage of self-other referential terms in their teaching scripts ($M = 2.32\%$, $SD = 2.22\%$) than study notes ($M = 1.01\%$, $SD = 3.12\%$), $t(198) = 5.08$, $p < .001$, $d = 0.36$, 95% CI [0.22, 0.50].

Concept difficulty ratings

Finally, we ascertained that students' perceived difficulty of the concepts did not differ across learning conditions. There was no significant difference in students' ratings of the standalone concepts' difficulty in the silent teaching ($M = 4.28$, $SD = 1.38$) versus study notes ($M = 4.19$, $SD = 1.43$) conditions, $t(198) = 0.86$, $p = .39$, $d = 0.06$, 95% CI [−0.08, 0.20]. Similarly, students' ratings of how difficult the pairs of juxtaposed concepts were did not differ across the silent teaching ($M = 3.51$, $SD = 1.45$) and study notes ($M = 3.63$, $SD = 1.53$) conditions, $t(198) = -0.93$, $p = .35$, $d = -0.07$, 95% CI [−0.21, 0.07].

Discussion

Asking good research questions is vital for scientific inquiry and discovery. In a well-powered field experiment, we found that learning-by-teaching enhanced students' research question generation performance in an undergraduate research and statistical methods course. When tested after a 1-month delay, students successfully generated more *create*-level research questions relating to statistical concepts for which they had written verbatim teaching scripts than study notes. This advantage of learning-by-teaching held reliably across both standalone concepts and pairs of juxtaposed concepts. In addition, writing teaching scripts improved students' application of the learned concepts to design a study that tests a given hypothesis, relative to writing study notes.

It is worth noting that students' poorer research question generation and concept application performance in the study notes condition was not due to a weaker understanding of the concepts. Indeed, we found that across both learning conditions, students not only rated the concepts as similarly difficult (or easy), but also displayed comparably high levels of understanding when asked to explain the learned concepts on the final exam. These findings demonstrate that study techniques that support basic learning outcomes such as remembering or understanding information may not necessarily yield the same benefits for more complex, higher-order outcomes such as applying or creating new knowledge (e.g., Agarwal, 2019; Wong et al., 2023).

Taken together, our data yield new theoretical and practical insights for extant learning-by-teaching research. First, our results add to the nascent empirical evidence that teaching others can improve more complex learning outcomes such as research question generation (Wong et al., 2023), beyond basic recall and comprehension or transfer of learning to new problems. This suggests that learning-by-teaching is a relatively potent strategy for enhancing deep learning in which students actively construct meaning (Fiorella & Mayer, 2016), toward generating their own research questions that create new knowledge. More broadly, these findings attest to the benefits of generating teaching materials for others (Ribosa & Duran, 2022) and learning-by-teaching even without interacting with one's audience (Lachner et al., 2022).

Second, whereas previous lab studies had observed a benefit of delivering video-recorded lectures for the tutor's research question generation on immediate and 48-hr delayed tests (Wong et al., 2023), the present field experiment has shown that the learning-by-teaching benefit persists even in a real-world classroom when students write verbatim teaching scripts and when the test is administered after a substantially longer delay of 1 month. From an educational stance, it is vital to examine whether learning techniques that are effective in the lab actually work in the classroom, and how they can be viably implemented with authentic educational materials (Roediger, 2013). The present study does so by offering an example of how learning-by-teaching can be translated in an efficient and inexpensive way via writing verbatim teaching scripts to improve meaningful, lasting learning in a university course. It is particularly encouraging that the benefits observed here relate to long-term learning, suggesting that silent teaching supports durable learning rather than merely short-term performance.

Educational and theoretical implications

In distilling the unique effects of "teaching", our data reveal that writing verbatim teaching scripts confers long-term learning benefits over writing study notes for higher-order outcomes such as research question generation and knowledge application. Overriding the need for logistical resources or others to take on the role of tutee when applying learning-by-teaching in the classroom, students may now simply write a full verbatim teaching script to learn effectively (Lim et al., 2021).

Why did silent teaching outperform writing study notes? In the present course, students were allowed to refer to their course material when writing their study notes and verbatim teaching scripts. As there was little, if at all any, basis for students to mentally retrieve and rehearse the material during the intervention, the retrieval hypothesis for the learning benefits of teaching (Koh et al., 2018; see also Kobayashi, 2022) is unlikely to apply in this instance. Rather, teaching is an intentional, generative process. During teaching preparation, the tutor presumably contemplates what material to include or exclude, and how best to organize and present the selected material (Fiorella, 2023; Fiorella & Mayer, 2016). The interpretation is that, as our students taught material to others, they initiated and maintained generative processing. In particular, they reasoned with the material and elaborated on it whilst teaching, producing more elaborations. They also reflected on the material and monitored their self-understanding, producing more monitoring statements in their teaching scripts than study notes. Altogether, such elaborative and metacognitive processes (reflective knowledge-building; Roscoe & Chi, 2007) could have helped students make better sense of the material (Roscoe, 2014).

Moreover, our students demonstrated more audience-directed utterances when writing teaching scripts than study notes, implicating stronger feelings of social presence. When directing their teaching to an audience that is perceived as real, the tutor may adapt their teaching to suit their audience's learning needs (e.g., offering more elaborations for a less knowledgeable audience; Wittwer et al., 2010), thus triggering greater generative processing and physiological arousal (for discussions, see Hoogerheide et al., 2016; Lachner et al., 2022) that benefit the tutor's learning. In principle, a combination of these processes that are unique to the intentionality of teaching others—purportedly different from those involved in producing egocentric content in study notes for one's own learning—may boost durable, meaningful learning.

Interestingly, though, extant studies have often observed no benefit of writing instructional explanations over restudying (Hoogerheide et al., 2016) or self-explaining (Lachner et al., 2021) or retrieval practice (Jacob et al., 2021; Lachner et al., 2021). This lack of effect has commonly been attributed to lower feelings of social presence and generative processing in written explaining, relative to oral explaining (Hoogerheide et al., 2016; Jacob et al., 2020; Lachner et al., 2018). In contrast, the present field experiment found that writing verbatim teaching scripts enhanced higher-order research question generation and concept application performance more than writing study notes, although both techniques did not differ in their effects on basic understanding.

Put together, these findings suggest that there could be important qualitative differences between writing instructional explanations versus verbatim teaching scripts. In particular, whereas written instructional explanations tend to adopt written discourse (e.g., expository prose) that has a relative focus on content, verbatim teaching scripts transcribe speech to text and are thus more likely to also adopt spoken discourse (e.g., conversation) that has a relative focus on interpersonal involvement (Lim et al., 2021; see also Tannen, 1983, 1985). Consequently, verbatim teaching scripts may induce higher levels of social presence and generative processing for the tutor's better learning. Indeed, in their lab study, Lim et al. (2021) found that writing verbatim teaching scripts was just as effective as oral teaching in increasing the tutor's social presence, elaboration, and learning, relative to a restudying control. Directly comparing the effects of writing verbatim teaching scripts versus instructional explanations, as well as their underlying mechanisms, presents an intriguing prospect for future laboratory work.

Limitations and future directions

The observed effect sizes ($d = 0.22$ and 0.26 for research question generation relating to standalone concepts vs. pairs of juxtaposed concepts, respectively) in the present field experiment might seem modest by traditional standards, but based on recently proposed reclassified benchmarks for effect sizes in educational interventions (Kraft, 2020), can be considered "large" relative to the median effect size ($d = 0.17$) for interventions using narrow measures of achievement outcomes. Moreover, we note that the effect sizes here were based on comparisons with an active learning control (writing study notes) as opposed to passive or "business-as-usual" learning methods that are known to be less effective (e.g., restudying). Indeed, recent meta-analyses have observed smaller benefits of learning-by-teaching when comparing this technique against relatively sophisticated control strategies that are known or expected to be effective (e.g., other generative activities, retrieval practice) than "business-as-usual" control strategies (Kobayashi, 2024; Ribosa & Duran, 2022).

The present field experiment adopted a within-subjects design, wherein all students were trained in and used both learning methods. They wrote study notes/teaching scripts for all concepts concurrently during the intervention (take-home open-book assignment), and were tested on all concepts concurrently at the final exam 1 month later. Given the nature of this field study in an actual university course with implications for students' grades, it was not ethically feasible to randomly assign half of the participants to write study notes whereas the other half wrote teaching scripts (i.e., use a between-subjects design). Otherwise, this could have inadvertently disadvantaged some students. While carryover effects are theoretically possible in a within-subjects design, it should be noted that the pairing of concepts with learning methods had been fully counterbalanced. Also, given the high-stakes nature of the final exam, there were no reasonable grounds to expect that students would intentionally try to learn some prescribed concepts better but not others. In other words, students would likely want to learn everything to their best abilities, independently of the specific learning methods prescribed for the various concepts.

Whereas the present research tested the learning benefits of teaching for students, it would be worthwhile to investigate the extent that these benefits apply to teachers too in the context of teacher education. According to the National Research Council (1997), teachers often develop their preferred methods of teaching and learning by experimenting with ways of teaching and observing student reactions. In going beyond one's own experiences, though, the science of education literature provides ideas for student learning and, by logical extension, the teacher's teaching. In view that teaching one's peers in turn improves the student's own understanding of the material, teaching other teachers or teacher trainees how to teach (e.g., how to revise one's presentation of a concept to students, or adapt a particular teaching strategy to become one's own in the classroom) could in turn improve the teacher's own teaching skills. Future research ought to test this hypothesis directly. As teachers prepare to teach, actually teach, and interact with their students or colleagues, they gain learning opportunities too. In this way, teachers may not only contribute to pedagogical support, but also become models of the teachable apprentice who is committed to continual development (see Duran & Topping, 2017 for a discussion).

Finally, the present field study subsumes under our broader goal of translating educational science into usable practice at scale. Drawing inspiration from a now-classic *backward design* philosophy (Wiggins & McTighe, 2005), educators would contemplate (a) what the desired student learning outcome is—in this case, the ability to ask good research questions, (b) what constitutes acceptable evidence of learning—do the research questions create new knowledge?, and (c) what literature-informed instructional activities might serve the learning best—teaching-based learning activities. We then established an inner circle of like-minded educators to promote this evidence-based educational philosophy, inviting the wider teaching and learning communities at our universities to do the same. Most recently, we reached out to national, regional, and international teaching communities beyond the university in calling for a translational educational science globally. Going forward, we are hopeful that more real-world applications of the learning sciences and their reports are underway.

Note

1. Submissions were received from six students two weeks in advance of the due date, nine students one week in advance, and the remaining 184 students during the last week.

Authors' contribution

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

Disclosure statement

No potential conflict of interest was reported by the authors. The authors declare no competing interests, and thank Esther Clare Xin Yi Ng and Sze Chi Lee for their invaluable assistance with data scoring.

Funding

This research was conducted with ethics approval from the Institutional Review Board of the National University of Singapore, and in part supported by a Chulalongkorn University Visiting Scholar Grant (Academic Year 2019, 2022) awarded to Stephen Wee Hun Lim.

ORCID

Stephen Wee Hun Lim http://orcid.org/0000-0003-3636-7587
Sarah Shi Hui Wong http://orcid.org/0000-0003-4243-212X
Piyawan Visessuvanapoom http://orcid.org/0000-0002-4378-8198

References

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
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, Inc.

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.

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: Comparing intended versus actual study behaviour. *Memory*, *25*(6), 784–792. https://doi.org/10.1080/09658211.2016.1221974

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

Bugg, J. M., & McDaniel, M. A. (2012). Selective benefits of question self-generation and answering for remembering expository text. *Journal of Educational Psychology*, *104*(4), 922–931. https://doi.org/10.1037/a0028661

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 35–53). Ablex.

Cheng, M., Wang, F., & Mayer, R. E. (2023). Benefits of asking students to make an instructional video of a multimedia lesson: Clarifying the learning-by-teaching hypothesis. *Journal of Computer Assisted Learning*, *39*(5), 1636–1651. https://doi.org/10.1111/jcal.12823

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

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

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

Di Vesta, F. J., & Gray, G. S. (1972). Listening and note taking. *Journal of Educational Psychology*, *63*(1), 8–14. https://doi.org/10.1037/h0032243

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

Duran, D., & Topping, K. (2017). *Learning by teaching: Evidence-based strategies to enhance learning in the classroom.* Routledge. https://doi.org/10.4324/9781315649047

Ebersbach, M., Feierabend, M., & Nazari, K. B. B. (2020). Comparing the effects of generating questions, testing, and restudying on students' long-term recall in university learning. *Applied Cognitive Psychology*, *34*(3), 724–736. https://doi.org/10.1002/acp.3639

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. (2023). Making sense of generative learning. *Educational Psychology Review*, *35*(2), Article 50. https://doi.org/10.1007/s10648-023-09769-7

Fiorella, L., & Kuhlmann, S. (2020). Creating drawings enhances learning by teaching. *Journal of Educational Psychology*, *112*(4), 811–822. https://doi.org/10.1037/edu0000392

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

Fukaya, T. (2013). Explanation generation, not explanation expectancy, improves metacomprehension accuracy. *Metacognition and Learning*, *8*(1), 1–18. https://doi.org/10.1007/s11409-012-9093-0

Graesser, A. C., & Olde, B. A. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*, *95*(3), 524–536. https://doi.org/10.1037/0022-0663.95.3.524

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

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

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

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., Renkl, A., Fiorella, L., Paas, F., & van Gog, T. (2019a). Enhancing example-based learning: Teaching on video increases arousal and improves problem-solving performance. *Journal of Educational Psychology*, *111*(1), 45–56. https://doi.org/10.1037/edu0000272
Hoogerheide, V., Visee, J., Lachner, A., & van Gog, T. (2019b). 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
Jacob, L., Lachner, A., & Scheiter, K. (2021). Does increasing social presence enhance the effectiveness of writing explanations? *PLoS ONE*, *16*(4), Article e0250406. https://doi.org/10.1371/journal.pone.0250406
Jacob, L., Lachner, A., & Scheiter, K. (2022). Do school students' academic self-concept and prior knowledge constrain the effectiveness of generating technology-mediated explanations? *Computers & Education*, *182*, Article 104469. https://doi.org/10.1016/j.compedu.2022.104469
Jahandarie, K. (1999). *Spoken and written discourse: A multi-disciplinary perspective*. Ablex Publishing Corporation.
Kedrick, K., Schrater, P., & Koutstaal, W. (2023). The multifaceted role of self-generated question asking in curiosity-driven learning. *Cognitive Science*, *47*(4), Article e13253. https://doi.org/10.1111/cogs.13253
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
Kiewra, K. A. (1985). Investigating notetaking and review: A depth of processing alternative. *Educational Psychologist*, *20*(1), 23–32. https://doi.org/10.1207/s15326985ep2001_4
King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal*, *29*(2), 303–323. https://doi.org/10.3102/00028312029002303
King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*, *31*(2), 338–368. https://doi.org/10.2307/1163313
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
Kobayashi, K. (2022). The retrieval practice hypothesis in research on learning by teaching: Current status and challenges. *Frontiers in Psychology*, *13*, Article 842668. https://doi.org/10.3389/fpsyg.2022.842668
Kobayashi, K. (2024). Interactive learning effects of preparing to teach and teaching: A meta-analytic approach. *Educational Psychology Review*, *36*(1), Article 26. https://doi.org/10.1007/s10648-024-09871-4
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
Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798
Kreijns, K., Xu, K., & Weidlich, J. (2022). Social presence: Conceptualization and measurement. *Educational Psychology Review*, *34*(1), 139–170. https://doi.org/10.1007/s10648-021-09623-8
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
Lachner, A., Ly, K.-T., & Nückles, M. (2018). Providing written or oral explanations? Differential effects of the modality of explaining on students' conceptual learning and transfer. *The Journal of Experimental Education*, *86*(3), 344–361. https://doi.org/10.1080/00220973.2017.1363691
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

Marbach-Ad, G., & Claassen, L. A. (2001). Improving students' questions in inquiry labs. *The American Biology Teacher*, *63*(6), 410–419. https://doi.org/10.2307/4451147

Marbach-Ad, G., & Sokolove, P. G. (2000a). 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

Marbach-Ad, G., & Sokolove, P. G. (2000b). Good science begins with good questions: Answering the need for high-level questions in science. *Journal of College Science Teaching*, *30*(3), 192–195.

Mayer, R. E. (1989). Models for understanding. *Review of Educational Research*, *59*(1), 43–64. https://doi.org/10.3102/00346543059001043

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

Miyatsu, T., Nguyen, K., & McDaniel, M. A. (2018). Five popular study strategies: Their pitfalls and optimal implementations. *Perspectives on Psychological Science*, *13*(3), 390–407. https://doi.org/10.1177/1745691617710510

Morehead, K., Dunlosky, J., Rawson, K. A., Blasiman, R., & Hollis, R. B. (2019). Note-taking habits of 21st Century college students: Implications for student learning, memory, and achievement. *Memory*, *27*(6), 807–819. https://doi.org/10.1080/09658211.2019.1569694

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. (1997). *Science teaching reconsidered: A handbook*. The National Academies Press. https://doi.org/10.17226/5287

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

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

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

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

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

Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes*, *7*(1), 43–55. https://doi.org/10.1080/01638538409544580

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

Ribosa, J., & Duran, D. (2023). Students' feelings of social presence when creating learning-by-teaching educational videos for a potential audience. *International Journal of Educational Research*, *117*, Article 102128. https://doi.org/10.1016/j.ijer.2022.102128

Roediger, H. L. (2013). Applying cognitive psychology to education: Translational educational science. *Psychological Science in the Public Interest*, *14*(1), 1–3. https://doi.org/10.1177/1529100612454415

Roediger, H. L., & Karpicke, J. D. (2006). 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

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

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research*, *66*(2), 181–221. https://doi.org/10.3102/00346543066002181

Sibley, L., Fiorella, L., & Lachner, A. (2022). It's better when I see it: Students benefit more from open-book than closed-book teaching. *Applied Cognitive Psychology*, *36*(6), 1347–1355. https://doi.org/10.1002/acp.4017

Sindoni, M. G. (2013). *Spoken and written discourse in online interactions: A multimodal approach*. Routledge. https://doi.org/10.4324/9780203587935

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

Tannen, D. (1983). Oral and literate strategies in spoken and written discourse. In R. W. Bailey & R. M. Fosheim (Eds.), *Literacy for life: The demand for reading and writing* (pp. 79–96). The Modern Language Association.

Tannen, D. (1985). Relative focus on involvement in oral and written discourse. In D. R. Olson, N. Torrance, & A. Hildyard (Eds.), *Literacy, language, and learning: The nature and consequences of reading and writing* (pp. 124–147). Cambridge University Press.

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

Topping, K. J. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education*, *32*(3), 321–345. https://doi.org/10.1007/BF00138870

Wang, F., Cheng, M., & Mayer, R. E. (2023). Improving learning-by-teaching without audience interaction as a generative learning activity by minimizing the social presence of the audience. *Journal of Educational Psychology*, *115*(6), 783–797. https://doi.org/10.1037/edu0000801

Weinstein, Y., McDermott, K. B., & Roediger, H. L. (2010). A comparison of study strategies for passages: Rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied*, *16*(3), 308–316. https://doi.org/10.1037/a0020992

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). Association for Supervision and Curriculum Development.

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

Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist*, *24*(4), 345–376. https://doi.org/10.1207/s15326985ep2404_2

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., Lim, K. Y. L., & Lim, S. W. H. (2023). To ask better questions, teach: Learning-by-teaching enhances research question generation more than retrieval practice and concept-mapping. *Journal of Educational Psychology*, *115*(6), 798–812. https://doi.org/10.1037/edu0000802

Wong, S. S. H., & Lim, S. W. H. (2023). Take notes, not photos: Mind-wandering mediates the impact of note-taking strategies on video-recorded lecture learning performance. *Journal of Experimental Psychology: Applied*, *29*(1), 124–135. https://doi.org/10.1037/xap0000375

Appendix

Table A.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

Note. Reprinted from "To Ask Better Questions, Teach: Learning-by-Teaching Enhances Research Question Generation More Than Retrieval Practice and Concept-Mapping" by S. S. H. Wong, K. Y. L. Lim, and S. W. H. Lim, 2023, *Journal of Educational Psychology*, *115*(6), p. 800. Copyright 2023 by American Psychological Association.