Online and Clicker Quizzing on Jargon Terms Enhances Definition-Focused but Not Conceptually Focused Biology Exam Performance

Steven C. Pan,[†]* James Cooke,[‡] Jeri L. Little,[§] Mark A. McDaniel,[‖] Erin R. Foster,[¶] Lisa Tabor Connor,[#] and Timothy C. Rickard[@]

[†]Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095; [‡]Division of Biological Sciences and [@]Department of Psychology, University of California, San Diego, La Jolla, CA 92093; [§]Department of Psychology, California State University, East Bay, Hayward, CA 94542; [‖]Department of Psychological and Brain Sciences and Center for Integrative Research on Cognition, Learning, and Education, Washington University in St. Louis, St. Louis, MO 63130; [¶]Program in Occupational Therapy, School of Medicine, Washington University in St. Louis, St. Louis, MO 63130; [#]Department of Occupational Therapy, MGH Institute of Health Professions, Boston, MA 02129

ABSTRACT

Mastery of jargon terms is an important part of student learning in biology and other science, technology, engineering, and mathematics domains. In two experiments, we investigated whether prelecture quizzes enhance memory for jargon terms, and whether that enhanced familiarity can facilitate learning of related concepts that are encountered during subsequent lectures and readings. Undergraduate students enrolled in neuroanatomy and physiology courses completed 10-minute low-stakes quizzes with feedback on jargon terms either online (experiment 1) or using in-class clickers (experiment 2). Quizzes occurred before conventional course instruction in which the terms were used. On exams occurring up to 12 weeks later, we observed improved student performance on questions that targeted memory of previously quizzed jargon terms and their definitions relative to questions on terms that were not quizzed. This pattern occurred whether those questions were identical (experiment 1) or different (experiment 2) from those used during quizzing. Benefits of jargon quizzing did not consistently generalize, however, to exam questions that assessed conceptual knowledge but not necessarily jargon knowledge. Overall, this research demonstrates that a brief and easily implemented jargon-quizzing intervention, deliverable via Internet or in-class platforms, can yield substantial improvements in students' course-relevant scientific lexica, but does not necessarily impact conceptual learning.

INTRODUCTION

In many science courses, students learn *jargon terminology*—discipline-specific technical or specialized vocabulary words that are not commonly used in other contexts—in concert with other course content. For example, when learning about the visual system, a student may acquire the meanings of "lateral geniculate body," "occipital lobe," "thalamic nuclei," and many other new words and phrases. There are many such jargon terms across the natural and physical sciences, with biology and its subdisciplines having among the most (at ~2000–17,000 terms per high school textbook, as catalogued by Yager, 1983; Groves, 1995). The process of learning jargon in science courses has been compared with that involved in acquiring a foreign language (e.g., Osborne, 2002), and by some accounts requires even more attention and time (Yager, 1983).

John Coley, *Monitoring Editor*

Submitted Jan 28, 2019; Revised Jun 14, 2019; Accepted Aug 6, 2019

CBE Life Sci Educ December 1, 2019 18:ar54

DOI:10.1187/cbe.18-12-0248

*Address correspondence to: Steven C. Pan (stevencpan@psych.ucla.edu).

Since at least the 1970s, researchers have pinpointed the "terminology problem" (Wandersee, 1988) as a major challenge for science instruction and for scientific literacy in the general population (e.g., Arons, 1973; Yager, 1983; Snow, 2010). The need to learn jargon terminology has been hypothesized to impose excessive cognitive load on learners (Groves, 1995), impede conceptual understanding and the formation of integrated mental models (Osborne, 2002; Britt *et al.*, 2014), reduce student motivation and interest (Yager, 1983), and especially impact struggling readers and nonnative speakers (Fang, 2006; Brown and Ryoo, 2008). The challenge is likely exacerbated by students' increasing use of Internet sources, which often eschew proper scientific language (Snow, 2010). Many instructors are likely aware of the issue as well. For example, several of the authors of this article have observed students struggling with scientific terminology in their courses and sought measures to assist those students.

Proposed Solutions for the "Terminology Problem"

Researchers have suggested various solutions for the jargon issue. These largely fall into two categories: *jargon-free* and *jargon-first* training (also known as the "content-first" and "jargon-training" approaches, respectively). In the former, jargon is removed from instructional content and withheld until after introductory concepts are learned. In the latter, jargon terms are directly trained, and the remainder of instruction occurs with jargon terms embedded per usual practice.

Two studies of the jargon-free approach have shown promising results. Brown and Ryoo (2008) had fifth-grade students learn about photosynthesis using software that introduced the process in plain, everyday language. On an immediate posttest, these students scored higher on measures of conceptual understanding than students who had used a software version that presented jargon throughout. Similarly, McDonnell *et al.* (2016) had undergraduate students learn about DNA structure and the genome via a jargon-free reading assignment and a conventional lecture on the topic. On an immediate posttest, these students earned higher scores on open-ended questions than students whose reading assignment included jargon (and with the correct usage of jargon terms on those questions comparable among both student groups). The findings from these two studies are consistent with the premise that removing jargon from course content, at least initially, can facilitate conceptual learning.

Despite those positive results, the jargon-free approach requires further research to determine generalizability and is not without cost. In particular, it necessitates the reworking of course materials to remove jargon. It also does not obviate the need to introduce jargon terms at a later point (such terms were indeed introduced by Brown and Ryoo, 2008; McDonnell *et al.*, 2016).

In contrast, a benefit of the jargon-first approach is that standard curricular materials can be used without alteration. Moreover, students complete the training with some (if not all) of the knowledge of the jargon terms that they ultimately need for a given topic. Proposed variants of this approach include studying term lists and definitions, learning the structure of scientific terms, and paraphrasing exercises (Fang, 2006). However, the effectiveness of these methods for learning scientific jargon per se has not been thoroughly investigated (although programs that employ a series of these techniques have shown promise for learning larger sets of vocabulary words, e.g., Snow, 2010).

In one relevant example, Mayer *et al.* (1984) had undergraduate participants learn about the concepts of "volume" and "mass" via informative handouts (which included definitions of both) before reading a passage about the concept of density. On an immediate test, those participants recalled more passage details and were better able to solve transfer problems than participants who had not received handouts. That study arguably did not directly test the jargon-first approach, however, given that jargon-term definitions were not the sole focus of training. Only two terms were trained, and the terms themselves were likely already somewhat familiar to participants. Yet the study's results are consistent with the premise that training that increases knowledge of scientific terms before they are encountered in instructional contexts can be beneficial for learning. Possibly, such training facilitates the construction of improved mental models of to-be-learned content.

Overall, in contrast with the conventional method of jargon terms being learned in concert with other course materials, the jargon-free and jargon-first approaches may foster conceptual learning that is unencumbered by unfamiliarity with jargon or the need to allocate cognitive resources in an effort to master them. In that sense, the two methods employ divergent strategies in an attempt to achieve the same goal. However, the efficacy and generalizability of either remains to be thoroughly investigated, and for the jargon-first approach, a crucial unanswered question is exactly what form that training should take.

Enhancing Learning via Retrieval Practice

In the learning sciences, the use of low-stakes practice quizzing and tests—a technique more formally known as retrieval practice or test-enhanced learning—is widely regarded as one of the most potent educational techniques uncovered to date (Dunlosky *et al.*, 2013; Brown *et al.*, 2014; Brame and Biel, 2015). Taking a practice quiz or test on information commonly improves memory for that information, relative to no activity or an equivalent amount of time spent engaged in studying, highlighting, or other nontesting activities (Roediger and Butler, 2011; Rowland, 2014). In some cases, practice quizzing even enhances learners' ability to transfer learning to new contexts, such as when solving application questions (for a review, see Pan and Rickard, 2018). Moreover, quizzing benefits have been observed for educationally relevant materials and in classroom settings (e.g., McDaniel *et al.*, 2007), including in biology courses (e.g., Bailey *et al.*, 2017; Carpenter *et al.*, 2017; Pape-Lindstrom *et al.*, 2018; Cooke *et al.*, 2019; Walck-Shannon *et al.*, 2019), and with formats ranging from free recall (e.g., McDaniel *et al.*, 2009) to cued recall (e.g., Pan *et al.*, 2015) and multiple-choice tests (e.g., Little *et al.*, 2012).

The evidence summarized above challenges the popular conception of quizzes and tests as solely instruments of assessment. They can also be deployed to enhance learning as well. This finding raises an intriguing question: Can practice quizzing be an effective way of addressing the jargon issue—as an implementation of the jargon-first approach—in science courses? Although a large portion of research on retrieval practice has occurred in the verbal learning tradition, with word lists as target materials, no study to date has specifically investigated use of the technique for acquiring scientific jargon in preparation

for learning course content. Moreover, the manner in which practice quizzes might be deployed to train on jargon, and how that could be done in a logistically feasible and efficient manner, has yet to be established. We delved into these and related issues in this article.

The Present Study

The current research was designed to address three primary questions: 1) Do practice quizzes on jargon terms (henceforth, jargon quizzing), which constitute a previously uninvestigated implementation of the jargon-first approach, enhance learning and comprehension of those terms at retention intervals of a week or more? 2) Are any such effects attainable with brief and easily implemented quiz methods that use existing learning platforms? 3) Do any effects of jargon quizzing translate to improved conceptual learning, as indexed by performance on conceptual exam questions that may not directly reference corresponding jargon terms?

Across two experiments—the first in a graduate-level neuroanatomy course and the second in two sections of an undergraduate-level neuroanatomy course—we implemented ~10-minute quizzes on jargon terms in online and in-class clicker format. Jargon terms were chosen in consultation with the course instructors and were defined as discipline-specific or specialized scientific vocabulary that students were unlikely to encounter outside the discipline (cf. McDonnell *et al.*, 2016). Quizzes occurred for one of two counterbalanced topics in experiment 1 and three of six counterbalanced topics in experiment 2. This design facilitated within-subjects comparisons of learning for jargon-quizzed topics versus topics that were not quizzed.

Experiment 1 involved online quizzes (three quizzes in total, administered before or during each of three consecutive class meetings), whereas experiment 2 involved in-class clicker quizzes (with one quiz occurring during each of 3 weeks of the course and before the first lecture of each week). The different quiz schedules reflected two common ways in which an instructor might implement the jargon-first approach in a course. Both online and clicker quizzes are generally easily implemented, although a possible disadvantage is that they may be less potent than more extensive formats (Rowland, 2014). Both methods have been investigated in the classroom in relatively few studies to date (e.g., McDaniel *et al.*, 2007; Glass, 2009; Mayer *et al.*, 2009; Anderson *et al.*, 2011).

We measured the effects of jargon quizzing on exams occurring at least 1 week and up to 12 weeks later (i.e., at longer retention intervals than those in prior studies involving the jargon-free approach). In experiment 1, assessment involved a single practice exam that occurred after all in-class content had been delivered but before the final exam. In experiment 2, assessment occurred via two midterms and a final exam. All assessments had questions from topics that were quizzed (quizzed condition) and not quizzed (control condition). Questions targeting directly quizzed content, namely, jargon terms (definition-focused questions), and questions targeting content that was not directly tested, namely, conceptual information (conceptually focused questions), were included. Definition-focused questions always involved jargon terms, thus requiring retrieval of knowledge about those terms, whereas conceptually focused questions did not always involve jargon but drew from the same topics. Definition-focused questions on quizzes and exam(s) were identical in experiment 1 but not in experiment 2. In experiment 2, we also assessed the effects of quizzing on students' lecture experiences and study habits.

If practice quizzing is an effective way to implement training on jargon terms, then improved memory for jargon terms and their definitions should be observed in the current research. Better transfer to conceptual test questions might result as well. If these outcomes are obtained, then they might be attributable to potential benefits of the jargon-first approach, such as decreased cognitive load and improved construction of mental models. Alternatively, if practice quizzing is ineffective or its effects are "washed out" in the course of regular instruction, then reduced or no benefits for either jargon-definition or conceptual summative test questions might be observed. Either outcome would also have important implications for the pedagogical utility of the jargon-first approach.

EXPERIMENT 1: ONLINE JARGON QUIZZING

Methods

Participants and Course Description. In the first experiment, the participants were 85 master's and doctoral students in the Occupational Therapy Program at Washington University in St. Louis who were taking a neuroanatomy course, OT 5782. The course took place over a 16-week academic semester in Spring 2014 and involved two weekly lectures of 80 minutes each. OT 5782 covers the structure and basic functions of the human nervous system as they support individuals engaging in activities of daily life, with course content supplied via lectures and supplemented via assigned textbook readings. The class followed a traditional lecture format wherein the instructor presented content using projected slides. Students earned points in the course for completing practice quizzes and a subsequent practice exam. One of the authors (L.T.C.) was the instructor of record for the course and another author (E.R.F.) taught lectures in the course pertaining to the quizzed topics. The study was approved by the Institutional Review Board (IRB) at Washington University in St. Louis. All students completed the quizzes and the practice exam.

Materials. Jargon terms (see Table 1) were chosen from three sections of the course: sensation and motor (which served as the experimental topics) and basic neuroanatomical terminology (which was used for filler items). There were 25 jargon terms chosen for the sensation topic and 27 chosen for the motor topic. These terms, which had a combined Flesch-Kincaid reading score of 50.2 (college graduate and above), were chosen on the basis of their use in lectures during a previous term. From these jargon terms, 16 definition-focused questions were designed for the sensation topic and 17 were designed for the motor topic. These fill-in-the-blank questions, in which one to three jargon terms were missing and had to be retrieved, were used for online quizzing. As described below in *Assessment of Learning Outcomes*, a portion of those questions also reappeared on the subsequent practice exam. Additionally, 34 conceptually focused questions (17 from each topic) were created based on the terms. These fill-in-the-blank or short-answer questions, which required single-word or short-phrase responses (per blank, with one to two blanks per question), were used on the practice exam and necessitated the recall of jargon terms or non-term information

TABLE 1. Jargon terms

Experiment	Topic	Terms
1: Online jargon quizzing	Sensory	Agnosia, ascending, ataxia, cones, descending, dorsal, equilibrium, hair cells, inferior temporal cortex, kinesthesis, lateral geniculate body, macula cells, medial geniculate body, Merkel discs, occipital, posterior parietal cortex, proprioception, rods, Ruffini corpuscles, somatosensory cortex, thalamus, tinnitus, transduction, ventral, vestibular cells
	Motor	Akinesia, anterior lateral, basal ganglia, caudate nucleus, cerebellum, chorea, direct pathway, dopamine, dystonia, globus pallidus, globus pallidus internal segment, hyperkinesia, hypokinesia, indirect pathway, lower motor neurons, medial, nigrostriatal, nucleus accumbens, posterior lateral, Purkinje cells, putamen, rigidity, substantia nigra pars reticulate, striatum, substantia nigra, subthalamic nucleus, upper motor neurons
2: In-class clicker jargon quizzing	Action potential	Absolute refractory period, action potential, closed Na+ channel, electrotonic current, graded potential, inactivated Na+ channel, intracellular/cytoplasmic resistance, membrane resistance, relative refractory period, repolarization
	Synapses	Action potential threshold, chemical synapse, end-plate potential, excitatory post-synaptic potential, inhibitory post-synaptic potential, ionotropic receptors, metabotropic receptors, quanta of neurotransmitter, saltatory conduction, summation
	Autonomic nervous system	Alpha 1 receptors, alpha 2 receptors, beta 1 receptors, beta 2 receptors, feedback loops, muscarinic receptors, parasympathetic nervous system, paravertebral ganglia, prevertebral ganglia, sympathetic nervous system
	Skeletal muscle	Actin, crossbridge cycle, motor unit, myosin, power stroke, recruitment, sarcomere, summation, tetanus, twitch
	Cardiac muscle	Afterload, baroreceptors, chemoreceptors, diastole, end diastolic volume, end systolic volume, Frank-Starling law, preload, stroke volume, systole
	Renal	Aquaporins, ascending limb of the Loop of Henle, clearance, counter-current exchange, descending limb of the Loop of Henle, excretion, reabsorption, renal clearance ratio, secretion, transport maximum

(as such, these questions could be classified under the "remember," "understand," or "apply" levels of Bloom's taxonomy, per Anderson and Krathwohl, 2001). Table 2 shows examples of the jargon definition–focused and conceptually focused questions developed for the sensation and motor topics.

Jargon quiz questions pertaining to basic neuroanatomy (e.g., ventral, caudal, axon, neuron) were also created to serve as filler quiz items, as discussed in the next section.

Procedure. An experiment timeline is presented in Figure 1, top panel.

Before beginning the study, participants were told that prior research had suggested that retrieving terminology might improve their ability to learn conceptual information in a course such as neuroanatomy. They were then told that, as part of their course, they would practice recalling general terminology for some of the course topics.

Quizzing. All participants took three quizzes pertaining to either sensation or the motor system outside class through Blackboard, a learning management system. These three quizzes were identical and were administered about 2 days apart, occurring immediately before and on the dates of that topic being covered in the course. Specifically, as each topic was covered for 2 days, a quiz was posted online at least 24 hours before each lecture; a third quiz was posted immediately after the second of the two lectures. Quizzes took ~10 minutes but did not have a strict time limit, and participants were notified that a quiz was available at least 24 hours before it needed to be completed. Correct-answer feedback was provided at the end of the quizzes. Credit in the class was awarded for taking the quizzes (i.e., not based on performance).

Counterbalancing. For counterbalancing purposes, a random half of the students took quizzes pertaining to sensation, and the other half of the students took quizzes pertaining to the motor system. Filler quizzes pertaining to general neuroanatomical terms were used to disguise the fact that some students were taking experimental quizzes and some were not. These quizzes were given when participants were not assigned to take the experimental quizzes. For example, before and during the sensation topic, half of the students took sensation quizzes, and the other half took a quiz on general neuroanatomical terms. Before and during the motor system topic, the students who had taken sensation quizzes took the basic neuroanatomical term quizzes, and those who had previously taken the basic neuroanatomical term quizzes took the motor system quizzes.

Assessment of Learning Outcomes. After lectures had concluded, but before the final exam, all participants took a comprehensive "practice exam" on Blackboard (~12 weeks after the final sensation quiz and 6 weeks after the final motor quiz). They were told that this practice exam would include questions from the sensation and motor topics. The practice exam included 16 previously tested questions from the sensation

TABLE 2. Example jargon definition–focused and conceptually focused questions

Experiment	Topic	Definition-focused questions	Conceptually focused questions
1: Online jargon quizzing	Sensory	The dorsal pathway ends in the _____ (part of the cortex), whereas the ventral pathway ends in the _____ (part of the cortex). Answers: posterior parietal cortex, inferior temporal cortex.	Henry recognizes a hammer sitting on the table, but is unable to reach for it. He doesn't have general motor impairment. Henry likely has impairment in what part of the cortex? Answer: posterior parietal cortex
		The receptor cells for pressure are _____. The receptor cells for temperature are _____. Answers: Ruffini corpuscles, Merkel discs	The Ruffini corpuscles help alert Patrick to a change in _____. Answer: pressure
	Motor	Increased muscle tone in some muscles, resulting in abnormal (bent, twisted) relatively fixed postures is called _____; increased tone in all muscles is called _____. Answers: dystonia, rigidity	Miguel has increased muscle tone in some muscles, resulting in abnormal but relatively fixed posture. Miguel likely has what movement disorder? Answer: dystonia
		_____ is the increase in muscular activity that can result in excessive abnormal movements, excessive normal movements, or a combination of both; _____ is a decrease in bodily movement. Answers: hyperkinesia, hypokinesia	Hypokinesia is _____from the basal ganglia. Answer: overinhibition
2: In-class clicker jargon quizzing[a]	Action potential	Electrotonic current is i. Local current consisting of similarly charged ions repelling each other in the cytoplasm. ii. Local current consisting of a single ion traveling through the cytoplasm. iii. Local current consisting of ions that occurs only in the dendrites and soma. iv. Local current consisting of electrons that move through the cytoplasm. Answer: i	Given a plot of two action potentials (not shown here), a series of questions need to be answered, including the following: • What has happened to the number of voltage-gated Na^+ channels? (circle one) increased no change decreased • Justify your response to the preceding question using two pieces of information provided in the "new" action potential.
		The passive movement of ions caused by similar electrical charges that oppose each other is i. electrotonic current ii. repolarization iii. intracellular/cytoplasmic resistance iv. membrane resistance Answer: i	• How has the duration of the absolute refractory period of the dashed-line action potential changed compared with control? • Tell me why you selected your answer to the preceding question. Your answer must include reference to the gating of Na^+ channels.

[a]In experiment 2, the definition-focused questions that appeared on the midterms and final exam were rephrased and had different answer choices.

topic and 15 previously tested questions from the motor topic. All of these questions were definition focused and were identical to those that had been used during quizzing. Because some questions necessitated more than one response (e.g., recalling both "ventral" and "dorsal" in response to a question about different neurological streams), each topic involved the recall of 25 terms for 25 possible points. The practice exam also included 17 conceptually focused questions for each topic (18 possible points each). The conceptual questions were tested first, with sensation tested before motor for both previously tested and conceptual questions. Students were given about a week to complete the test and were told to complete it in a single session.

Data Analysis. Statistical analyses of practice exam data (i.e., paired-samples *t* tests on quizzed versus not-quizzed topics) were performed separately for the definition-focused and conceptually focused questions.

Results

Quizzes. IRB restrictions precluded the availability of quiz data for the first experiment. All participants completed the assigned quizzes.

Practice Exam. Results are depicted in the left panel of Figure 2. Performance (henceforth reported in percentages rounded to the nearest whole number) was significantly better for the definition-focused questions that were previously tested (M = 53%, SE = 2%) than for those that were not (M = 43%, SE = 2%), as indicated by a paired-samples *t* test, $t(84) = 5.28$, $p < 0.001$, $d = 0.60$. Performance for previously nontested conceptual questions was marginally better when that topic was tested (M = 41%, SE = 2%) than when it was not (M = 37%, SE = 2%), $t(84) = 1.94$, $p = 0.06$, $d = 0.21$.

Some of the conceptually focused questions in this experiment had terms as answers and some did not. For example, as shown in Table 2, term-based conceptual questions include

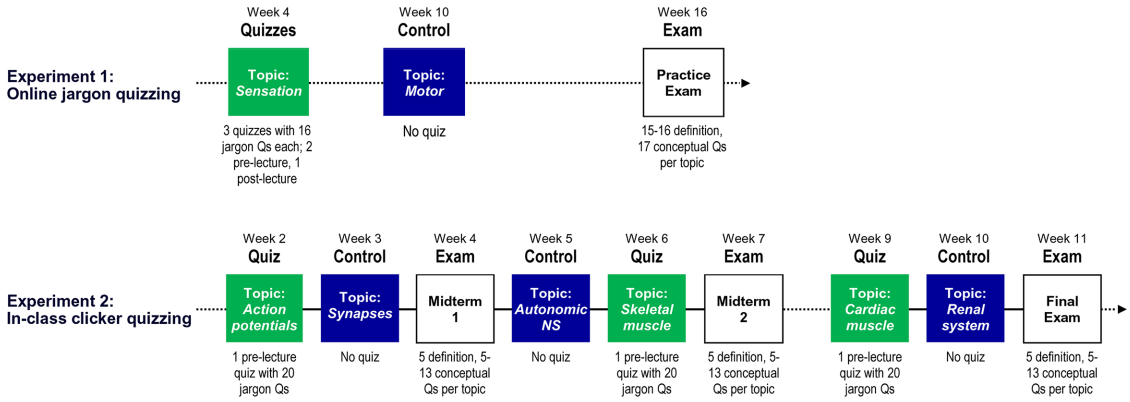

FIGURE 1. Schematic timelines of experiments 1 and 2. Note: in both timelines, one of two counterbalanced orders (assignment of topic to quizzing vs. control) is depicted. There was a university holiday during week 8 in experiment 2.

the ones for which "posterior parietal cortex" and "dystonia" are answers. Non–term based conceptual questions include the ones for which "pressure" and "over-inhibition" are answers. As a post hoc prediction, we hypothesized that the practice quizzes may have improved performance more for the conceptual questions for which a term was the answer than for the conceptual questions for which a term was not the answer. We thus divided the conceptual items into two sets such that both sensation and motor topics had 10 points of term-based conceptual questions (most but not all of these were terms that had been explicitly tested during the quizzes) and 8 points of non–term based conceptual questions. This post hoc analysis, the results of which are depicted in the right panel of Figure 2, revealed that term-based conceptual questions were recalled better when that topic had been tested (M = 38%, SE = 2%) than when that topic had not been tested (M = 32%, SE = 2%), $t(84) = 2.85$, $p < 0.01$, $d = 0.31$. Conceptual questions for which a term was not the correct answer were recalled with similar frequency when that topic had been tested (M = 46%, SE = 2%) and when it had not (M = 44%, SE = 2%), $t(84) = 0.66$, $p = 0.51$, $d = 0.07$.

For the term-based conceptually focused questions, a concern could be that participants were simply providing the practiced term where it might be appropriate. As can be seen in Table 2, however, aspects of the materials made this possibility unlikely, because competitive terms (i.e., terms that were also potentially plausible answers) were also studied. Nevertheless, to assess this concern, we examined intrusions for the term-based questions. For example, we assessed the likelihood of participants recalling "akinesia" instead of "dystonia" for the question "Miguel has increased muscle tone in some muscles, resulting in abnormal but relatively fixed posture. Miguel likely has what movement disorder?" Intrusions of this type were not significantly more likely to occur when those terms had been tested (M = 8%, SE = 1%) than when they had not (M = 7%, SE = 1%), $t(84) = 0.40$, $p = 0.70$, $d = 0.04$.

EXPERIMENT 2: IN-CLASS CLICKER JARGON QUIZZING

Methods

Participants and Course Description. In the second experiment, the participants were 406 undergraduate students at the University of California, San Diego, who were taking either of two Spring 2017 sections of BIPN 100, an introductory human physiology course. This course, which is a "gateway" prerequisite for several majors, occurs over an 11-week academic quarter and covers the organ systems of the human body and their regulation via the nervous and endocrine systems. Both sections occurred on the same days of the week but at different times of day (i.e., a morning and an afternoon section) and were taught by the same instructor (J.C.).

BIPN 100 involves twice-weekly lecture sections of 80 minutes each, supplemented by a 1-hour weekly discussion period. Each week focuses on a different topic (e.g., skeletal muscle), and although there are some common principles shared between topics, there is relatively modest content overlap between them. In Spring 2017, the course was taught using an

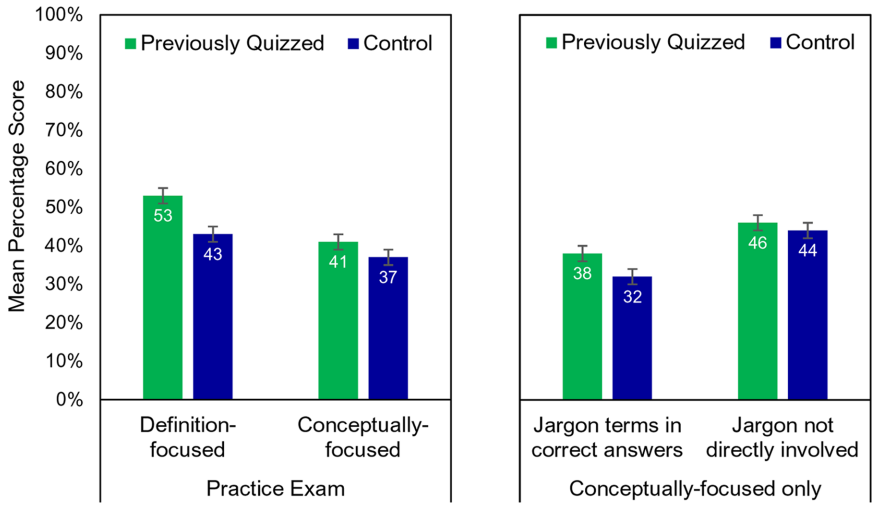

FIGURE 2. End-of-semester practice exam results after online jargon quizzing in experiment 1 (error bars = SEM).

active-learning format wherein traditional lecture was combined with problem-solving group activities. These were supplemented by assigned textbook readings. Students earned points for their performance on in-class clicker quizzes, two midterms, and a final exam.

The study was approved by the IRB at the University of California, San Diego, which also authorized collection of anonymized student data and exam performances. Students in both sections had similar grade point averages (GPAs; M = 3.06 in both), Scholastic Aptitude Test (SAT) Math scores (M = 663 vs. 678), and ACT composite scores (M = 28.8 vs. 29.0). Data from about one-third of the students were excluded from data analysis due to nonattendance at one or more of the in-class quizzes or dropping of the course, yielding a total sample size of 277 (morning section, $n = 117$; afternoon section, $n = 160$). Independent-samples t tests revealed no significant differences among included students from the two sections in GPA or ACT scores (p values ≥ 0.78) and a nonsignificant trend toward better SAT Math scores for the afternoon section ($p = 0.077$).

Materials. Jargon terms (see Table 1) were chosen from six weekly course topics: action potentials, synapses, autonomic nervous system, skeletal muscle, cardiac muscle, and renal system. These terms had a combined Flesch-Kincaid reading score of 65.8 (college graduate and above). There were 10 jargon terms per topic. The instructor chose the terms on the basis of their being challenging for students to learn in prior iterations of the course. Using these terms, 20 definition-focused clicker questions were constructed per topic (i.e., two questions per jargon term; one provided the definition and asked for the jargon term, and the other involved the reverse, per McDaniel *et al.*, 2013; Pan and Rickard, 2017). Each clicker question consisted of a question and four randomly ordered answer choices (including the correct answer and three competitive lures, cf. Little *et al.*, 2012). Ten definition-focused exam questions were also constructed per topic (60 total questions), with five appearing on each of a subsequent midterm or the final exam, and with no question appearing on more than one exam. All of these definition-focused questions were modified versions of questions used during quizzing. Specifically, each had different answer choices (e.g., for the question on electrotonic current, the correct answer of "the passive movement of ions caused by similar electric charges that oppose each other" was changed to "local current consisting of similarly charged ions repelling each other in the cytoplasm"). As such, students could not rely on rote memorization of correct quiz question answers to respond correctly to definition-focused exam questions.

In addition, 60 conceptually focused questions (five to 13 per topic) were constructed for use on the midterms and the final exam. These questions, which drew from the same topics as the definition-focused questions but did not require the use of jargon terms, involved evaluating data, making predictions based on new scenarios, or generating and justifying inferences (and hence could be classified under the "apply," "analyze," or "evaluate" levels of Bloom's taxonomy, per Anderson and Krathwohl, 2001). The conceptually focused questions for a given topic were designed to be answered in succession, with a given question often referring to the next. These questions also resembled in-class problem-solving group activities but did not duplicate them.

Table 2 shows examples of the jargon definition–focused questions and conceptually focused questions for the action potential topic. It should be noted that the midterm and final test questions, which counted toward the majority of the actual course grade, would have been designed in a very similar (if not identical) manner had no experimentation occurred in the course. Thus, the exam materials had high "ecological validity."

Procedure. An experiment timeline is presented in Figure 1, bottom panel.

On the first day of class, the instructor announced that in-class clicker quizzes would be administered at various points throughout the course. It was also stated that each quiz date would be posted online beforehand, would cover key terms from the assigned readings, and would count toward 5% of the overall course grade. The jargon terms would be provided in a study sheet posted online before each quiz. The instructor justified the implementation of the quizzes as a tool to help drive learning.

Quizzing. Participants completed three in-class clicker quizzes, one for a topic that was taught before each high-stakes exam (midterm 1, midterm 2, and the final exam), in three separate weeks. On the Friday before a quiz was to be administered, the instructor announced it on the course's Blackboard page. A list of the 10 jargon terms to be quizzed was also posted (students were left to their own devices regarding quiz preparation; e.g., it was possible to look those terms up online or in the textbook). The quiz took place on the following Monday at the start of the lecture period. During the quiz, the instructor projected jargon definition–focused clicker questions for students to answer, one at a time. There were 20 multiple-choice questions and ~30 seconds was allotted per question. Over the first 10 questions, each jargon term was presented once with the definition to be retrieved (i.e., students had to select among four possible definition answers), and over the next 10 questions, each definition was presented once with the term to be retrieved (i.e., students had to select among four possible jargon-term answers). That design contributed to students' generally high quiz scores, with most students attaining a perfect score on the last 10 questions. The instructor presented brief correct-answer feedback after each question. After the quiz ended, the remainder of the class period proceeded normally.

Clicker quizzing involved the iClicker remote system produced by Macmillan Learning. The instructor had used this system in prior iterations of the course. All students were required to bring an iClicker remote to class and to activate their clickers before the start of each quiz. Each clicker was linked to an individual student ID. During the quizzes, five to six instructional assistants patrolled the lecture hall to ensure that students were completing the quizzes individually.

During class periods in which no quizzing occurred, instruction proceeded as normal. Students in each section received nearly identical lectures on each course topic regardless of whether quizzing did or did not occur.

Counterbalancing. For counterbalancing purposes, students in the morning section had quizzes in weeks 2, 6, and 9 (on the action potentials, skeletal muscle, and cardiac muscle topics),

whereas students in the afternoon section had quizzes in weeks 3, 5, and 10 (on the synapses, autonomic nervous system, and renal topics). This schedule enabled a within-subjects comparison of the quizzing versus control conditions across the entire sample in conjunction with counterbalanced assignment of topic to those conditions across successive weeks of the course. Unlike experiment 1, no filler quizzes were given. This was due to the fact that counterbalancing occurred at the level of entire sections, and all students in each section took quizzes at the same time.

Assessment of Learning Outcomes. The midterm exams (midterm 1 during week 4; midterm 2 during week 7) assessed learning on topics covered during the 3 weeks immediately preceding their administration. These included the two topics that had been subject to the quizzing versus control manipulation (i.e., the assessment of learning from the quizzes occurred 1–2 weeks after they had been administered). There were also other questions that assessed topics that had not been subject to the quizzing versus control manipulation (e.g., for midterm 1, material from week 1). These questions were not considered part of the experiment and were not analyzed. The final exam, which occurred during the 11th week of the course, was cumulative and had questions assessing content from each week of the course (including topics from the 3 weeks immediately preceding its administration, which were only assessed on that exam). Both midterms took place during a regular class period, while the final exam was scheduled for a longer 3-hour period.

On each of the three exams, each of the topics from the preceding three weeks was assessed with five definition-focused questions (i.e., questions that targeted jargon terms but were not identical to those used during quizzing, as previously described) and five to 13 conceptually focused questions. Moreover, on the final exam, additional definition-focused and conceptually focused questions assessed topics from weeks 2, 3, 5, and 6 (which had already been assessed on either of the preceding midterms) as well as week 1. On each exam, each topic's questions were presented in succession, with definition-focused questions preceding conceptual questions (the questions were lettered as parts "a," "b," "c," and so on of a given topic's question set).

All exams were graded by instructional assistants who were blind to condition and using an instructor-furnished rubric. One point was allotted per question. Each midterm accounted for 22.5% of the course grade, and the final exam accounted for 50%.

Surveys. During weeks 6 and 10, at the conclusion of the Monday lecture, a brief clicker survey was administered. That survey asked students to provide a rating on a scale of 1–5 for three questions, namely 1) how well they understood the lecture for that day, 2) how engaged they were during the lecture, and 3) how difficult they found the lecture material to be. The final exam also included a multiple-choice exit survey. That survey measured 1) how much time students used to prepare for each in-class quiz, 2) the types of study techniques they used to prepare, 3) whether students studied other jargon-term definitions beyond those that had been assigned online, and 4) whether students incorporated jargon-term study into their midterm and final exam preparation.

Data Analysis. All analyses were conducted on data aggregated across both sections of the course and separately for each exam and question type. Analyses of midterm and final exam data were restricted to questions on the topics from weeks 2–3, 5–6, and 9–10, during which the quizzing versus control manipulation occurred. As with experiment 1, analyses (paired-samples *t* tests) were performed separately for the definition-focused and conceptually focused questions. Final exam data involving topics from weeks 2, 3, 5, and 6, which were previously assessed on midterms 1 or 2, were also analyzed separately. Finally, as described later in this article, we performed supplementary analyses to examine potential effects of course section.

Results

Quizzes. Participant mean performance was 95%, SE = 0.4%, 95%, SE = 0.6%, and 95%, SE = 0.6%, for the first, second, and third in-class quizzes (collapsed across counterbalanced topics). Across all quizzes, performance on the first 10 questions averaged 85% and rose to 98% for the second 10 questions. The patterns were nearly identical across sections.

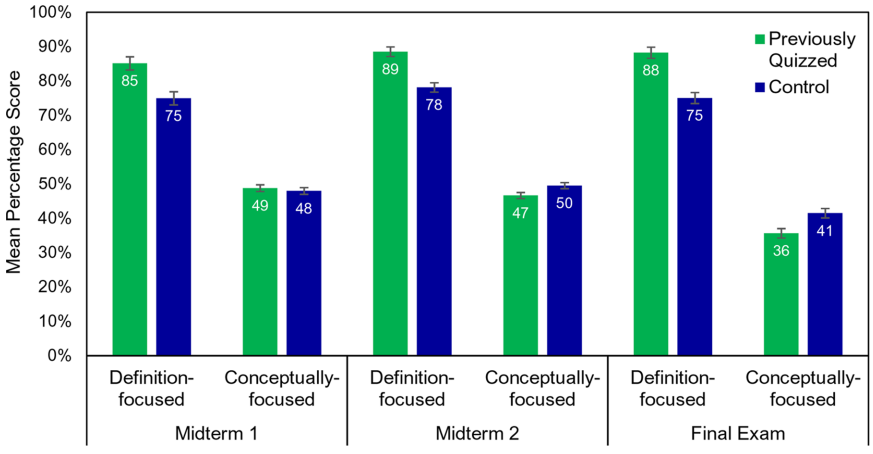

FIGURE 3. Midterm and final exam results after in-class clicker jargon quizzing in experiment 2 (data collapsed across course sections; error bars = SEM).

Midterm and Final Exams. Results for the entire sample (collapsed across sections) are depicted in Figure 3. As is evident upon inspection of the figure, mean performance on definition-focused questions was higher on all exams in the quizzed versus control conditions. Those differences were statistically significant in each case, as indicated by paired-samples *t* tests: midterm 1 (M = 85%, SE = 1% vs. M = 75%, SE = 1%), $t(276) = 5.32$, $p < 0.0001$, $d = 0.32$; midterm 2 (M = 89%, SE = 1% vs. M = 78%, SE = 1%), $t(276) = 7.58$, $p < 0.0001$, $d = 0.46$; and the final exam (M = 88%, SE = 1% vs. M = 75%, SE = 1%), $t(276) = 8.27$, $p < 0.0001$, $d = 0.50$. By contrast, mean performance on conceptually focused questions was either equivalent or

somewhat lower in the quizzed versus control conditions. There was no significant mean difference for midterm 1 (M = 49%, SE = 1% vs. M = 48%, SE = 1%), $t(276) = 0.87$, $p = 0.39$, $d = 0.052$, but there were significant mean differences for midterm 2 (M = 47%, SE = 1% vs. M = 50%, SE = 1%), $t(276) = 3.27$, $p = 0.0012$, $d = 0.20$, and the final exam (M = 36%, SE = 1% vs. M = 41%, SE = 1%), $t(276) = 4.28$, $p < 0.0001$, $d = 0.26$.

Regarding topics that were previously assessed on midterms 1 or 2, and were assessed again on the final exam (unpublished data), mean performance on definition-focused questions was higher for the quizzed versus control conditions (M = 87%, SE = 1% vs. M = 81%, SE = 1%), $t(276) = 5.09$, $p < 0.0001$, $d = 0.31$, again showing a benefit for quizzing on jargon terms (but persisting over a longer retention interval and after repeated testing). There was no difference in mean conceptual question performance between the quizzed versus control conditions (M = 67%, SE = 1% vs. M = 68%, SE = 1%), $t(276) = 1.04$, $p = 0.30$, $d = 0.062$, mirroring the patterns observed for midterm 1.

Unlike the first experiment, the conceptually focused questions in experiment 2 could not be explicitly categorized into those that had jargon terms as answers or not. All could be answered without directly using a jargon term. Thus, the conceptual question data in this experiment were not analyzed separately in terms of term-based and non–term based question subgroups.

Surveys. In-class surveys collected during weeks 6 and 10 revealed no significant differences in the distribution of ratings between the quizzed versus control conditions with regard to understanding of the lecture, engagement in the lecture, or ease of the lecture material (χ^2 test, p values ≥ 0.14). On the final exam survey, the majority of participants (84%) reported spending between 15 minutes and 1 hour preparing for jargon quizzes, during which studying (21%), self-testing (26%), or a combination of both (39%) were common. Further, 84% of participants reported incorporating jargon-term lists into their exam preparation, and 57% reported that the quizzing prompted them to study other jargon terms in the course.

DISCUSSION

Did Jargon Quizzing Enhance the Learning and Comprehension of Jargon Terms?

Across two experiments involving courses with different content and different quizzing procedures, practice quizzing on jargon terms generally yielded improvements in students' learning and comprehension of those terms as assessed on definition-focused exam questions. This pattern was observed in cases in which exam questions were identical to (experiment 1) and different from (experiment 2) those used during quizzing. The mean observed gain in percentage terms on those questions, relative to a no-quizzing control condition, ranged from 10 to 13% (and in effect size terms, from $d = 0.32$ to $d = 0.60$). When considering the time that had elapsed from the practice quizzes to exam administration (on the order of weeks, wherein forgetting processes inevitably took place) and the intervening study that students typically performed (on both quizzed and nonquizzed topics), this pattern of improvement—with effect sizes approaching or exceeding a medium size effect (by conventional standards and percentage gains that are equivalent to that of an entire letter grade; cf. Cohen, 1988)—can be regarded as substantial.

It is also notable that our jargon-quizzing intervention in experiment 2 was successful in prompting most students to devote at least some time to studying jargon terms before their use in class and to incorporate jargon terms into their exam preparation activities. Thus, beyond the direct effects of the jargon quizzes on learning and memory, our intervention had additional effects on students' learning activities and the relative emphasis that students placed on jargon terms.

Were Jargon-Quizzing Benefits Attainable Using Online and In-Class Clicker Quizzes?

In this study, both online and in-class clicker quizzing improved learning. Experiment 1 reinforces the conclusion that the effectiveness of online quizzes for learning is comparable to quizzes conducted in classrooms and in controlled laboratory environments. Experiment 2 confirms the efficacy of in-class clicker questions at achieving similar results. In addition, the quizzing benefit observed after 12 weeks in experiment 1 aligns with prior results showing that the effects of online and in-class retrieval practice persist across periods of several months or more (e.g., Carpenter *et al.*, 2009; Roediger *et al.*, 2011). It is also notable that quizzing benefits persisted across repeated high-stakes testing in experiment 2. The fact that learning benefits emerged with different quizzing platforms, training schedules, biology courses, and student populations indicates that the effects of jargon quizzing are robust. Overall, we can conjecture that the present findings are not specific to a fortuitously selected quizzing implementation or a particular student level and are likely to generalize to other implementations of jargon quizzing, including in various biology and other science, technology, engineering, and mathematics (STEM) courses.

Importantly, the results of experiment 2 reveal that jargon-quizzing benefits for definition-focused questions are not restricted to cases in which quiz and exam questions are identical (wherein a rote memorization strategy would have been effective). Rather, benefits can occur when subsequent course exams contain reworded questions with entirely different answer options. Successfully answering those questions requires a deeper understanding of jargon-term definitions. That finding represents a successful case of near transfer (Perkins and Salomon, 1994).

Did Any Effects of Jargon Quizzing Translate to Improved Conceptual Learning?

In both experiments, we observed inconsistent results on assessments of conceptual understanding. Only when recall of previously trained jargon terms was required was there any evidence of a quizzing benefit for conceptually focused exam questions (experiment 1). There was no reliable benefit for conceptual questions that did not directly involve jargon terms (experiments 1 and 2), and on midterm 2 and the final exam in experiment 2, there was evidence of a modest quizzing deficit for conceptual questions across both course sections. This raises the possibility that jargon quizzing may have deleterious effects, although such effects were not as consistent or as large as the benefits for definition-focused questions. Given that

inconsistency, we hesitate to engage in extended speculation as to the source of a jargon-quizzing deficit for conceptual question performance, although one possibility is that students may have focused their exam preparation on jargon definitions at the expense of conceptual content. It is important to note, however, that in experiment 1, a conceptual question benefit was observed in a post hoc analysis that identified questions for which the jargon terms were the correct answers; this constitutes a second example of transfer for a case wherein practice quiz and subsequent exam questions are not identical. Overall, the quizzing approaches investigated in this study appear to have limitations with regard to transfer to indirectly related course content, and particularly under training conditions such as those employed in experiment 2.

Jargon-First Training, Cognitive Processes, and Course Instruction

Ideally, the jargon-first approach should 1) enhance students' knowledge of relevant jargon terms, 2) enhance their ability to use those terms in relevant contexts, and 3) alleviate any negative effects that may occur when jargon is learned in concert with other course materials (e.g., high cognitive load). The present results provide strong evidence of efficacy for the first objective, indications of some potential for the second, and little evidence for the third. Further consideration of the cognitive processes that training on jargon terms may evoke, and the effects of such training on the subsequent learning of course materials, suggests possible explanations for this pattern.

One possibility is that learning jargon terms before those terms are used in other contexts may yield a primarily lexical representation that is not strongly grounded in conceptual knowledge. This outcome would be especially likely if students have no relevant conceptual information in semantic memory with which to associate newly learned jargon terms. As a result, they can recognize and define those jargon terms, but may not necessarily know how to apply them in a variety of other contexts. Both steps are likely required for successful transfer to occur (Barnett and Ceci, 2002).

A second and related possibility is that the jargon-first approach, at least as implemented in this study, may yield learning that is not necessarily fluid and automatic. As a result, when previously trained jargon terms are later encountered in other course materials, students must mentally disengage to recall their definitions, and there may be little improvement in terms of allocation of cognitive resources. This may have especially been the case in experiment 2, wherein each jargon term was trained only once. Notably, our in-class surveys in experiment 2 recorded no evidence of improved lecture understanding, engagement, or difficulty. This suggests that the jargon-quizzing intervention in that experiment, while effective at enhancing the learning of jargon terms, was relatively limited in its downstream effects on students' processing of subsequently encountered course content.

How then might all three objectives be attained? Four avenues hold promise. The first is to incorporate more conceptual information into the jargon-training process. For example, the informational handouts provided by Mayer *et al.* (1984), which yielded improved transfer performance, not only included jargon definitions, but also contained examples and diagrams. Adding that information may provide conceptual "anchor points" with which to associate jargon terms. Second, a more extensive jargon-training procedure may be helpful. This might involve training akin to the repeated and spaced quizzing method employed in experiment 1, in which there was some evidence of transfer to conceptually focused questions (in contrast to that obtained with a one-session training procedure in experiment 2). Such training could create the fluency that is needed (much as practicing on multiplication tables repeatedly confers fluency of retrieval of multiplication facts) to enhance cognitive resources available for subsequent encounters with material that includes jargon terms. Third, a jargon-quizzing method that incorporates higher-order quiz questions (e.g., not just recalling exact definitions but also applying jargon terms to new examples) may yield learning that is better grounded in conceptual knowledge and more able to transfer to other contexts (for a related example, see Jensen *et al.*, 2014). The fourth avenue is to abandon the jargon-first approach entirely in favor of the jargon-free approach. As demonstrated by Brown and Ryoo (2008) and McDonnell *et al.* (2016), that method has shown promise at enhancing conceptual understanding (possibly because learners are free to devote their full attention during the initial study).

Study Limitations and Future Directions

As with any study that occurs in an authentic educational context, we were not able to fully control for students' outside study activities or the effects of other aspects of regular course instruction. Most (if not all) students in both experiments likely engaged in at least some studying in the time period between quizzing and exam(s), and sometimes in groups; such study activities may have attenuated some of the benefits of quizzing. Conversely, the practice exam in experiment 1 did not count toward students' grades, which may have reduced student motivation and contributed to the relatively low level of overall performance. Additionally, in experiment 2, more than half of the students in our sample chose to study additional jargon terms beyond those that were directly trained. In that experiment, students also engaged in weekly problem-solving group activities that served as training for answering the conceptually focused questions on course exams. Either of these activities may have attenuated effects of jargon quizzing. Further, experiment 2 suffered from 31.7% attrition; it is possible that the lowest-performing students, for whom retrieval practice may or may not yield benefits (e.g., Carpenter *et al.*, 2016), comprised the bulk of dropouts and were not represented in the final analyses.

There are several promising directions for follow-up work. These include an investigation of other forms of jargon quizzing, perhaps including aforementioned types as well as others, such as writing out explanations (e.g., Hinze *et al.*, 2013), multiple-choice questions wherein the answer choices are highly competitive with one another (e.g., Little *et al.*, 2012), the use of more sophisticated forms of feedback (e.g., Pan *et al.*, 2019), and comparisons with nonquizzing methods (e.g., Pan *et al.*, 2015). Another possibility is to focus on the effects of jargon quizzing for students with beginning or nonnative English ability (an issue that could not be separately examined in the present work due to insufficient data). A direct comparison of the jargon-free versus jargon-first approaches, as well as more direct measures of cognitive load, may be revealing as well. Future work might also incorporate free-response assessments

wherein the usage of jargon terms can be more extensively analyzed (Zukswert *et al.*, 2019). Finally, the selection of jargon terms according to normed data or a detailed analysis of their specific usage in course materials may also yield insights.

CONCLUSION

The present results serve as a proof of concept for a practical and potentially effective method of partially addressing the "terminology problem" in biology and other STEM courses. Our results reveal that a total time investment of ~10–30 minutes in jargon quizzing per topic—with minimal work required to implement practice quizzing using existing online and in-class learning platforms and no changes to other course materials—yields long-lasting improvements in students' scientific lexica, albeit with limited transfer to conceptual assessments that do not directly involve jargon terms.

ACKNOWLEDGMENTS

The completion of experiment 1 was supported by a James S. McDonnell Foundation Collaborative Activity Award. Portions of this research were presented in September 2018 at the Center for Integrative Research on Cognition, Learning, and Education's (CIRCLE) Conference in St. Louis, MO, and in July 2019 at the Society for the Advancement of Biology Education Research (SABER) meeting in Minneapolis, MN.

REFERENCES

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

Anderson, L. S., Healy, A. F., Kole, J. A., & Bourne, L. E. (2011). Conserving time in the classroom: The clicker technique. *Quarterly Journal of Experimental Psychology*, *64*(8), 1457–1462.

Arons, A. (1973). Toward wider public understanding of science. *American Journal of Physics*, *41*(6), 769–782.

Bailey, E. G., Jensen, J., Nelson, J., Wiberg, H. K., & Bell, J. D. (2017). Weekly formative exams and creative grading enhance student learning in an introductory biology course. *CBE—Life Sciences Education*, *16*(1), ar2.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637.

Brame, C. J., & Biel, R. (2015). Test-enhanced learning: The potential for testing to promote greater learning in undergraduate science courses. *CBE—Life Sciences Education*, *14*(2), es4.

Britt, M. A., Richter, T., & Rouet, J. F. (2014). Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist*, *49*(2), 104–122.

Brown, B. A., & Ryoo, K. (2008). Teaching science as a language: A "content-first" approach to science teaching. *Journal of Research in Science Teaching*, *45*(5), 529–553.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick*. Cambridge, MA: Harvard University Press.

Carpenter, S. K., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review*, *28*(2), 353–375.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology*, *23*(6), 760–771.

Carpenter, S. K., Rahman, S., Lund, T. J., Armstrong, P. I., Lamm, M. H., Reason, R. D., & Coffman, C. R. (2017). Students' use of optional online reviews and its relationship to summative assessment outcomes in introductory biology. *CBE—Life Sciences Education*, *16*(2), ar23.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooke, J. E., Weir, L., & Clarkston, B. (2019). Retention following two-stage collaborative exams depends on timing and student performance. *CBE—Life Sciences Education*, *18*(2), ar12.

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

Fang, Z. (2006). The language demands of science reading in middle school. *International Journal of Science Education*, *28*(5), 491–520.

Glass, A. L. (2009). The effect of distributed questioning with varied examples on exam performance on inference questions. *Educational Psychology*, *29*(7), 831–848.

Groves, F. H. (1995). Science vocabulary load of selected secondary science textbooks. *School Science and Mathematics*, *95*(5), 231–235.

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language*, *69*(2), 151–164.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test … or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, *26*(2), 307–329.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*(11), 1337–1344.

Mayer, R. E., Dyck, J. L., & Cook, L. K. (1984). Techniques that help readers build mental models from scientific text: Definitions pretraining and signaling. *Journal of Educational Psychology*, *76*(6), 1089–1105.

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., … & Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, *34*(1), 51–57.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read–recite–review study strategy: Effective and portable. *Psychological Science*, *20*(4), 516–522.

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. *Applied Cognitive Psychology*, *27*(3), 360–372.

McDonnell, L., Barker, M. K., & Wieman, C. (2016). Concepts first, jargon second improves student articulation of understanding. *Biochemistry and Molecular Biology Education*, *44*(1), 12–19.

Osborne, J. (2002). Science without literacy: A ship without a sail? *Cambridge Journal of Education*, *32*(2), 203–218.

Pan, S. C., Hutter, S., D'Andrea, D., Unwalla, D., & Rickard, T. C. (2019). In search of transfer following cued recall practice: The case of biology concepts. *Applied Cognitive Psychology*, *33*(4), 629–645.

Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts? *Journal of Experimental Psychology: Applied*, *23*(3), 278–292.

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

Pan, S. C., Rubin, B. R., & Rickard, T. C. (2015). Does testing with feedback improve adult spelling skills relative to copying and reading? *Journal of Experimental Psychology: Applied*, *21*(4), 356–369.

Pape-Lindstrom, P., Eddy, S., & Freeman, S. (2018). Reading quizzes improve exam scores for community college students. *CBE—Life Sciences Education*, *17*(2), ar21.

Perkins, D. N., & Salomon, G. (1994). Transfer of learning. In Husen, T., & Postelwhite, T. N. (Eds.), *International handbook of educational research* (2nd ed., Vol. 11, pp. 6452–6457). Oxford, UK: Pergamon.

Roediger, H. L., III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*, 382–395.

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

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.

Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science*, *328*(5977), 450–452.

Walck-Shannon, E. M., Cahill, M. J., McDaniel, M. A., & Frey, R. F. (2019). Participation in voluntary re-quizzing is predictive of increased performance on cumulative assessments in introductory biology. *CBE—Life Sciences Education*, *18*(2), ar15.

Wandersee, J. H. (1988). The terminology problem in biology education: A reconnaissance. *American Biology Teacher*, *50*(2), 97–100.

Yager, R. E. (1983). The importance of terminology in teaching K–12 science. *Journal of Research in Science Teaching*, *20*(6), 577–588.

Zukswert, J. M., Barker, M. K., & McDonnell, L. (2019). Identifying troublesome jargon in biology: Discrepancies between student performance and perceived understanding. *CBE—Life Sciences Education*, *18*(1), ar6.