Scalable Science Education via Online Cooperative Questioning

Courtney B. Hilton,[†‡*] Micah B. Goldwater,[‡] Dale Hancock,[§] Matthew Clemson,[§] Alice Huang,[§] and Gareth Denyer[§]

[†]Department of Psychology, Harvard University, Cambridge, MA, 02138; [‡]School of Psychology and [§]School of Life and Environmental Sciences, University of Sydney, Sydney, NSW 2006, Australia

ABSTRACT

A critical goal for science education is to design and implement learning activities that develop a deep conceptual understanding, are engaging for students, and are scalable for large classes or those with few resources. Approaches based on peer learning and online technologies show promise for scalability but often lack a grounding in cognitive learning principles relating to conceptual understanding. Here, we present a novel design for combining these elements in a principled way. The design centers on having students author multiple-choice questions for their peers using the online platform PeerWise, where beneficial forms of cognitive engagement are encouraged via a series of supporting activities. We evaluated an implementation of this design within a cohort of 632 students in an undergraduate biochemistry course. Our results show a robust relationship between the quality of question authoring and relevant learning outcomes, even after controlling for the confounding influence of prior grades. We conclude by discussing practical and theoretical implications.

INTRODUCTION

The path to developing scientific expertise is not simply an accumulation of knowledge, but a change in how knowledge is structured and applied. This change is reflected in a shift from a focus on the superficial aspects of problems and phenomena to the deeper structure or principles that foster making connections across disparate exemplars and contexts (Chi *et al.*, 1981; Goldwater and Schalk, 2016). For example, organic chemistry experts recognize when reactions share common mechanisms despite diverse constituent molecules. Novices, on the other hand, focus on the specific features of the molecules themselves and often fail to notice these deeper relationships (Galloway *et al.*, 2018).

This aspect of the development of expertise may be most familiar to educators and learning designers through Bloom's enduringly popular taxonomy of educational objectives (Bloom *et al.*, 1956). This taxonomy, and its subsequent revision (Anderson and Krathwohl, 2001), lay out a hierarchy of verbal descriptors for the sorts of behavior that reflect these cognitive changes as expertise develops. The base level describes the ability to *remember* facts and procedures—to simply accumulate potentially "undigested" knowledge. Remembering is important, of course, but it does not demonstrate understanding. To reach the higher Bloom levels—to *understand*, to *apply*, to *analyze*, to *evaluate*, and to *create*—one must increasingly integrate different sorts of knowledge, building coherent connections between facts, procedures, concepts, and experiences: enacting deeper changes to how knowledge is structured and applied.

How can we help students to develop such expert-like knowledge and thinking, and to do so at scale? In this paper, we report a collaboration between cognitive scientists (authors C.B.H. and M.B.G.) and life science educators (authors D.H., M.C., A.H., G.D.) aimed at developing methods to achieve this. Applying principles derived from

Ido Davidesco, *Monitoring Editor*

Submitted Nov 22, 2019; Revised Nov 4, 2021; Accepted Nov 16, 2021

CBE Life Sci Educ March 1, 2022 21:ar4

DOI:10.1187/cbe.19-11-0249

*Address correspondence to: Courtney B. Hilton (courtneyhilton@g.harvard.edu).

the cognitive science of learning, along with educational technologies, we design and evaluate a novel learning activity based upon scaffolded and collaborative authoring of multiple-choice questions (MCQs), applied to scalably foster deep conceptual knowledge in the domain of tertiary-level biochemistry. We end by enumerating the benefits of such a collaborative approach, both for advancing practical educational outcomes and for ensuring more generalizable and practically relevant theories in the sciences of learning.

Evidence-based Principles for Effective Learning

Application of cognitive science to education has developed several evidence-based principles for how to design effective learning experiences. One synthesis of these principles is the ICAP framework (Chi and Wylie, 2014), whose acronym describes a hierarchy of modes of cognitive engagement in terms of their benefit to learning—interactive > constructive > active > passive. Complementary to Bloom's taxonomy, which describes the *goals* or *ends* of learning, the ICAP framework describes the *processes* or *means* of learning and knowledge change; it claims that higher learning outcomes (e.g., higher levels of Bloom's taxonomy) are best supported by drawing more frequently upon higher levels of cognitive engagement.

Here, we make the case that a series of activities centered around *authoring* and *answering* MCQs, embedded within supporting social and technical systems, can effectively draw upon all levels of cognitive engagement described in the ICAP framework and can do so in a coherent and scalable learning design. We start by reviewing these evidence-based principles in the context of answering and authoring MCQs.

Answering MCQs as an Active Form of Learning

One of the simplest and most effective learning principles is the "testing effect": after initial study, *actively testing* yourself on the same material leads to better long-term memory than *passively rereading* those materials (Karpicke and Roediger, 2008; Kornell *et al.*, 2009; Richland *et al.*, 2009; Huelser and Metcalfe, 2012; Bjork *et al.*, 2013; Rowland, 2014). As classically described by William James (1890, p. 646):

> A curious peculiarity of our memory is that things are impressed better by active than by passive repetition … it pays better to wait and recollect by an effort within, than to look at the book again. If we recover the words the former way, we shall probably know them the next time; if in the latter way, we shall likely need the book once more.

The crucial insight is that the process of retrieving information from memory is itself a distinct learning event that modifies existing memories. Human memory is quite unlike computer memory in this respect: we can check whether a file is saved to a computer as many times as we want, but this will not change its contents or the ease of its retrieval. This is counterintuitive for many learners and educators, who often have metacognitive misconceptions about what study practices are effective for learning. Students often endorse the opposite of what is effective, preferring activities that give the illusion of fluency (e.g., rereading notes or primary sources) over those that actually challenge and improve their understanding (e.g., self-testing; Bjork *et al.*, 2013)—a challenge for which Robert Bjork has coined the term "desirable difficulties."

In addition to boosting memory consolidation, testing also serves to identify knowledge gaps. Educators can use this to provide targeted explanation and feedback for their students (Metcalfe, 2017). But perhaps more importantly, learners can use this information themselves to self-regulate their own studying. The extent and sophistication of self-regulation is one of the most reliable predictors of student success generally (recent meta-analysis: Dent and Koenka, 2016). An important element of self-regulated learning is performance monitoring (e.g., identifying knowledge gaps), for which regular testing can be an invaluable tool (Butler and Winne, 1995; Tanner, 2012; Panadero, 2017).

While testing is typically thought of as a means of consolidating *previously learned* material, it can also potentiate future learning of *new* material—the "pre-testing effect" (Kornell *et al.*, 2009; Richland *et al.*, 2009; Grimaldi and Karpicke, 2012; Little and Bjork, 2016). That is, instead of giving a lecture and then testing students on the lectured content to boost their memory (although this may still be a good idea), sometimes it may be more effective to test students on this content before the lecture. This actively engages students in thinking about what they do know about a new topic and particularly what they do not. The gaps identified can then be more deliberately amended in the following lecture.

While the most frequently cited benefit of testing is on memory (the bottom level of Bloom's taxonomy), testing can also improve spontaneous knowledge transfer and application (the higher Bloom levels)—helping learners develop deeper understanding by changing how existing knowledge is structured and applied. The importance of deeper understanding of life sciences concepts, going beyond rote memorization, is underscored by the 2009 *Vision and Change* report on undergraduate biology education, stating (p. viii-ix):

> Biology in the 21st century requires that undergraduates learn how to integrate concepts across levels of organization and complexity and to synthesize and analyze information that connects conceptual domains.

This deeper understanding can be supported when testing involves elements of "transfer": where understanding of the learning material is extended in some manner (Johnson and Mayer, 2009; Jacoby *et al.*, 2010; McDaniel *et al.*, 2013; Goldwater and Schalk, 2016; Pan and Rickard, 2018). These benefits are less likely to occur when testing involves only superficial recall of previously encountered facts (although, see Rohrer *et al.*, 2010).

In other words, the benefits of testing hinge upon whatever mental processes are tested. If, for example, an MCQ asks, "Which of the following is the correct definition of a competitive inhibitor?," and presents a set of definitions that includes the previously studied textbook definition, then the ability to *recall* this definition and distinguish it from others is improved. But if instead the question asks, "What might account for low phosphatase activity in samples prepared with phosphate-buffered saline?," then the ability to *transfer* this knowledge about competitive inhibitors to new situations (e.g., those involving phosphate-buffered saline) is improved.

This latter type of question can be further broken down into tests of either near or far transfer. When one extends understanding to new examples within the same domain or in contexts similar to the one in which something was learned—this is near transfer. Testing near transfer can improve the ability to apply understanding within these familiar contexts. And indeed, much of how professional knowledge is used amounts to instances of near transfer—applying old knowledge in new but predictable ways. A key part of developing domain expertise involves building a robust network of context-dependent experiences for how such knowledge can be used (National Research Council, 1999; Markauskaite and Goodyear, 2016).

Far transfer occurs when one further extends knowledge to contexts or domains that are superficially unrelated but connect at a deeper, more abstract level. Although more difficult, and not a routine part of professional work, far transfer can support many aspects of the higher Bloom levels that require knowledge to be applied in flexible and novel ways. Practice with far transfer (e.g., through testing) can also have the more general effect of reorganizing conceptual knowledge in terms of more coherent mental networks of facts, concepts, and experiences (Goldwater and Schalk, 2016).

In summary, testing is an effective evidence-based method for reinforcing memories, prompting meta-cognitive reflection, and for restructuring knowledge—it is a method for active learning par excellence. Testing can take many forms, but regular low-stakes quizzes using MCQs represent a simple and familiar option whose beneficial effects reliably translate from lab-based studies to real-world teaching (McDaniel *et al.*, 2011; Agarwal *et al.*, 2012) and can be extended to test deeper aspects of conceptual understanding in many of the ways we described.

Authoring MCQs as a Constructive Form of Learning

Closely related to the testing effect is the "generation effect," which describes the benefit of having learners actively *generate* information as compared with studying the same information by other means (Slamecka and Graf, 1978; Foos *et al.*, 1994). For example, you are more likely to remember a word pair when there are missing elements requiring completion (e.g., "cat | d__") than when reading a complete version (e.g., "cat | dog").

The mechanisms underlying these benefits are still debated, but one important aspect appears to be how generation draws attention to *relations between elements* (McCurdy *et al.*, 2020). Actively generating the word "dog" when given the prompt "cat | d__" draws attention to and requires thinking about how cats *relate* to other things: activating a broad range of prior knowledge about cats to infer that the relationship with dogs is the most likely, ruling out other alternatives like "dot" or "dam."

Simple forms of generation like this (i.e., inferring an occluded word) dominate the experimental psychology literature on the generation effect and its specific effects on recall memory (reviewed in Bertsch *et al.*, 2007; McCurdy *et al.*, 2020). But generation can also take more elaborate forms that are relevant to real classroom learning and conceptual understanding. For example, students can generate an outline of a text instead of just reading it; or, as in the present study, students can author MCQs rather than just being tested on them (Kelley *et al.*, 2019).

The key to these more elaborate forms of generation is that they motivate learners to *constructively* make sense of learning materials (Fiorella and Mayer, 2016). That is, inferring an occluded word may be more cognitively active ("cat | d__") than mere reading, but the relations that get drawn to make this inference are still likely to be mostly implicit: not requiring much, if any, explicit reflection. By contrast, the task of authoring a high-quality MCQ is more complex and requires explicit and iterated reflection and engagement with ideas. This complexity arises from the many relations between question components (question stem, answer options, examples used, appropriateness to audience, etc.) that must be harmonized while abiding within certain constraints. Unlike simple forms of generation, therefore, question authoring can motivate deeper forms of cognitive engagement than question answering.

One such form of engagement known to support the development of conceptual understanding is "self-explanation" (Chi, 2000): a constructive activity involving verbalizing (or externalizing by other means) explanations of learning materials and thought processes while completing a task. In the context of question authoring, it may be particularly helpful for students to self-explain *why* they made various design decisions in their questions: *Why* this topic? *Why* this phrasing of the question? *Why* this example? *Why* this set of distractors? This line of questioning helps to explicate their thought processes and knowledge about a topic, making it easier to identify misconceptions or tacitly justified choices that would motivate further study or revisions of their questions. Resolving the conflicts they identify not only helps to produce a better question but crucially fosters their own learning and conceptual change (Chi *et al.*, 1994).

Another strategy that builds on self-explanation and that may particularly help with authoring deep conceptual questions is *analogical comparison*—comparison, in the sense of comparing similarities and differences among two or more examples; analogical, in the sense that examples are superficially contrasting and only relate at a deeper level. For example, glycolysis (the breakdown of glucose) and the synthesis of fatty acids are two superficially contrasting processes (one a breakdown and one a synthesis) but in fact are united in how they are driven by positive feedback interactions (the well-fed state). When comparing such examples that lack *superficial* commonalities, learners are forced to find *abstract* commonalities and, as a result, are biased to notice the shared structural principles often missed by novices (Gentner and Markman, 1997). By getting practice in explicitly identifying these more abstract relationships in diverse contexts, learners become better able to notice them in new examples and contexts (Gentner *et al.*, 2003). As part of a brainstorming process of authoring MCQs, analogical comparisons may be an effective way to develop understanding of the deeper causal patterns that underlie important concepts and processes that are not only relevant within specific subdomains in the life sciences, but that cut across science, technology, engineering, and mathematics domains (Goldwater and Schalk, 2016; Jacobson *et al.*, 2020; Gray and Holyoak, 2021).

A particularly important component of MCQs is the set of distractors that a question answerer must choose between when answering the question. But not just any incorrect answer serves as a distraction. A good distractor distinguishes meaningfully

different levels of understanding—distinguishing full understanding from incomplete or superficial understanding or from a common misconception. For example, for the question "Which of the following contributes to the stability of DNA compared with RNA?," the distractor "super glue" can be easily dismissed, even by somebody without much chemistry knowledge. Contrastingly, the distractor "thymine is more stable than uracil" separates shallow from deep understanding (i.e., DNA does differ from RNA in its use of thymine/uracil as nucleobases, and this is important, but it does not explain their differences in stability). Authoring MCQs, and in particular authoring effective distractors, has the potential to draw attention to the important details that make particular scientific truths true and distinguish them from partial truths or misconceptions. Strategies such as self-explanation and analogical comparison can help with this.

Although the conceptual processes of comparison, self-explanation, and confronting misconceptions can be leveraged while authoring questions, it is also quite possible for students to just query superficial understanding. Students are indeed known to have a superficial "knowledge telling" bias to learning activities like peer tutoring if they are not given additional support (knowledge telling is opposed to a more elaborative "knowledge building" approach: Roscoe and Chi, 2007). And although the basic notion of a MCQ is simple and likely familiar to students, the qualities that make a *high-quality* question are harder to concretely define, and there is no simple recipe for making them. Having students author questions with minimal guidance is therefore unlikely to fully elicit the beneficial cognitive processes described earlier, as has been shown for other learning designs that overwhelm students with task complexity (Kirschner *et al.*, 2006). As such, the question of how to support question authoring for effective learning will be a key part of our study.

In summary, the process of authoring a MCQ constitutes a generative learning activity that not only produces a helpful learning resource (to test oneself on) but can serve as a rich learning experience in its own right. While testing is a clear example of active learning (as opposed to a passive one, like simply listening to a lecture), authoring goes one step further and is an example of a constructive type of cognitive engagement, because it involves generating, manipulating, and combining knowledge in novel ways.

Cooperating via PeerWise as an Interactive Form of Learning

To reach the highest level of the ICAP hierarchy, a socially *interactive* form of learning is needed, wherein learners work cooperatively together to co-construct knowledge. We now describe a simple way to support this in ways that engage both question answering and authoring through an online educational technology called PeerWise.

PeerWise is an online application designed to facilitate students in creating a bank of MCQs for their peers and supporting cycles of question authoring, answering, and discussion (Denny *et al.*, 2008). It has primarily been used in higher education (reviewed in Kay *et al.*, 2020), including in the biochemical sciences (Ryan, 2013; McQueen *et al.*, 2014; Galloway and Burns, 2015; Hancock *et al.*, 2018).

An important element of PeerWise is that it provides a mechanism to scalably provide students with feedback on their question authoring. It does this by encouraging students to rate the quality of their peers' questions on a simple five-point scale. Aggregated ratings of question quality (from students) correlate well with expert ratings (McQueen *et al.*, 2014) and thus provide a way to crowdsource basic feedback. The reliability of student ratings of question quality is consistent with research showing that, although question authoring is difficult, people readily appreciate a good question when they see it (Rothe *et al.*, 2018). In addition to providing feedback on overall question quality, PeerWise also tracks how students answer each question, allowing the effectiveness of distractors to be empirically rather than subjectively assessed, providing further means to curate study or assessment resources (Huang *et al.*, 2021).

When supported over a teaching semester, these dynamics of question authoring, answering, and feedback can eventuate in an online learning community whose shared goal and motivation is to curate relevant and useful study resources for the collective good. Such cooperative "contributor" (de Boer and Collis, 2002; Collis and Moonen, 2006) or "knowledge-building" (Scardamalia and Bereiter, 1994) approaches to learning can be more intrinsically motivating for students, as the outputs of their efforts have both tangible social and practical consequences. This is very different from completing more traditional study assignments, which often at best are only briefly engaged with by an instructor. Indeed, studies of PeerWise usage show increased student engagement (Casey *et al.*, 2014; Biggins *et al.*, 2015; Hancock *et al.*, 2018) with associated gains in learning (McQueen *et al.*, 2014; Kay *et al.*, 2020).

However, the peer question-rating scheme is a sparse form of feedback that may be suboptimal for improving question quality. This may be especially true in large early-year undergraduate cohorts with variable abilities. Some groups have addressed this by extending the question-rating system to use more sophisticated schemes (Bates *et al.*, 2014; Galloway and Burns, 2015). But improving question quality is not just about the resulting questions but the learning that happens while authoring them (Chin and Osborne, 2008). Specifically of interest here: How might *interactive* forms of cognitive engagement, which could benefit learning, be supported in the context of providing peer feedback on MCQ authoring?

Potentially serving this function, PeerWise does allow an additional form of peer feedback via a discussion forum–style comment field under each question, where students can, in principle, provide feedback on question quality, express confusion if they do not understand some aspect of the question, or discuss other aspects of the question or topic more generally. But just as students rarely author high-quality questions without additional support (as discussed in the previous section), it is also unlikely that students will engage in deep and meaningful interactive discussions simply because a comment field is available to them. And despite a growing body of applied educational research on PeerWise usage, there is a lack of research on how to optimize such interactions and their potential learning benefits.

Addressing this gap may be particularly impactful, because prior research on the ICAP framework has found that teachers struggle with implementing interactive learning experiences in particular (Chi *et al.*, 2018). We believe that educational technologies like PeerWise have the potential to make it easier for educators to design and implement learning experiences that

engage active, constructive, and interactive forms of cognitive engagement. Online communication technologies facilitate interactive learning in particular by making it easier to share knowledge and to collaborate asynchronously (for recent meta-analysis on computer-supported collaborative learning, see Chen *et al.*, 2018). This paper explores a novel design for how PeerWise can be used to optimize learning in this way.

METHODS

The following learning design takes the online educational software PeerWise (Denny *et al.*, 2008), with a few small modifications, and embeds it in a series of learning activities designed to encourage beneficial modes of cognitive engagement in the context of answering and authoring MCQs.

Participants

A total of 712 students took part in an introductory biochemistry class at the University of Sydney. Of the 712 students, 634 students both authored questions in PeerWise for course credit and completed the final exam. Data from these 634 students are analyzed in the following analysis (mean age = 19.2; 405 female, 229 male). This study was approved by a Human Research Ethics Committee at the University of Sydney (reference number: 2017/131).

Scaffolding the Use of PeerWise Student Question Online Platform

Our team of biochemistry educators and cognitive science researchers developed iterative exercises to scaffold the question-authoring process. These exercises leverage analogical comparison, self-explanation, and confronting misconceptions. Further, embedded within peer collaboration and instructor feedback facilitated through PeerWise (Denny *et al.*, 2008), the students are guided through the beneficial knowledge-building processes that result in the previously described learning gains and are most likely to improve their ability to craft questions that query deeper levels of knowledge.

Over the course of the semester, there were five steps to this process; each step was 2 weeks apart. The students were all assigned a learning outcome as a target for their questions. There was a list of learning outcomes assigned to 712 students (634 students are in the analysis; see *Participants* section). These were assigned randomly by an instructor to ensure approximately equal distribution of outcomes across the cohort. This resulted in a study bank for students that covered the content of the exam.

Step 1 was for each student to write two true-false statements (a false example: "Chemicals such as AZT are able to terminate growing viral DNA chains because DNA polymerases have an inherent proofreading function."[1]) and post to PeerWise on the first few lectures of the class. This was to practice writing clear concise statements covering one concept. The students did not get individualized feedback on their true-false statements. Feedback with good and bad examples was given to the entire cohort (in the step 3 tutorial) to illustrate how to write clear options for a MCQ.

[1]This statement is false. AZT is a modified nucleotide that can act as a substrate for the viral DNA polymerase. Because it lacks a 3′-OH group, it blocks elongation of the DNA chain, preventing viral replication. This blocks the polymerization function (but is not related to the proofreading function) of DNA polymerase.

Step 2 was to go onto the Web and find a MCQ (on a particular allocated topic within molecular biology) and post on PeerWise. Students were then told to answer at least 40 from the bank of more than 600 questions and were given bonus points for answering more than 40 over the course of the semester. Crucially, students then had to evaluate their chosen MCQs and make some suggestions for improvements. We used the number of questions answered here as a predictor variable in our analyses. This allowed us to directly contrast the role of using a bank of questions as a study aid from contributing an original question to the bank.

Step 3 required the students to individually author a MCQ on an allocated learning outcome within the domain of molecular biology. In addition to their practice with authoring a true-false statement (step 1), and with evaluating other MCQs on the Internet (step 2), they had a tutorial session (led by authors D.H. and C.B.H.) on how to effectively author questions that involved a lecture on Bloom's taxonomy and an explanation of how engaging in reflection during questioning can both help one write better questions and boost one's own learning.

The specific instructions for authoring questions were to:

1. Look up a learning outcome on the syllabus and find the relevant lecture and online material.
2. Write down at least five statements about this topic.
3. Start by writing an MCQ from these statements and share this with a peer. Examples were presented using true-false statements from step 2.
4. Then think about how to present the question with experimental data, information about a mutation, inhibitor, predictor, etc.

Instructions indicated that this was only to get them started. To then write a conceptually deep and effective question, they were to follow further (generative) steps that all involve considering how to apply an understanding of biochemical mechanisms to research or problem-solving contexts:

1. Consider how the information was first discovered.
2. Ask yourself: What techniques would have been used or would exploit this information?
3. Would an inhibitor or a mutation produce interesting results you would need this information to explain?
4. In what contexts would this information be particularly applicable?
5. Identify common misconceptions to use as plausible distractors.

The students' MCQs produced in step 3 were evaluated against a modified version of Bloom's taxonomy by an instructor (author D.H.). The details of the scoring system are given in the following section. These data are the primary predictor variable in our main analyses.

Steps 4 and 5 also involved a question-authoring activity, this time for allocated topics in the domain of metabolism. Unlike with step 3, step 4 had students *interactively* work in groups, evaluating and refining their questions. They also got feedback from instructors in between steps 4 and 5. The initial question produced in step 4 (before instructor feedback) was given an authoring score as in step 3, and this was used in our analysis. The refined question at the end of step 5 was submitted for credit in the class (and was marked by the instructors).

The group discussions were scaffolded by the following generative prompts. Like the prompts in step 3, these prompts were designed to further engage the evidence-based methods discussed in the *Introduction*, such as generating explanations and analogies with a focus on molecular causal processes (e.g., the prompt about inhibiting or activating steps), and to encourage considering possible misconceptions.

- What are the facts?
- Why are things the way they are?
- Does anything strike you as strange or noteworthy?
- What didn't you understand about this?
- What don't your peers understand about this?
- How would you explain it to someone else?
- What would happen if a step were inhibited/activated?
- What would happen if a component was in excess or short supply?
- Can you think of an analogy for this?
- Can you build a scenario or context around this?

Examples of the sorts of comments students provided one another as part of this *interactive* cooperative learning activity are provided in the Supplemental Material. Now we turn to the primary analyses of the paper: how the quality of the question the students authored predicts performance on the final exam.

Materials and Scoring

The final exam contained both MCQs and short-answer questions (SAQs). Each of these can be further broken down into questions relating to the topic of molecular biology (30 MCQs; two SAQs; covering 97 learning outcomes) or metabolism (34 MCQs; two SAQs; covering 259 much more granular learning outcomes). The four dependent variables used in our analyses are the average mark for each of these combinations of question type (MCQ or SAQ) and topic (molecular biology or metabolism). The split by topic is particularly important, because steps 2 and 3 of our intervention were targeting molecular biology concepts (individual authoring), and steps 4 and 5 were targeting those for metabolism (collaborative authoring). This allows us to investigate the more interactive mode of question authoring in steps 4 and 5 (where students collaboratively worked on questions in groups) and the more individual (merely constructive) mode of question authoring in step 3 separately, although the comparison between them is confounded by differences in topic, so we do not focus on this (see *Discussion*).

A.H. and D.H. analyzed the exam questions for their Bloom levels as well. For the molecular biology MCQs, the average Bloom level was 3.05, and for the metabolism ones, it was 3.09. For the SAQs, there were only two questions for each topic, with each containing aspects testing Bloom levels of at least 5 and 6 (the questions had multiple parts, some of which asked about lower-level aspects to establish a baseline to build up to the high-level aspects). Thus, performance on the short-answer section in particular will likely reflect deeper aspects of conceptual understanding that go beyond superficial memorization of terminology and procedures.

Our key predictor variable for analyzing the step 3 data (corresponding to the molecular biology topics on the exam) was an assessment of the MCQ that each student authored. Every question was scored on a 0–3 continuous scale by one of two instructors (D.H. & M.C.). This score reflected both the quality of the question stem and the answer choices; the scheme was developed by D.H. and represents a coarsened version of Bloom's taxonomy. To assess the reliability of the question scoring by the instructors, 63 randomly selected questions (~10% of total) were doubled marked. These pairs of scores had high consistency (intraclass correlation coefficient = 0.841, $p < 0.001$), indicating that the scoring procedure was reliable.

- <1.5 was given if the question stem was confusing, options were ambiguous, or there was more than one right answer.
- 1.5–1.8 was assigned to a straightforward but clear question stem and four to five answer options that were unambiguous with one correct option: level 1 or 2 of Bloom's taxonomy.
- 1.8–2.0 was given if the question included some context or more thoughtful answer options: levels 2 to 3 of Bloom's taxonomy.
- 2.0–2.5 was given if there were greater experimental context; more creative answer options, including plausible distractors; and common misconceptions: levels 3 to 4 of Bloom's taxonomy.
- 2.5–3.0 was assigned for well-described relevant experimental context in the question, clear options with very plausible distractors, and concepts relevant to key learning outcomes in the class being covered: levels 4 to 5 Bloom's taxonomy.

Our key predictor variable for the step 4 (i.e., before receiving instructor feedback) analysis was a combined cooperative question-authoring score, taking into account both the quality of the question produced on a metabolism topic, as well as the number of meaningful comments each student provided for peers' questions. This score is intended to reflect the degree of cognitive engagement arising from both question authoring and providing constructive peer feedback. We chose to analyze these aspects in combined form, rather than separately, because we believe that these components of step 4 function as a holistic activity, wherein it would be difficult (if not impossible) to disentangle their causal influences.

Hypotheses

Our primary hypothesis is that students who engage more deeply in the question-authoring process will do better on the final exam. In the step 3 analysis, this is operationalized as the question-authoring score and its relation to the molecular biology sections of the final exam; and in step 4, the cooperative question-authoring score and its relation to the metabolism sections of the final exam.

One problem, however, in attributing this putative relationship to the effect of question authoring is that it is confounded by the prior abilities of the students. Higher-performing students may be both more likely to author higher-quality questions, engage in more constructive feedback with peers, and ultimately to do better on the final exam, regardless of any additional learning benefits accrued through our design. Performance in the molecular biology section of the exam is also confounded by the number of PeerWise MCQs (which were all on molecular biology topics) that students tested themselves on in step 2. Thus, we also hypothesized that this relationship would hold after statistically controlling for these other factors in ways we now describe.

TABLE 1. Marginal structural model predicting exam mark (molecular biology, MCQ section) from question-authoring score, with effect modification from number of questions self-tested

	Estimate	SE	*t* value	*p* value
Intercept	36.434	3.741	9.739	<0.001***
Question-authoring score	6.545	1.658	3.948	<0.001***
Number of questions self-tested	0.036	0.026	1.356	0.176

***$p < 0.001$.

Statistical Analysis

We accounted for the confounding relationships described in the previous section using inverse-probability weighting (Austin and Stuart, 2015), using stabilized weights to account for continuous (rather than categorical) treatment exposure (Naimi *et al.*, 2014). This statistical procedure is common in epidemiological research in which causal patterns are to be identified from observational data in which there are confounding relationships like the ones described here. In short, this procedure estimates a scalar weight value for each observation (each student, in our case), such that the confounding relationships are removed (i.e., after applying weights, removing any correlation between prior grades and question-authoring quality).

Students' prior grades were defined as the average mark for each student over the eight first-year units (the unit described in this study is a second-year unit).

The number of tested questions was taken from each student's usage of PeerWise during step 2 of the activities described in the previous section. The median number of answered questions was 51, and the majority clustered around this number, but there was also a long tail of the distribution, with one student answering as many as 489 questions, and some students at the other extreme who answered only a handful of questions. It seems implausible that there would be a linear relationship across this whole range (of 0–489 tested questions), thus, we applied a winsorization approach to attenuate the influence of outliers. This worked by replacing values outside the 5th and 95th quantiles with the values at these quantiles (see Supplemental Figure 1).

Before the inverse-probability weighting procedure was applied, the data were highly unbalanced, especially with respect to the students' prior grades (i.e., prior grades were indeed correlated with authored-question quality and engagement). After this procedure was applied, acceptable covariate balance was achieved (i.e., statistically removing this relationship; see Supplemental Figure 1). Before using this weighted data set for analyses, we trimmed extreme weights beyond the 99th percentile. Extreme (outlier) weight values can cause artificially inflated standard errors and indicate violations to the positivity assumption. Using this final weighted data set allows us to estimate the effect of authored-question quality on exam performance and to interpret this as not just being a reflection of these other confounding factors.

We then regressed the question quality ratings on exam performance. As a result of the applied weights, these models can be interpreted as *marginal structural models* (Robins *et al.*, 2010)—marginal, because the parameter estimates are no longer conditional on confounders (it marginalizes over them); structural, because instead of modeling *observed outcomes* (the original data), it models *potential outcomes* (the weighted data). We additionally included the number of questions answered as an effect modifier in the model. This serves to determine 1) whether there is also a testing effect, and if there is, 2) whether this effect is distinct from the authoring effect.

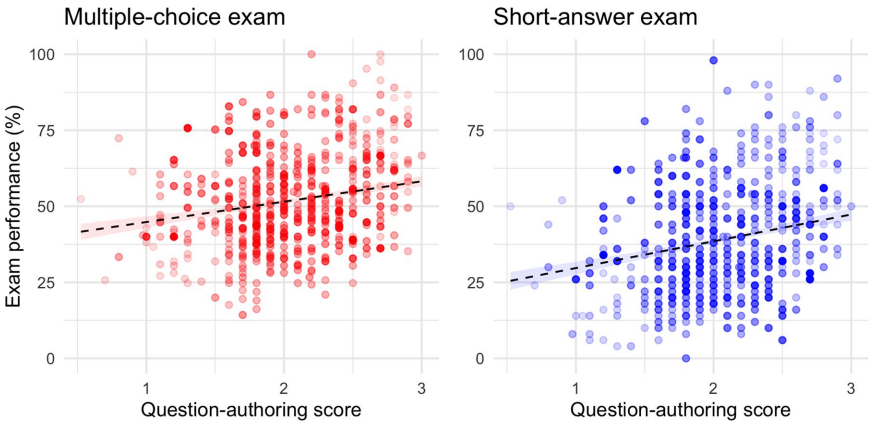

FIGURE 1. Results in the molecular biology section of the exam (corresponding to the solo question-authoring activity in step 3) for the multiple-choice (left) and short-answer (right) sections. The *y*-axis on both plots is the percent grade for this section of the exam, the *x*-axis is the question-authoring score. Each point represents a student, and the transparency of the point represents the estimated weight used in the inverse-probability weighting analysis used to control for confounding (i.e., the more transparent, the less that student's score influences the result). The dotted lines represent linear regressions of the marginal effect of the question-authoring score on exam performance (i.e., controlling for the influence of prior grades), with shaded regions representing the standard error of the mean.

RESULTS

The results show significant effects of question authoring on exam performance, in line with our hypotheses. Starting with the step 3 (molecular biology) results, for the multiple-choice section of the exam (Table 1 and Figure 1), a person who scores a zero on the authored question (i.e., did not complete it) and does not answer any PeerWise questions has a predicted average mark of 36.4% (for reference, the average mark across the whole cohort for this section of the exam is 51.0%). Holding the number of questions answered constant, for each unit increase in the question quality score (up to a maximum score of 3), there is a predicted 6.5% improvement in the mark (i.e., a predicted 19.6% boost in exam performance if the

TABLE 2. Marginal structural model predicting exam mark (molecular biology, SAQ section) from question-authoring score, with effect modification from number of questions self-tested

	Estimate	SE	*t* value	*p* value
Intercept	17.828	4.121	4.326	<0.001***
Question-authoring score	8.542	1.869	4.570	<0.001***
Number of questions self-tested	0.064	0.032	2.001	0.046*

*$p < 0.05$.
***$p < 0.001$.

question scores a 3). There was no statistically significant effect of question answering on exam performance in the multiple-choice section.

For the short-answer section (Table 2 and Figure 1), the predicted mark for those who did not author or answer a question was 17.8% (the average mark in the cohort for this section was 38.1%). Holding answering constant, each unit increase in authored-question quality brought a predicted 8.5% increase in exam performance (i.e., accruing up to a maximum boost of 25.5%) and, accordingly, a 0.06% increase for each question used for self-testing (with a predicted 3.4% boost overall if the median 51 questions were answered).

For step 4 (metabolism topics), there was a similar trend (Tables 3 and 4 and Figure 2). For deeper levels of engagement in the cooperative question-authoring task, there was a predicted 7.8% increase for each rise in level of engagement for the multiple-choice exam mark, and a 9.7% increase in the mark on the short-answer section of the exam.

DISCUSSION

A large class of students participated in a cooperative online community of MCQ authoring and answering as part of a tertiary biochemistry class. Their participation was supported through the software PeerWise (Denny *et al.*, 2008) and by learning activities that we designed in light of the cognitive principles discussed in the *Introduction*. Our results show that, within this context, both answering and authoring questions had beneficial and independent effects on student performance, with authoring having the largest effect, consistent with the predictions of the ICAP framework (Chi and Wylie, 2014). These effects not only extended to doing better on the multiple-choice section of the final exam, but also to doing better in the short-answer section, which tested a deeper conceptual understanding of the learning materials and required actively generating explanations of phenomena (as opposed to just selecting the correct multiple-choice option).

TABLE 3. Marginal structural model predicting exam mark (metabolism, MCQ section) from cooperative question-authoring score

	Estimate	SE	*t* value	*p* value
Intercept	34.365	2.564	13.405	<0.001***
Question-authoring score	7.825	1.187	6.593	<0.001***

***$p < 0.001$.

TABLE 4. Marginal structural model predicting exam mark (metabolism, SAQ section) from cooperative question-authoring score

	Estimate	SE	*t* value	*p* value
Intercept	20.472	3.249	6.300	<0.001***
Question-authoring score	9.684	1.499	6.459	<0.001***

***$p < 0.001$.

The large effect of question authoring is critical. While the testing effect is a well-known and frequently cited finding in cognitive and educational research, in its most typical uses, the focus is on more superficial kinds of knowledge tested just with memory recall. Our results emphasize that question authoring is a learning activity that pairs naturally with the benefits of testing but additionally leverages deep and elaborative cognitive processes that, with the aid of tools like PeerWise, can be readily implemented at scale in large tertiary science classes to support conceptual understanding.

In characterizing this authoring effect, the deeper and more meaningful the engagement with question authoring (step 3: higher on Bloom's taxonomy; better distractors; relevant examples; step 4: all this plus cooperative interaction in the form of feedback on one another's questions), the better the student performed on the final exam. This held true even while statistically controlling for prior grades and question-answering activity. This effect was replicated for both the molecular biology and metabolism sections of the exam, each of which corresponded with a separate question-authoring activity (steps 3 and 4, respectively, in our design). These results are consistent with our account, described in the *Introduction*, that question authoring is an engaging task that taps into several cognitive processes widely known to increase learning, such as the generation effect and self-explanation (Chi *et al.*, 1994; Rosner *et al.*, 2013).

It is worth further noting that the students were assigned one of 97 learning outcomes in molecular biology to author a question in step 3, and one of 259 learning outcomes to author a question in steps 4 and 5. So each student only authored two questions (although they gave feedback on other people's questions on different topics in step 4) but saw benefits on exam performance across the board. This speaks to how these elaborative processes use broad conceptual knowledge even when the task is quite focused. Prior work by this research group examining a previous year's cohort (Hancock *et al.*, 2018) specifically analyzed questions by their learning outcomes and showed that students saw just as big a benefit for learning outcomes they were not assigned as the learning outcomes they were assigned. To get this kind of broad conceptual benefit from question answering (or other forms of testing), we suggest that a deep form of conceptual engagement is required.

It is important to emphasize that unscaffolded question authoring is potentially not enough to elicit these broad and deep conceptual benefits. That is, the goal of our educational design was not just to show the positives of PeerWise for

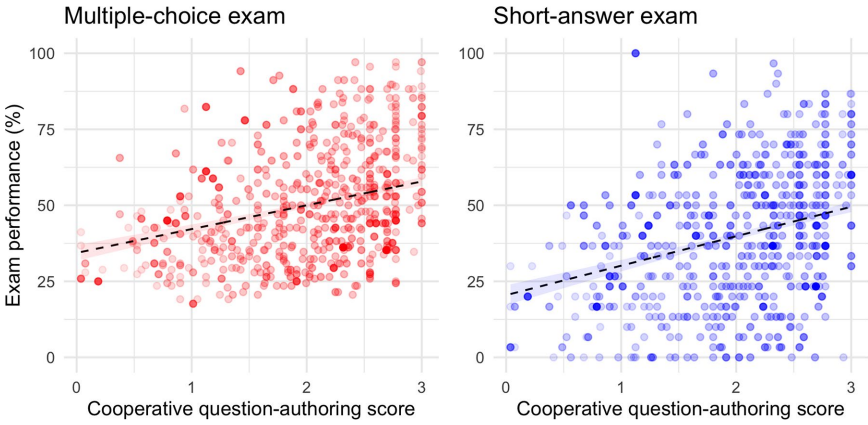

FIGURE 2. Results in the metabolism section of the exam (corresponding to the interactive question-authoring activity in step 4) for the multiple-choice (left) and short-answer (right) sections. The *y*-axis on both plots is the percent grade for this section of the exam, the *x*-axis is the cooperative question-authoring score. Each point represents a student, and the transparency of the point represents the estimated weight used in the inverse-probability weighting analysis used to control for confounding. The dotted lines represent linear regressions of the marginal effect of the cooperative question-authoring score on exam performance (i.e., controlling for the influence of prior grades), with shaded regions representing the standard error of the mean.

question authoring generally, but to outline and test a multistep scaffolding process to increase the quality of engagement in the question-authoring process. Our results show greater learning benefits for those who authored deeper questions and engaged more in peer feedback on question authoring, but it remains for future research to more systematically test and compare different approaches for encouraging deeper forms of engagement.

The comparison between steps 3 (solo authoring of molecular biology questions) and 4 (cooperative authoring of metabolism questions) of our design comes close to providing such a test. However, due to the naturalistic nature of our experiment, we were not able to randomize the ordering of the *interactive* (steps 4 and 5) and *noninteractive* (step 3) authoring activities or the assignment of topic domain (molecular biology or metabolism), among other potential issues. So while there was a numerically larger improvement in exam performance as a function of authoring engagement in step 4 compared with step 3 (and indeed, more students clustered up in the higher end of the distribution of authoring engagement scores), we choose to only interpret this as a promising hint, motivating further research to assess this question more rigorously.

Another limitation of our study is that, although we showed that authoring deeper questions brought more learning gains, we were not able to assess the efficacy of our question-answering and question-authoring design overall in comparison to a reference standard teaching method (such as standard lectures followed by practice exercises). Given the extensive literature concerning the testing effect (Rowland, 2014) and the benefits of generative/constructive forms of engagement (Chi and Wylie, 2014; Fiorella and Mayer, 2016), as discussed in the *Introduction*, we suspect that the specific design used here would be at least as effective as more standard approaches, and quite likely more effective (see especially Chi *et al.*, 2018; Kelley *et al.*, 2019). But once again, more research is needed for firmer conclusions to be made.

Many of these limitations arise from the challenges of doing scientific research outside the lab. Most of the studies we reviewed in our *Introduction* analytically decomposed learning activities into simpler parts such that causality could be more clearly established between components and eventual learning outcomes, typically within well controlled lab environments rather than messy real-world classrooms. However, while this helps achieve certain scientific goals, it does not always support the best practical outcomes for education, where often the best designs will be of a more holistic nature.

More importantly, while it is easier to establish causal connections in the lab, the types of learning that are possible under these restricted conditions may not generalize straightforwardly to real-world contexts. This is increasingly recognized as a serious concern for traditional psychological experiments, which typically study only narrow and unrepresentative groups of people (Henrich *et al.*, 2010), in narrowly constrained contexts (Baribault *et al.*, 2018)—leading to a putative "generalizability crisis" (Hilton and Mehr, 2021; Yarkoni, 2021). These concerns have motivated the need to systematically embed educational experiments within real-world learning contexts (Motz *et al.*, 2018; Fyfe *et al.*, 2021) and for research teams to collaborate deeply with educators in both research and implementation (Penuel *et al.*, 2011)—as we have done here. So while there are challenges in conducting research in real-world teaching contexts, there are potentially equally as many opportunities.

Finally, we emphasize the promise this approach has for scalability. By centering learning activities on having students contribute to a cooperative and intrinsically motivating enterprise (producing study questions that they may themselves benefit from in exam preparation), the effort of each individual student not only benefits personal learning but potentially amplifies that of others. Online applications like PeerWise are specifically designed to take advantage of such self-reinforcing dynamics and to enhance them through design mechanics such as gamification (Indriasari *et al.*, 2020). This approach to educational design, and to understanding contextually situated learning, fits well with recent attempts to conceptualize learning as a complex system (Jacobson *et al.*, 2016, 2019). Achieving large-scale and deep conceptual learning from this perspective need not always require proportional effort, but rather, in our case at least, is about tapping the latent potential of the crowd through well-designed socio-technical systems aligned to the peculiarities of human cognition.

CONCLUSION

It is relatively easy to instruct basic factual knowledge at scale, but scaling up the teaching of deeper conceptual understanding is more challenging. Our paper was aimed at addressing this challenge from both practical and theoretical perspectives. First, from a cognitive science perspective, we reviewed how

and why answering and authoring MCQs benefits learning and specifically when these activities might support deeper forms of knowledge change through tapping deeper forms of cognitive engagement. Applying these principles, we designed a novel series of question-answering and question-authoring activities that were embedded within the free online cooperative tool PeerWise (Denny *et al.*, 2008). We then evaluated the extent to which this system of activities, cooperating students, and technologies scalably supports conceptual understanding. Our initial results were promising, motivating future research to verify and extend our findings and address the limitations. Ultimately, we hope that this work provides a useful design template for educators and motivates more collaborations between cognitive scientists and domain-expert educators.

ACKNOWLEDGMENTS

We would like to thank Paul Denny for help with various aspects relating to PeerWise and Peter Reimann and Judy Kay for several helpful discussions, as well as all the students.

REFERENCES

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review*, *24*(3), 437–448. https://doi.org/10.1007/s10648-012-9210-2

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679. https://doi.org/10.1002/sim.6607

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., … & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences USA*, *115*(11), 2607–2612. https://doi.org/10.1073/pnas.1708285114

Bates, S. P., Galloway, R. K., Riise, J., & Homer, D. (2014). Assessing the quality of a student-generated question repository. *Physical Review Special Topics—Physics Education Research*, *10*(2), 020105. https://doi.org/10.1103/PhysRevSTPER.10.020105

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

Biggins, D., Crowley, E., Bolat, E., Dupac, M., & Dogan, H. (2015). Enhancing university student engagement using online multiple choice questions and answers. *Open Journal of Social Sciences*, *3*(9), 71–76. https://doi.org/10.4236/jss.2015.39011

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *46*, 417–444.

Bloom, B. S., Engehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals*. New York, NY: Longmans.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281.

Casey, M. M., Bates, S. P., Galloway, K. W., Galloway, R. K., Hardy, J. A., Kay, A. E., … & McQueen, H. A. (2014). Scaffolding student engagement via online peer learning. *European Journal of Physics*, *35*(4), 045002.

Chen, J., Wang, M., Kirschner, P. A., & Tsai, C.-C. (2018). The role of collaboration, computer use, learning environments, and supporting strategies in CSCL: A meta-analysis. *Review of Educational Research*, *88*(6), 799–843.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual process of generating inferences and repairing mental models. In Glaser, R. (Ed.), *Advances in Instructional Psychology* (pp. 161–238). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Chi, M. T. H., Adams, J., Bogusch, E. B., Bruchok, C., Kang, S., Lancaster, M., … & Yaghmourian, D. L. (2018). Translating the ICAP theory of cognitive engagement into practice. *Cognitive Science*, *42*(6), 1777–1832. https://doi.org/10.1111/cogs.12626

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2

Chi, M. T. H., Leeuw, N. D., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477.

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., & 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

Collis, B., & Moonen, J. (2006). The contributing student: Learners as co-developers of learning resources for reuse in Web environments. In Hung, D., & Khine, M. S. (Eds.), *Engaged learning with emerging technologies* (pp. 49–67). Dordrecht, Netherlands: Springer-Verlag. https://doi.org/10.1007/1-4020-3669-8_3

de Boer, W., & Collis, B. (2002). A changing pedagogy in e-learning: From acquisition to contribution. *Journal of Computing in Higher Education*, *13*(2), 87–101. https://doi.org/10.1007/BF02940967

Denny, P., Luxton-Reilly, A., & Hamer, J. (2008). The PeerWise system of student contributed assessment questions. *Proceedings of the 10th Australasian Computing Education Conference*, *78*, 6.

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, *28*(3), 425–474. https://doi.org/10.1007/s10648-015-9320-8

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

Foos, P. W., Mora, J. J., & Tkacz, S. (1994). Student study techniques and the generation effect. *Journal of Educational Psychology*, *86*(4), 567–576.

Fyfe, E. R., de Leeuw, J. R., Carvalho, P. F., Goldstone, R. L., Sherman, J., Admiraal, D., … & Motz, B. A. (2021). ManyClasses 1: Assessing the generalizable effect of immediate feedback versus delayed feedback across many college classes. *Advances in Methods and Practices in Psychological Science*, *4*(3), 251524592110275. https://doi.org/10.1177/25152459211027575

Galloway, K. R., Leung, M. W., & Flynn, A. B. (2018). A comparison of how undergraduates, graduate students, and professors organize organic chemistry reactions. *Journal of Chemical Education*, *95*(3), 355–365. https://doi.org/10.1021/acs.jchemed.7b00743

Galloway, K. W., & Burns, S. (2015). Doing it for themselves: Students creating a high quality peer-learning environment. *Chemistry Education Research and Practice*, *16*(1), 82–92. https://doi.org/10.1039/C4RP00209A

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*(2), 393–408. https://doi.org/10.1037/0022-0663.95.2.393

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45–56.

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*(7), 729–757. https://doi.org/10.1037/bul0000043

Gray, M. E., & Holyoak, K. J. (2021). Teaching by analogy: From theory to practice. *Mind, Brain, and Education*, mbe.12288, *15*(2), 1–14. https://doi.org/10.1111/mbe.12288

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513. https://doi.org/10.3758/s13421-011-0174-0

Hancock, D., Hare, N., Denny, P., & Denyer, G. (2018). Improving large class performance and engagement through student-generated question banks: Student-generated question banks. *Biochemistry and Molecular Biology Education*, *46*(4), 306–317. https://doi.org/10.1002/bmb.21119

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hilton, C. B., & Mehr, S. A. (2021). Citizen science can help to alleviate the generalizability crisis. *Behavioral and Brain Sciences*.

Huang, A., Hancock, D., Clemson, M., Yeo, G., Harney, D., Denny, P., & Denyer, G. (2021). Selecting student-authored questions for summative assessments. *Research in Learning Technology*, *25*, 1–25. https://doi.org/10.25304/rlt.v29.2517

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, *40*(4), 514–527. https://doi.org/10.3758/s13421-011-0167-z

Indriasari, T. D., Luxton-Reilly, A., & Denny, P. (2020). Gamification of student peer review in education: A systematic literature review. *Education and Information Technologies*, *25*(6), 5205–5234. https://doi.org/10.1007/s10639-020-10228-x

Jacobson, M. J., Goldwater, M., Markauskaite, L., Lai, P. K., Kapur, M., Roberts, G., & Hilton, C. (2020). Schema abstraction with productive failure and analogical comparison: Learning designs for far across domain transfer. *Learning and Instruction*, *65*, 101222. https://doi.org/10.1016/j.learninstruc.2019.101222

Jacobson, M. J., Kapur, M., & Reimann, P. (2016). Conceptualizing debates in learning and educational research: Toward a complex systems conceptual framework of learning. *Educational Psychologist*, *51*(2), 210–218. https://doi.org/10.1080/00461520.2016.1166963

Jacobson, M. J., Levin, J. A., & Kapur, M. (2019). Education as a complex system: Conceptual and methodological implications. *Educational Researcher*, *48*(2), 112–119. https://doi.org/10.3102/0013189X19826958

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1441–1451. https://doi.org/10.1037/a0020636

James, W. (1890). *The principles of psychology* (Vol. I). Cambridge, MA: Harvard University Press.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*(3), 621–629. https://doi.org/10.1037/a0015183

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kay, A. E., Hardy, J., & Galloway, R. K. (2020). Student use of PeerWise: A multi-institutional, multidisciplinary evaluation. *British Journal of Educational Technology*, *51*(1), 23–35. https://doi.org/10.1111/bjet.12754

Kelley, M. R., Chapman-Orr, E. K., Calkins, S., & Lemke, R. J. (2019). Generation and retrieval practice effects in the classroom using PeerWise. *Teaching of Psychology*, *46*(2), 121–126. https://doi.org/10.1177/0098628319834174

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729

Little, J. L., & Bjork, E. L. (2016). Multiple-choice pretesting potentiates learning of related information. *Memory & Cognition*, *44*(7), 1085–1101. https://doi.org/10.3758/s13421-016-0621-z

Markauskaite, L., & Goodyear, P. (2016). *Epistemic fluency and professional education: Innovation, knowledgeable action and actionable knowledge*. Dordrecht, Netherlands: Springer.

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*, *27*(6), 1139–1165. https://doi.org/10.3758/s13423-020-01762-3

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. https://doi.org/10.1037/a0021782

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams: Quizzing and successful transfer. *Applied Cognitive Psychology*, *27*(3), 360–372. https://doi.org/10.1002/acp.2914

McQueen, H. A., Shields, C., Finnegan, D. J., Higham, J., & Simmen, M. W. (2014). PeerWise provides significant academic benefits to biological science students across diverse learning tasks, but with minimal instructor intervention: Diverse learning benefits of PeerWise for biology students. *Biochemistry and Molecular Biology Education*, *42*(5), 371–381. https://doi.org/10.1002/bmb.20806

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*(1), 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Motz, B. A., Carvalho, P. F., De Leeuw, J. R., & Goldstone, R. L. (2018). Embedding experiments: Staking causal inference in authentic educational contexts. *Journal of Learning Analytics*, *5*(2), 47–52. https://doi.org/10.18608/jla.2018.52.4

Naimi, A. I., Moodie, E. E. M., Auger, N., & Kaufman, J. S. (2014). Constructing inverse probability weights for continuous exposures: A comparison of methods. *Epidemiology*, *25*(2), 292–299. https://doi.org/10.1097/EDE.0000000000000053

National Research Council. (1999). *How people learn: Brain, mind, experience, and school*. National Academy Press.

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

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Penuel, W. R., Fishman, B. J., Haugan Cheng, B., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educational Researcher*, *40*(7), 331–337. https://doi.org/10.3102/0013189X11421826

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. https://doi.org/10.1037/a0016496

Robins, J. M., Hernán, M. A., Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology 11, 550–560. https://doi: 10.1097/00001648-200009000-00011.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233–239. https://doi.org/10.1037/a0017678

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.

Rosner, Z. A., Elman, J. A., & Shimamura, A. P. (2013). The generation effect: Activating broad neural circuits during memory encoding. *Cortex*, *49*(7), 1901–1909. https://doi.org/10.1016/j.cortex.2012.09.009

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89. https://doi.org/10.1007/s42113-018-0005-5

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

Ryan, B. J. (2013). Line up, line up: Using technology to align and enhance peer learning and assessment in a student centred foundation organic chemistry module. *Chemistry Education Research and Practice*, *14*(3), 229–238. https://doi.org/10.1039/C3RP20178C

Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge-building communities. *Journal of the Learning Sciences*, *3*(3), 265–283. https://doi.org/10.1207/s15327809jls0303_3

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604.

Tanner, K. D. (2012). Promoting student metacognition. *CBE—Life Sciences Education*, *11*(2), 113–120. https://doi.org/10.1187/cbe.12-03-0033

Yarkoni, T. (2021). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685