Received: 20 April 2018 | Revised: 22 August 2018 | Accepted: 29 November 2018

DOI: 10.1002/acp.3506

RESEARCH ARTICLE

WILEY

In search of transfer following cued recall practice: The case of process-based biology concepts

Steven C. Pan[1,2] | Sarah A. Hutter[1] | Dominic D'Andrea[1] | Daanish Unwalla[1] | Timothy C. Rickard[1]

[1] Department of Psychology, University of California San Diego, San Diego, California, USA

[2] Department of Psychology, University of California Los Angeles, Los Angeles, California, USA

Correspondence
Steven C. Pan, Department of Psychology, University of California Los Angeles, Los Angeles, CA 90095.
Email: stevencpan@psych.ucla.edu

Summary

Previous work has demonstrated that cued recall of a term from a fact yields learning that does not transfer, relative to a restudy control, to recall of another term from the same fact. Here we report six experiments in which a series of manipulations during the initial study and training phases of learning, hypothesized to increase transfer for process-based biology concepts, were investigated. In Experiments 1 and 2, fill-in-the-blank questions combined with immediate or delayed and repeated correct answer feedback improved learning but not transfer. In Experiments 3 and 4, practice questions that involved recalling process steps, understanding ordinal relationships, or making inferences did not improve transfer. Positive transfer was produced, however, in Experiments 5 and 6 via *retrieval–verification–scoring*, a new method in which difficult fill-in-the-blank questions were combined with extensive feedback processing. We discuss implications for transfer in both theoretical and applied contexts.

KEYWORDS

cued recall, feedback, retrieval practice, testing effect, transfer

1 | INTRODUCTION

Across virtually all areas of science, technology, engineering, and mathematics (STEM) education, the learning and retention of *process-based concepts* is an essential step for achieving mastery. For current purposes, a process-based concept is defined as a sequence of events that occur over time, leading to a predictable end state. Such concepts are ubiquitous in the target domain of this manuscript, namely, introductory biology. A classic example is the concept of *protein synthesis*. It involves the following process: DNA is first copied into RNA via transcription, and then RNA is coded into protein via translation (Freeman, Quillin, & Allison, 2014). In that concept, two sequential processes, transcription and translation, result in the formation of protein.

Given the ubiquity and foundational nature of such materials, it is important to understand how students can learn them efficiently (i.e., with the least time invested), durably (i.e., yielding long-term retention), and comprehensively. One promising approach is *retrieval practice with feedback*. Retrieving information from memory, as occurs on a practice test, yields robust learning benefits over restudy and other nonretrieval methods. That finding, commonly known as the *retrieval practice effect* or *testing effect*, has been observed across a wide range of educationally relevant materials (for discussion see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013), using different test formats (e.g., Kang, McDermott, & Roediger, 2007), with diverse learners (e.g., Pan, Pashler, Potter, & Rickard, 2015), and under different levels of motivation (e.g., Kang & Pashler, 2014). The effect is enhanced if *correct answer feedback* is provided after attempting retrieval, and it persists over time; the relative advantage of retrieval practice over restudy even increases with longer retention intervals (e.g., Carpenter, Pashler, & Cepeda, 2009; Rowland, 2014). Given these findings, retrieval practice would appear to be well suited to enhance the learning and transfer of process-based concepts—a possibility that we explore in this manuscript.

1.1 | Specificity of learning through cued recall practice

In the research literature (see Rickard & Pan, 2017), and quite likely also in current educational practice, the most commonly used variant

of retrieval practice is *cued recall*, in which part of a studied item is presented on the initial test as a cue, and the remainder of the item is to be retrieved from memory. For the concept given above, an example cued recall question (in fill-in-the-blank format) is: "In protein synthesis, DNA is first copied into RNA via _____?" There is overwhelming evidence that cued recall can facilitate, relative to restudy, subsequent retrieval of the same term that was retrieved on the initial test (i.e., the test that occurred during training). However, there is correspondingly strong evidence that for educationally relevant materials (e.g., fact learning), that facilitated learning does not extend, or *transfer*, to questions from the same stimulus that require a different term response (henceforth, *stimulus–response rearrangement*). For example, in Pan, Gopal, and Rickard (2015), cued recall involving the fill-in-the-blank question, "Thomas Jefferson purchased _____ from France," for which the answer is Louisiana, yielded substantial learning relative to restudy when that same question was again asked after a delay, but did not yield transfer to "Thomas Jefferson purchased Louisiana from WHOM?," relative to restudy. Rather, criterial test performance on such items was equivalent to performance after restudy. Similar findings have been reported by Hinze and Wiley (2011) and Pan and Rickard (2017); one exception is McDaniel, Anderson, Derbish, and Morrisette (2007), which is discussed later in this manuscript. Thus, although retrieval practice appears to be at least as effective as restudy in nearly all instances, it is, in work to date, evidently superior to restudy only for the initially tested term in cases involving stimulus–response rearrangement. This contrasts with some findings for other transfer contexts (e.g., application questions as in McDaniel, Howard, & Einstein, 2009; for a review of retrieval practice and transfer effects, see Pan & Rickard, 2018). When considering that it is not always feasible to implement retrieval practice in a manner that trains on all to-be-learned information (Roediger, Putnam, & Smith, 2011), and given that commonly available sets of practice questions (such as in textbook review sections) often include little more than a single test question (often cued recall) per concept, the amount of transfer following cued recall practice is an important consideration for instructors, students, and other users of the technique.

1.1.1 | Is specificity of learning universal to all cued recall methods?

It remains to be determined whether the extreme specificity of learning through cued recall testing that occurs for the cases outlined above is universal for facts and concepts or whether there are conditions, heretofore unexplored, wherein positive transfer will be observed. For instance, in many retrieval practice studies, target materials are studied relatively briefly and without supporting information (e.g., explanatory diagrams); it has yet to be established whether the same patterns would manifest for more enriched and more extensively studied information. Further, researchers have hypothesized that some forms of retrieval practice, including the use of questions that invoke processing of conceptual relationships, can yield more transferrable learning (e.g., Jensen, McDaniel, Woodard, & Kummer, 2014). There has also been some research and theorizing which suggests that feedback that is processed for greater lengths of time and/or contains more than just the correct answer can enhance various types of transfer (e.g., Butler, Godbole, & Marsh, 2013; see also Eglington & Kang, 2018; McDaniel & Little, in press). If feedback is the critical factor, then that would potentially constitute an *indirect* transfer effect (i.e., transfer that does not stem wholly from the retrieval event itself; for discussion see Roediger & Karpicke, 2006; Pan & Rickard, 2018). Learning benefits that stem solely from retrieval (e.g., as can occur for correctly answered cued recall without feedback) constitute *direct effects* of retrieval practice, whereas learning benefits stemming from activities besides the retrieval event, such as post-retrieval processing of feedback, constitute indirect effects.

1.1.2 | Does specificity of learning occur for process-based concepts?

In the studies of cued recall and transfer discussed thus far, the target materials were not process-based. For instance, none of the facts in Pan et al. (2015), which included materials drawn from AP History and Biology courses, described a sequence of processes or events. Rather, for history they focused on details of the "who, what, when, and where" variety, and for biology they typically took the form of *x of y and z* (e.g., "The exoskeletons of most insects are made of chitin."). By contrast, the current experiments explicitly involved only process-based biology concepts. There is evidence that learners better remember sentences and/or paragraphs that describe cause-and-effect relationships (e.g., Beck, McKeown, Sinatra, & Loxterman, 1991; Fillenbaum, 1971), link beginning and end states in a coherent matter (e.g., McNamara, Kintsch, Songer, & Kintsch, 1996), and/or incorporate descriptions of events in chronological order (e.g., Blount & Johnson, 1973; Clark & Clark, 1968), relative to text materials that lack such information. These effects have been attributed to improved comprehension, memory organizational processes, and other mechanisms. Process-based concepts define cause-and-effect relationships in a coherent and sequential manner. Moreover, such concepts are highly suited to be learned with explanatory diagrams (Mayer & Gallini, 1990). It is possible that more thorough learning of process-based concepts during initial study or testing will yield a richer or deeper level of knowledge representation than does learning of other types of concepts or facts, which in turn might support positive transfer of cued recall relative to restudy.

1.2 | Overview of the present experiments

Across the six experiments detailed here, we explored three distinct types of cued recall with multiple forms of feedback that might promote transfer (see Table 1 for a summary of training methods). In each experiment, subjects first studied 24 or 36 process-based concepts one at a time in random order. The concepts were drawn from a widely used undergraduate biology textbook (Freeman et al., 2014; see Appendix A for a list). To enhance conceptual understanding and to promote more comprehensive learning, each concept was presented with a visual–conceptual diagram and a glossary of term definitions (see Figure 1 for an example). Subjects were instructed to study each concept carefully and to do so at their own pace. The decision to provide diagrams and definitions was motivated by a preliminary

TABLE 1 Summary of training methods used in Experiments 1–6

Exp.	Question type	Feedback condition	Restudy condition
Term retrieval			
1	Fill-in-the-blank	Correct answer only	Whole concept
2	Fill-in-the-blank	Whole concept, process timeline	Whole concept, process timeline
Relational questions			
3	Process step + order	Whole concept	Whole concept
4	Process step + inference	Whole concept	Whole concept
Retrieval–verification–scoring			
5	Difficult fill-in-the-blank	Terms scored, whole concept copied	Whole concept copied
6	Difficult fill-in-the-blank	Terms scored, whole concept copied	Whole concept copied, with or without terms listed for study

Note. Exp., experiment.

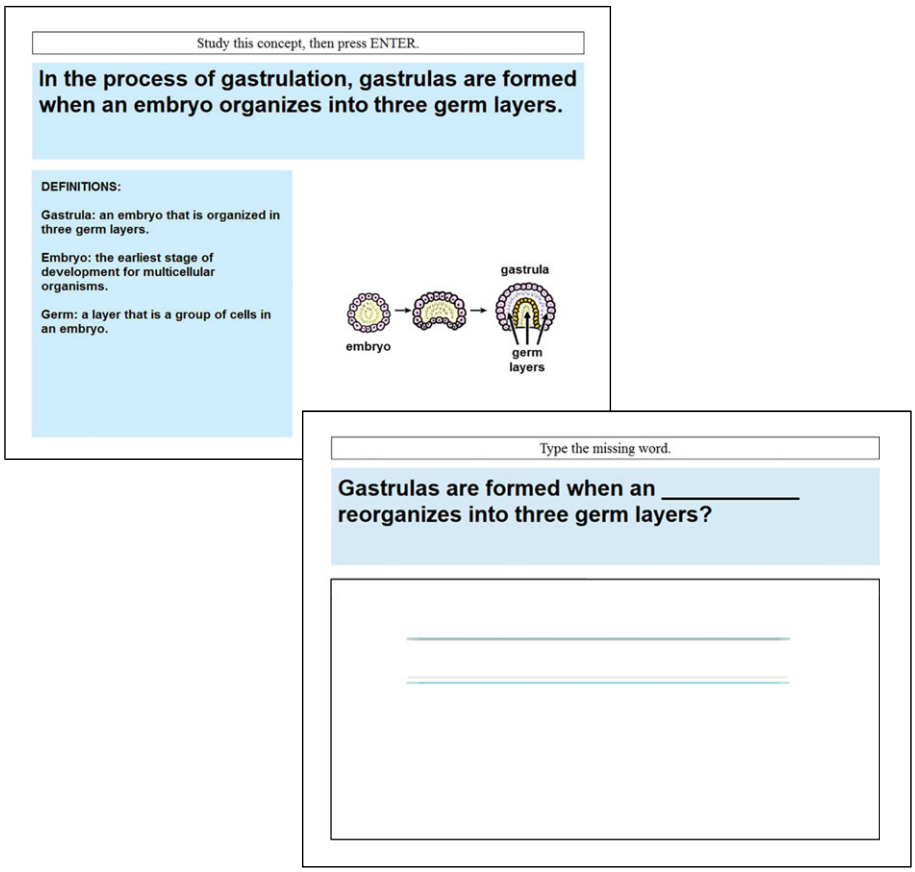

FIGURE 1 Example initial study and cued recall training test trials (top and bottom panels, respectively). During initial study, concepts are shown with diagrams and definitions. During the training phase, concepts are trained using cued recall (fill-in-the-blank example from Experiments 1 and 2 displayed) or restudied (not shown) [Colour figure can be viewed at wileyonlinelibrary.com]

experiment in which those materials were not provided during initial study and in which no transfer was observed. Although this series of experiments is systematic and is largely motivated by prior empirical or theoretical work, it is also exploratory, especially in the latter experiments.

1.2.1 | Training methods

After the initial study period, subjects completed a training phase in which each concept was trained with cued recall or restudy. Across six experiments, three retrieval practice methods were used (see Appendix B for examples): *term retrieval* (Experiments 1 and 2; fill-in-the-blank questions in which a missing term was to be retrieved), *relational questions* (Experiments 3 and 4; primarily short answer questions in which a process step needed to be retrieved, the relationship between terms needed to be identified, or inferences about the concept had to be made), and *retrieval–verification–scoring* (Experiments 5 and 6; a new method in which the entire concept, excepting one or two cue terms, had to be retrieved, followed by a series of feedback processing methods).

1.2.2 | Feedback methods

We implemented feedback in six different ways. This included three feedback types (see Appendix B for examples): *simple correct answer feedback* (Experiment 1), which involved just the term that was supposed to be retrieved (e.g., *transcription*); *process timelines* (Experiment 2), which involved a timeline of the underlying conceptual process (e.g., *DNA—transcription—RNA—translation—protein*); and *whole concept feedback* (all except Experiment 1), which involved presentation of the entire concept in text form (this could be described as explanatory feedback given that it explains an entire concept). Feedback was presented and processed using three different methods: *immediate feedback* (all except Experiment 2), *delayed and repeated feedback* (Experiment 2), and feedback in which subjects checked terms, scored their performance on those terms, and then copied the entire presented concept (i.e., *retrieval–verification–scoring*; Experiments 5 and 6). Those which involved more than immediate correct answer feedback can be classified as *elaborative* forms of feedback (Butler et al., 2013; see Kulhavy & Stock, 1989, for a taxonomy of feedback methods). Any of these methods might enhance indirect effects of retrieval practice.

1.2.3 | Criterial test

Two days after training, subjects completed a self-paced short answer criterial test. This criterial test, modeled after that used in Pan et al. (2015) and Pan, Wong, Potter, Mejia, and Rickard (2015), was identical across all experiments and featured three contiguous blocks in which each concept was tested once per block and with a different to-be-retrieved term per block (e.g., for *protein synthesis*, *transcription*, *translation*, and *RNA* were assessed; see Appendix B for examples). The instructions directed subjects to recall the exact terms that they had previously learned for each concept. No feedback was provided.

Based on prior literature, it was expected that there would be a substantial benefit of cued recall on the criterial test for terms that were previously tested (*tested-same* condition), relative to terms from concepts that were restudied (*restudied* condition). The primary question of interest was whether there would be a benefit of cued recall for previously untested terms from tested concepts (*tested-different* condition) versus the restudied condition and under what training conditions such a benefit would manifest.

2 | EXPERIMENT 1

The first experiment served to assess whether the cued recall and feedback methods used in prior studies of cued recall and transfer across terms from facts could yield transfer for process-based biology concepts (that were supplemented by diagrams and definitions during initial study). It was the first of two experiments in which *term retrieval* questions were used during training. *Simple correct answer feedback* was also implemented in this experiment only.

2.1 | Method

2.1.1 | Subjects

In all experiments, subjects were recruited from the same subject pool and identically compensated. In Experiment 1, 58 undergraduate students recruited from the subject pool at the University of California, San Diego participated for course credit. Data from one subject were excluded due to computer error, leaving 57 subjects' data for analysis. That sample size well exceeded our per-experiment target (a sample size of 45 is needed to achieve 0.85 power to detect a mean proportion correct difference between the tested-same or tested-different condition and the restudied condition of 0.04 or greater on a one-tailed, one-sample t test at $\alpha = 0.05$, based on the a priori power analysis reported in Pan et al., 2015).

2.1.2 | Design and procedure

In this and all subsequent experiments, the independent variable, namely, training via cued recall testing versus restudy, was manipulated within-subjects. There were two experimental sessions. In the first session, subjects were first informed that their task was to learn a series of biology concepts. An initial study period followed in which 36 concepts were studied one at a time as previously described. This was followed by a training phase in which each concept was tested or restudied once for 10 s each. Testing and restudy trials were randomly ordered within a single contiguous training block. Testing involved a single fill-in-the-blank test question with immediate correct answer feedback (presented for 8 and 2 s, respectively) and restudy involved viewing the concept in sentence form. After a 2-day delay, subjects returned for the criterial test as previously described. At the outset of both sessions, subjects were told to spell as accurately as possible when typing their responses, but if they were not sure, to still make their best attempt at an answer.

2.1.3 | Materials

The materials consisted of 36 process-based biology concepts drawn from Freeman et al. (2014). Each concept was a biological process involving at least two discrete steps and took the form of a single sentence with a mean length of 18 words. The mean Flesch–Kincaid reading grade level of the concepts was 12.8. Three essential terms that represented crucial components or steps were identified for each concept (e.g., *transcription*, *translation*, and *RNA*). There was largely minimal overlap in content and no terms shared between concepts. Three fill-in-the-blank training questions and three corresponding short answer criterial test questions, each targeting one of those terms, were created for each concept. The criterial test questions were identical to the training questions except for the use of the word *WHAT* in place of the answer blank. In summary, the training questions, restudy training items (i.e., concepts in sentence form), and criterial test questions were identical for each concept except: (a) one term was missing in each training and criterial test question, with the choice of missing term dependent on criterial test condition (tested-same, tested-different, or restudied); (b) the missing term

being replaced by a blank or the word *WHAT*; and (c) the use of a concluding period or question mark.

Visual–conceptual diagrams and term definitions

During initial study, each concept was presented in sentence form and accompanied by a visual–conceptual diagram and a glossary of term definitions (see Figure 1). Both appeared only during that portion of the experiment and were intended to aid subjects' understanding of the concepts, many of which included unfamiliar jargon (e.g., terms such as *cochlea*, *effector*, and *phospholipid*). The diagram consisted of a colour drawing of conceptual processes with arrows specifying individual steps or movements. Labeled representations of each of the three essential terms, plus any other jargon terms or important steps, were included with each diagram. The glossary defined each of the three essential terms for each concept plus any other jargon terms that were also present. Each term on the diagram and in the glossary was present in exact matching form in the concept sentence.

Training and criterial test lists

Each subject trained using one of six counterbalanced training lists. There was one training trial per concept on each list. Half of the concepts were tested using a fill-in-the-blank question, and the remainder was restudied. Assignment of concept to cued recall testing or restudy, and the question that was used for each concept (i.e., the term needed to be retrieved), was counterbalanced using a Latin square. Six additional test lists, each corresponding to one of the training lists, were used for the criterial test. Each criterial test list contained three blocks of 36 short answer questions each, with each concept assessed once per block and on one of the three terms per block. The three-block design enabled assessment of all three essential terms per concept and further allowed us to investigate the stability of any observed learning patterns across repeated criterial test trials per concept. There were 6 tested-same questions, 12 tested-different questions, and 18 restudied questions per block on the criterial test—numbers reflecting the fact that half of the items were restudied during training and that of the other half that were tested, one essential term out of three per concept was tested.

2.1.4 | Data coding and analysis

Training and criterial test data were computer scored. Typed responses that exactly matched the correct answer (ignoring capitalization) were scored as correct. In this and all subsequent experiments, we performed two planned orthogonal contrasts that were motivated by prior findings of no transfer: (a) an analysis of variance (ANOVA) with factors of tested-same vs. not tested (i.e., the tested-same condition vs. the tested-different and restudied conditions combined) and block and (b) an ANOVA limited to data from the tested-different and restudied conditions only, with factors of tested-different vs. restudied, and block. This latter contrast directly tested for transfer relative to restudy. We used a significance criterion of $\alpha = 0.05$ for all statistical analyses.

2.2 | Results

2.2.1 | Initial study and training

Subjects spent an average of 19 min studying the concepts or about 31 s per concept. During training, they typically retrieved the correct answer to less than one-third of the questions ($M = 0.27$) prior to viewing feedback. This relatively low rate of retrieval success (as compared with typical rates in the literature; catalogued in Pan & Rickard, 2018) underscores the difficulty of learning process-based biology concepts despite the presence of diagrams and definitions during initial study.

2.2.2 | Criterial test

Results are presented in Figure 2a. An important initial observation is that criterial test performance across all blocks in the tested-same condition was substantially higher than that during training ($M = 0.50$ vs. $M = 0.27$), confirming the expectation that training yielded substantial learning in that condition. In the first ANOVA contrast with factors as previously described, there was a main effect of tested-same vs not tested (Table 2). This is consistent with a retrieval practice effect for previously tested items as is evident in Figure 2. There was also a main effect of Block and no interaction. This is consistent with a performance improvement across blocks, an expected pattern given that the answers to criterial test questions in blocks 2 and 3 were viewable in preceding blocks. In the second contrast, there was no main effect of tested-different vs. restudied, which is consistent with there being no benefit of cued recall for untested terms relative to restudy. In other words, there was no evidence of positive transfer.

2.2.3 | Effect of prior course experience

In this and subsequent experiments, exit surveys revealed that 53–77% of subjects had university level biology or AP Biology course experience. Those with experience tended to perform better overall but transfer patterns did not differ substantially between groups (similar to Pan et al., 2015; Pan & Rickard, 2017). Thus, we do not further discuss results as a function of course experience.

2.3 | Discussion

In this experiment, although there was a benefit of cued recall for tested terms, there was no evidence of transfer of that learning to the tested-different condition. This result replicates the findings of Pan et al. (2015) using nonprocess–based materials supplemented by diagrams and definitions and with similar training methods. Further, in two unpublished follow-up experiments not detailed here, we observed the same results when a different term had to be retrieved on each of two training blocks, similar to Pan, Wong, et al. (2015; Experiment 2), and when two terms had to be retrieved per trial, similar to Hinze and Wiley (2011; Experiments 1 and 2) but with feedback. Thus, the benefits of cued recall combined with simple correct answer feedback appears to be specific to tested terms for

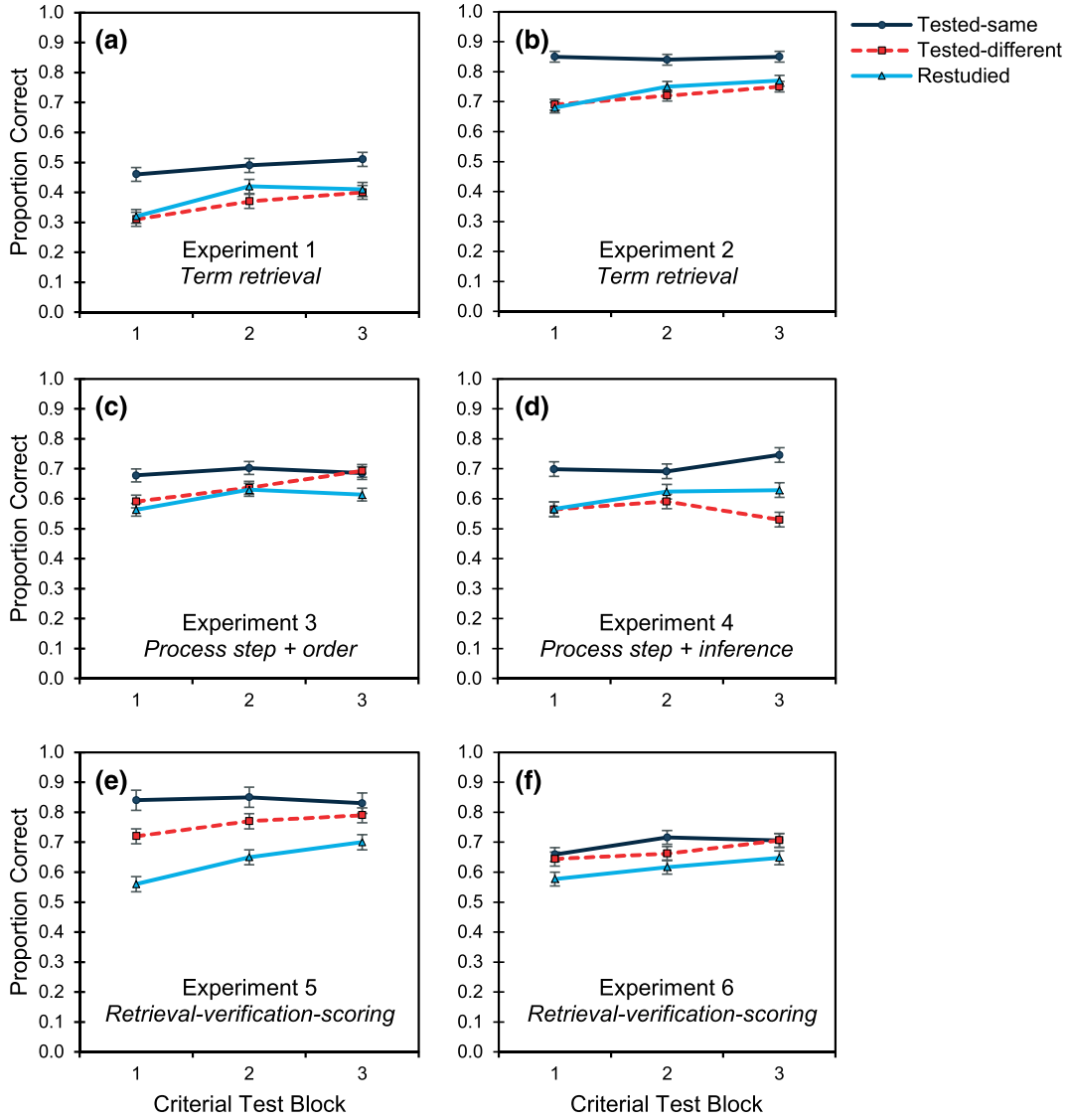

FIGURE 2 Criterial test results of Experiments 1–6. The error bars are standard errors based on the interaction error term of a within-subjects analysis of variance on subject-level mean accuracy scores (based on Loftus & Masson, 1994) [Colour figure can be viewed at wileyonlinelibrary.com]

the case of process-based concepts even in the context of diagram- and definition-augmented study, extending our prior results.

3 | EXPERIMENT 2

In the literature there has been one successful demonstration of transfer from tested to untested biology terms, relative to restudy, using fill-in-the-blank cued recall tests. In McDaniel et al. (2007), students enrolled in a university-level brain and behavior course initially learned target materials via assigned readings and classroom lessons. They then completed practice quiz questions online over a period of 3 weeks and in preparation for a later unit test. Question-by-question feedback consisting of the attempted answer and the correct answer was provided only after each quiz was submitted. That feedback could be viewed repeatedly until the date of the unit test, on which positive transfer was observed relative to restudy (albeit without strict controls for time-on-task during training in the testing and restudy conditions). None of those design features were present in our initial experiment. Thus, to bridge the differences between Experiment 1 and that prior work, and to further improve the level of learning that subjects achieved for each concept with the aim of observing transfer, in Experiment 2 we implemented multiple exposures to each concept during a standalone initial study session (designed to approximate lessons and readings that students may complete before using retrieval practice), had subjects train on those concepts with *delayed and repeated feedback*, and included *whole concept feedback* in which *process timelines* were shown. This design was intended to conceptually replicate McDaniel, Anderson, et al.'s experiment in a laboratory setting.[1]

[1]In McDaniel, Anderson et al. (2007), cued recall and multiple-choice quiz formats were used across different weeks, with assignment of materials to each quiz counterbalanced. We focused on the cued recall results here as they are most relevant.

TABLE 2 Criterial test analysis of variance (ANOVA) results

Exp.	ANOVA type	Factor	*df*	*F*	*MSE*	*p*	η_p^2
Term retrieval							
1	Contrast: tested-same vs not tested						
		Main effect of tested-same vs. not tested	1.56	34.14	1.55	<0.0001***	0.38
		Main effect of block	2.112	9.096	0.29	0.00022***	0.14
		Interaction	2.112	0.60	0.015	0.55	0.011
	Contrast: tested-different vs. restudied						
		Main effect of tested-different vs. restudied	1.56	2.43	0.04	0.13	0.042
		Main effect of block	2.112	16.23	0.26	<0.0001***	0.22
		Interaction	2.112	0.89	0.018	0.46	0.019
2	Contrast: tested-same vs not tested						
		Main effect of tested-same vs. not tested	1.42	50.4	1.26	<0.0001***	0.55
		Main effect of block	2.84	3.11	0.078	0.050	0.069
		Interaction	2.84	2.37	0.056	0.099	0.053
	Contrast: tested-different vs. restudied						
		Main effect of tested-different vs. restudied	1.42	0.88	0.013	0.35	0.021
		Main effect of block	2.84	7.044	0.13	0.0015**	0.14
		Interaction	2.84	0.83	0.010	0.44	0.019
Exp.	**ANOVA type**	**Factor**	***df***	***F***	***MSE***	***p***	η_p^2
Relational questions							
3	Contrast: tested-same vs not tested						
		Main effect of tested-same vs. not tested	1.43	14.17	0.40	0.00050***	0.25
		Main effect of block	2.86	5.78	0.11	0.0044**	0.12
		Interaction	2.86	2.046	0.35	0.14	0.045
	Contrast: tested-different vs. restudied						
		Main effect of tested-different vs. restudied	1.43	3.14	0.095	0.083	0.068
		Main effect of block	2.86	5.48	0.14	0.0058**	0.11
		Interaction	2.86	1.21	0.031	0.30	0.027
4	Contrast: tested-same vs not tested						
		Main effect of tested-same vs. not tested	1.43	46.79	1.45	<0.0001***	0.52
		Main effect of block	2.86	1.15	0.029	0.32	0.026
		Interaction	2.86	2.77	0.051	0.068	0.061
	Contrast: tested-different vs. restudied						
		Main effect of tested-different vs. restudied	1.43	9.21	0.13	0.0041***	0.18
		Main effect of block	2.86	1.42	0.041	0.25	0.032
		Interaction	2.86	1.76	0.054	0.18	0.040
Exp.	**ANOVA type**	**Factor**	***df***	***F***	***MSE***	***p***	η_p^2
Retrieval–verification–scoring							
5	Contrast: tested-same vs not tested						
		Main effect of tested-same vs. not tested	1.44	80.00	1.77	<0.0001***	0.65
		Main effect of block	2.88	5.95	0.14	0.0038**	0.12
		Interaction	2.88	6.13	0.11	0.0032***	0.12
	Contrast: tested-different vs. restudied						
		Main effect of tested-different vs. restudied	1.44	20.77	1.029	<0.0001***	0.32
		Main effect of block	2.88	8.55	0.25	0.00040***	0.13
		Interaction	2.88	1.085	0.024	0.34	0.024
6	Contrast: tested-same vs not tested						
		Main effect of tested-same vs. not tested	1.39	8.40	0.22	0.0061**	0.18
		Main effect of block	2.78	4.029	0.11	0.022*	0.094
		Interaction	2.78	0.60	0.015	0.55	0.015
	Contrast: tested-different vs. restudied						
		Main effect of tested-different vs. restudied	1.39	7.54	0.19	0.0091***	0.16
		Main effect of block	2.78	3.18	0.089	0.047*	0.075
		Interaction	2.78	0.078	0.0022	0.93	0.0020

Note.

*$p < 0.05$. **$p < 0.01$. ***$p < 0.0001$.

3.1 | Method

3.1.1 | Subjects

Forty-seven subjects participated for course credit. Data from 4 students were excluded, 2 due to a computer error and 2 due to not following instructions, leaving 43 subjects that were included in the data analysis

3.1.2 | Design and procedure

Uniquely in this experiment, there were three experimental sessions. The first session was entirely devoted to initial study of the concepts. Subjects began that session by cycling through each concept repeatedly for 30 min. During that period the concepts and term definitions were presented in paragraph form (see Appendix C for an example) with supporting diagrams. Next, subjects cycled through each concept

repeatedly for an additional 30 min. During this time each concept was presented only in sentence form

Training methods

The second session occurred 48 hr after the first. During this session, subjects trained on the concepts using fill-in-the-blank tests or restudy across separate blocks, presented in random order. Each training block lasted 12 min. During the test block, subjects answered one fill-in-the-blank question per tested concept. After all questions were answered, they viewed question-by-question feedback. That feedback consisted of the whole concept in sentence form, their attempted answer, and a process timeline. They were instructed to cycle through that feedback repeatedly at their own pace until time had elapsed.

The restudy block had a similar structure. Subjects first viewed each restudied concept once in sentence form. They next viewed those concepts again in sentence form and with process timelines, similar to the test condition. They were instructed to cycle through those concepts repeatedly at their own pace until time had elapsed. The second session ended after both training blocks were finished. Finally, after a second 48-hr delay, subjects returned for a criterial test that used the same design as that of the prior experiment.

Learning instructions

Uniquely in this experiment, the instructions at the outset of session two informed subjects that they would be training for an exam in session three. Moreover, at the outset of the test block, they were told that the practice questions resembled exam questions, but that any part of a tested concept could be assessed on the exam; thus, they should learn entire concepts. During the restudy block, subjects were told to study.

3.1.3 | Materials

The materials consisted of 24 biology concepts drawn from the set used in the preceding experiment, a reduction due to the logistics of experiment scheduling, and with training and criterial lists altered to match. For initial study, a textbook paragraph-style version of each concept, using short sentences and with term definitions embedded within the paragraph, was created. The paragraphs had a mean length of 79 words and a Flesch–Kincaid reading grade level of 11.3. For feedback, a timeline of process steps was created for each concept as previously described. All other materials were identical to those in the prior experiment.

3.2 | Results and Discussion

3.2.1 | Initial study and training

During the first and second halves of the initial study period, respectively, subjects viewed each concept an average of 6 and 4 times (for 21 and 31 s each). During training, subjects cycled through delayed feedback an average of 8 times per tested concept (11 s per concept) and 7 times each per restudied concept (16 s per concept). Subjects' training accuracy was $M = 0.62$, a substantial improvement over Experiment 1. This is especially notable given that training took place 48 hr after initial study, which allowed more time for forgetting to have taken place, and suggests that the stand-alone initial study session enabled subjects to attain a higher level of understanding of each concept.

3.2.2 | Criterial test

Results are presented in Figure 2b; ANOVA results are listed in Table 2. As is evident in the figure, criterial test performance was improved over the preceding experiment across all conditions. Moreover, for the tested-same condition, performance is near ceiling. These results likely reflect the more extensive initial study and training methods that were used. However, the now-familiar pattern, namely, a retrieval practice effect for tested-same terms and minimal transfer to tested-different terms relative to the restudied condition, was still observed. Thus, in this experiment we were unable to fully reproduce the findings of McDaniel et al. (2007) in a laboratory setting and despite using many of the same training methods (albeit without learning in the context of an actual class, or training on multiple sessions across several weeks—issues that we further address in section 8). Overall, the results of both Experiments 1 and 2 reinforce our conclusion that cued recall in the form of simple fill-in-the blank questions tends to yield highly specific learning for tested stimulus–response combinations. Moreover, the results of Experiment 2 rule out the hypothesis that delayed and repeated feedback that includes more than just the correct answers (i.e., to fill-in-the-blank questions) is always sufficient to yield indirect transfer to untested terms.

4 | EXPERIMENT 3

The third experiment was the first of two experiments in which we investigated retrieval practice methods that require learners to consider connections between different components of a concept. In the literature, some researchers have suggested that questions, which require subjects to infer relationships or retrieve more than just an isolated portion of a fact or concept, yield better overall understanding and may improve certain types of transfer (e.g., Jensen et al., 2014; Karpicke & Aue, 2015). Accordingly, in this experiment each tested concept was trained 4 times using two question types: *process step* questions, which required retrieval of half of each concept from memory, and *order* questions, which required the determination of the ordinal relationship of two terms from a concept. These can be categorized as *relational questions* in that they involve thinking about the sequential or chronological nature of a process or a cause-and effect relationship. The use of two question types also introduced variation (i.e., potential encoding variability) in the types of retrieval practice performed for each tested concept.

4.1 | Method

4.1.1 | Subjects

Fifty-four subjects participated for course credit. Data from four subjects was excluded due to computer errors, and seven subjects did not return for the second session, leaving 43 subjects' data for analysis.

4.1.2 | Design and procedure

The primary difference in this experiment was the training phase design. Subjects completed four successive training blocks in which each concept was tested or restudied once per block for 20 s (and for test trials, 10 s for retrieval and 10 s for feedback). Test and restudy trials were randomly intermixed in each block, with each concept trained four times. Moreover, for tested concepts, the process step and order question for that concept was presented a total of 2 times (once in every other block). With this experiment we also reverted to having both initial study and training occur in an initial session.

Training methods

For process step questions, subjects were given the name of the concept and asked, "What occurs in the first step? Answer completely with all details." (or that question but with s*econd* instead of *first*). For order questions, subjects were presented with the name of the concept, two terms in alphabetical order, and asked, "Which of these occurs first in the process?" (or that question but with *second* instead of *first*). Thus, order questions constituted a form of two-alternative forced choice question, marking the sole deviation from cued recall in the present experiments. Importantly, each tested concept was trained with one process step question and one order question, with each concept tested once per block and on one question type per block. Whole concept feedback was provided immediately after each retrieval attempt. On restudy trials, subjects viewed entire concepts in sentence form.

Learning instructions

To answer test questions, subjects were told to think of the entire concept, including its constituent processes, to recall exact words, and to check feedback word-by-word for complete comprehension. For restudy trials, subjects were told to study.

4.1.3 | Materials

To facilitate process step questions, all 36 concepts were reworded into two separate sentences that identified the two main steps for each concept in sequence using the connective words *first* and *second* or *as a result* (e.g., protein synthesis involves two steps. In the first, DNA is first copied into RNA via transcription. In the second, RNA is coded into protein via translation.). The sentences had a combined mean length of 17 words and a Flesch–Kincaid reading grade level of 7.6, the reduction in grade level due to the added connective words. Each concept was presented in two-sentence form throughout the experiment

Two process step and two order questions were created for each concept, one per step, with the assignment of question to concept counterbalanced across four training lists. The names and terms of several concepts were modified to avoid overlapping between cues and responses. Each order question required subjects to select between two terms, one from each step. For each subject, the correct answer to the process step and order questions for each tested concept referred to the same step (e.g., if retrieval of the first step was required for the former question, then identifying a term from the first step was required for the latter question). On the criterial test, which was identical at the trial level to that of the prior experiments, two terms from that step were assessed in the tested-same condition and one term from the other step (i.e., the lure in the order question) was assessed in the tested-different condition (cf. Little, Bjork, Bjork, & Angello, 2012). There were 12 tested-same questions, 6 tested-different questions, and 18 restudied questions per criterial test block.

4.1.4 | Data coding

All process step questions were scored by raters blind to condition using a rubric of idea units for each concept (following the method detailed in Pan & Rickard, 2017). To assess consistency, 5% of all process step questions were rescored by additional raters, with ambiguous cases resolved by discussion; 85% of scores matched between raters. All other questions were computer scored in the same manner as in the preceding experiments.

4.2 | Results and Discussion

4.2.1 | Initial study and training

Subjects spent an average of 23 min, or 38 s per concept, during initial study. During training, mean performance improved from the first to second administrations of each process step and order question (from M = 0.36 to 0.54 and from M = 0.78 to 0.87, respectively).

4.2.2 | Criterial test

Results are presented in Figure 2c; ANOVA results are detailed in Table 2. The same overall pattern as in the prior experiments was again observed. Mean performance in the tested-different condition reached parity with the tested-same condition by block 3, but the interaction with block was not significant. Moreover, across blocks 1 and 2, there was little indication of positive transfer, and the first block constitutes the purest measure of learning from the first session. Thus, although process step and order questions required a greater understanding of the relationship between different components of each tested concept than term retrieval questions and also targeted each concept in different ways (albeit with some overlap in content), using those questions in conjunction with whole concept feedback still yielded specific learning as in the earlier experiments.

5 | EXPERIMENT 4

In the fourth experiment, we continued our investigation of relational questions by having subjects train on concepts using process step questions and *inference questions*. This latter question type required subjects to integrate information about each tested concept in a different way from that in which it was originally presented, which might promote greater conceptual understanding and improve transfer (Jensen et al., 2014).

5.1 | Method

5.1.1 | Subjects

Forty-nine subjects participated for course credit. Data from five subjects were excluded due to computer or experimenter error and 1 subject did not return for the second session, leaving 43 subjects' data for analysis.

5.1.2 | Design and procedure

The design and procedure were identical to that of the preceding experiment, including four training repetitions per concept, except that inference questions were used in place of order questions. Directions to think of entire concepts, recall exact words, and completely check feedback were provided for test trials, just as in Experiment 3

5.1.3 | Materials

One of the authors with experience as a biology section instructor created an inference question for each of the 36 concepts. These questions replaced the order questions on the training lists. Depending on the concept, the inference question required integrating multiple pieces of information (e.g., "what is the purpose of plasmogamy in fungal cells?"), identifying a critically important component or process (e.g., "the process of gene flow affects what aspect of a population?"), or determining a hypothetical state of a conceptual process (e.g., "if protein synthesis could be reversed, what molecule would you end up with?"). The correct answers to these questions, which averaged between one and two words in length, did not necessarily refer to any of the essential terms for the tested concept (and never referred to all essential terms per concept, thus leaving those terms to be assessed in the tested-different condition), but could be derived from the provided whole concept feedback.

5.1.4 | Data coding

All process step questions were scored using identical procedures as in the prior experiment, with those procedures adapted for scoring inference questions. Of the 5% of questions rescored by additional raters, scores matched for 81% and 91% of process step and inference questions, respectively.

5.2 | Results and Discussion

5.2.1 | Initial study and training

Subjects spent an average of 23 min, or 37 s per concept, during initial study. During training, mean performance improved from the first to second administrations of each process step and inference question (from M = 0.38 to 0.57 and from M = 0.51 to 0.70, respectively).

5.2.2 | Criterial test

Results are presented in Figure 2d; ANOVA results are detailed in Table 2. The same overall pattern as in the prior experiments was observed yet again. Moreover, mean performance in the tested-different condition was lower than the tested-same and restudied conditions in block 3. That apparent negative transfer was evident in the orthogonal contrasts (Table 2) but is an anomalous result in the context of all prior results and may thus reflect random variance. Overall, the results of Experiments 3 and 4 are inconsistent with the hypothesis that questions which involve inferring relationships or retrieving elements generally yield transferrable test-based learning, at least for the case of transfer to untested terms.

6 | EXPERIMENT 5

Given the prior failures to observe positive transfer, we next investigated more extensive cued recall and feedback techniques. We hypothesized that subjects (a) had to be made aware of the importance of including multiple exact terms in their retrieval attempts through actual experience and not just through instructions, (b) their attention had to be specifically directed to each essential term in each tested concept, and (c) they had to focus on both while processing feedback. To achieve these goals, we drew inspiration from Rawson and Dunlosky's (2011) use of *retrieval-monitoring-feedback* trials in which subjects had to check each of several main ideas per retrieved definition; without such checking, subjects could erroneously regard incomplete answer attempts as fully correct. We ultimately developed a new training technique, *retrieval–verification–scoring*, which involved recalling nearly entire concepts and then checking terms, scoring terms, and viewing and copying whole concepts. This was a far more time-consuming procedure than in prior experiments—potentially boosting transfer at the cost of efficiency and complexity. It involved more extensive, and likely more effortful, retrieval attempts for each tested concept. Further, given the extensive use of feedback on each test trial, any observed transfer could be attributed at least partly to indirect effects of retrieval practice.

Uniquely in this experiment, rather than implement a total time limit across the test and restudy conditions (which given the more time-consuming nature of retrieval–verification–scoring trials under self-paced training would have yielded large disparities in the number of trial repetitions per concept), we equated the total trial repetitions per concept across training conditions (i.e., three repetitions each). Further, to equate the amount of word-for-word copying that

occurred in both conditions, we implemented copying during both test and restudy trials.

6.1 | Method

6.1.1 | Subjects

Fifty-seven subjects participated. Data from six subjects was excluded due to computer or experimenter errors and five subjects did not complete the second session, leaving 46 subjects' data for analysis.

6.1.2 | Design and procedure

Only the training phase substantially differed from the preceding experiments. Subjects completed three successive test blocks and three successive restudy blocks (with the starting order of test or restudy randomly assigned) and with each block having exactly one trial per concept with no time limit. Thus, subjects completed a total of three tests or three restudy trials per concept.

Training methods

Each retrieval–verification–scoring trial involved the following steps (see Figure 3 for an example). First, subjects attempted to answer a difficult fill-in-the-blank question, which required retrieval of nearly the entire concept except one essential term. After submitting their answer, the feedback screen appeared. On that screen, the answer attempt was displayed along with the three essential terms for that concept, one of which was the cue in the preceding question. Subjects had to verify whether they had correctly typed each of those terms by marking yes or no. Although subjects received instructions to spell accurately but still attempt responses if unsure (just as in the preceding experiments), they were told that

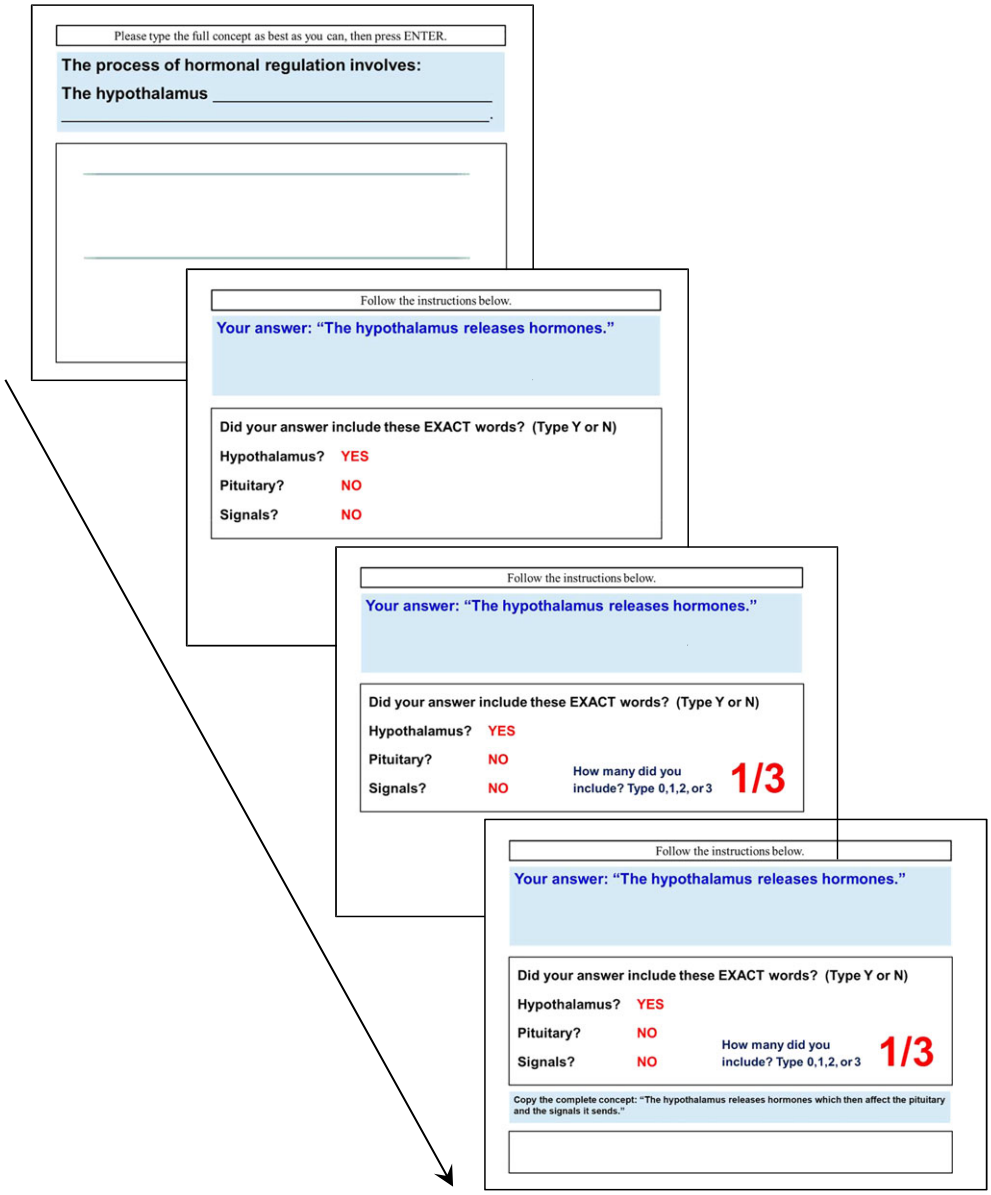

FIGURE 3 Example *retrieval–verification–scoring* trial. First panel: difficult fill-in-the-blank question. Second panel: subjects check their answer for essential terms. Third panel: numerical scoring. Fourth panel: whole concept presented for word-for-word copying [Colour figure can be viewed at wileyonlinelibrary.com]

their answer had to exactly match to be scored as correct. Next, they counted how many terms out of three that they had correctly recalled and typed that number as their score. This included the essential term that was present in the question. Finally, the whole concept was presented for subjects to copy. By comparison, in the restudy condition, subjects simply viewed and copied whole concepts.

Learning instructions

In the test condition, subjects were encouraged to fully learn each concept and to improve their scores over multiple attempts. In the restudy condition, subjects were told to study.

6.1.3 | Materials

The materials consisted of 24 biology concepts in one-sentence form as in the first two experiments. Three difficult fill-in-the-blank test questions were created for each concept. In each, all but one essential term and, in many cases, one article word (i.e., *a*, *an*, or *the*) that preceded that term, was replaced with a continuous blank line or several lines interspersed with punctuation marks. The choice of essential term that was presented as a retrieval cue for each concept was counterbalanced across six training lists. On the criterial test, the terms that were to be retrieved during training were assessed in the tested-same condition and the terms that were cues to training questions were assessed in the tested-different condition. There were 8 tested-same, 4 tested-different, and 12 restudied questions per criterial test block. All other materials were largely identical to those used in prior experiments.

6.1.4 | Data coding

Difficult fill-in-the-blank questions were computer scored in terms of the number of essential terms correctly recalled per concept.

6.2 | Results and Discussion

6.2.1 | Initial study and training

Subjects spent an average of 14 min, or 34 s per concept, during initial study. During training, mean subject performance improved (from $M = 0.33$ to 0.51 and 0.59) across each of three test trials per concept (i.e., recall of the two essential terms that were not present in the training questions). As expected, the more time-consuming nature of retrieval–verification–scoring trials (subjects spent an average of 84 s each vs. 26 s for restudy) resulted in a large time-on-task difference; subjects spent an average of 47 min to complete three repetitions per tested concept, versus only 15 min for restudy.

6.2.2 | Criterial test

Results are presented in Figure 2e; ANOVA results are detailed in Table 2. As is evident in the figure, there was evidence of positive transfer relative to restudy. Performance was best overall in the tested-same condition, second best in the tested-different condition, and lowest in the restudied condition. Thus, the use of retrieval–verification–scoring trials—a method that directed subjects to attend to the essential terms of each concept, highlighted the importance of correctly recalling and verifying exact terms and encouraged and provided the opportunity to closely examine entire concepts (albeit with much greater time required under self-paced training conditions)—elevated performance for both tested and untested terms above that of conventional restudy.

7 | EXPERIMENT 6

Having demonstrated a method of difficult cued recall coupled with extensively processed feedback that successfully yielded transfer relative to restudy, for the final experiment we attempted to replicate that finding under conditions of strict control for time-on-task, fixed-paced training, and equal training phase item repetitions. We also included a modified restudy control that was potentially more competitive with testing.

7.1 | Method

7.1.1 | Subjects

Forty-eight subjects participated. Data from five subjects were excluded due to computer or experimenter errors, and three subjects were unable to fully comply with instructions, leaving 40 subjects' data for analysis.

7.1.2 | Design, procedure, and materials

This experiment incorporated the design of its predecessor with the following changes: there was a fixed time limit of 70 s for both test and restudy trials (and for retrieval–verification–scoring trials, 30 s for testing and 40 s for feedback), and there were two repetitions of either testing or restudy trials for each concept across contiguous blocks (four training blocks in total; two training repetitions per concept rather than three as in the prior experiment). Thus, with this design, subjects that had not fully completed all steps of a retrieval–verification–scoring trial by the trial time limit were not permitted to do so and immediately begin the next trial. Time limits were chosen based on prior experiment data and experiment timeslot logistics. Additionally, in the restudy condition, the three essential terms were shown alongside the whole concept for one-third of restudied concepts (with the selection of concepts for which terms were shown counterbalanced across training lists). Subjects were informed that those terms, where present, were additional information for them to study. This was intended to further control for the display of terms during retrieval–verification–scoring. Other procedures and materials were otherwise identical to that of the prior experiment.

7.2 | Results

7.2.1 | Initial study and training

Subjects spent an average of 14 min, or 36 s per concept, during initial study. Performance improved over the two test trials per tested concept, from M = 0.21 to 0.41.

7.2.2 | Criterial test

Because criterial test performance in the restudied condition for concepts with essential terms displayed was not better than restudied concepts without terms displayed (M = 0.61 for both; data averaged across criterial test blocks), which suggests that the presence of those terms during some restudy trials was ineffective at improving performance, data from restudied concepts that did or did not have terms separately displayed during training were combined in the analyses. Results are presented in Figure 2f; ANOVA results are detailed in Table 2. We again observed evidence of positive transfer relative to restudy, replicating the results of the preceding experiment. Thus, the results of the Experiments 5 and 6 indicate that difficult fill-in-the-blank cued recall questions combined with extensive feedback processing—a substantially more intensive training method than that used in earlier experiments—can yield transfer from tested to untested terms.

8 | GENERAL DISCUSSION

In the forgoing experiments we explored whether a series of cued recall-based training and feedback methods could yield transfer for process-based biology concepts. This research encompassed a variety of cued recall question subtypes, the majority of which had not previously been investigated for transfer to stimulus–response rearranged items. None appeared to be sufficient alone to induce transfer. Additionally, enrichment of initial study using diagrams and definitions, extended study periods, and several different types of feedback appeared to be insufficient to generate transfer. Rather, criterial test performance in the tested-different condition was in most cases (Experiments 1–4) nearly equivalent or numerically lower than that in the restudied condition, just as in most prior work on this topic. These results underscore the nontrivial challenge of designing a task that produces transfer of learning following cued recall on a process-based concept to previously untested terms from that concept. Moreover, they suggest that specificity of learning following cued recall practice commonly extends to process-based concepts, despite their previously established different learning properties, and may indeed be a general property of cued recall practice. Ultimately, the use of far more extensive cued recall and feedback processing methods (Experiments 5 and 6) was necessary to generate transfer.

8.1 | Comparing implementations of cued recall practice with feedback

Why was the retrieval–verification–scoring method used in Experiments 5 and 6 effective at yielding transfer to stimulus–response rearranged items, whereas other methods were not? Further consideration of the cued recall and feedback methods employed across the six experiments enables us to draw conclusions about each.

8.1.1 | Term retrieval questions with correct answer or whole concept feedback

The results for simple fill-in-the-blank questions in Experiments 1 and 2 are consistent with prior results in the literature (e.g., Hinze & Wiley, 2011; Pan et al., 2015) and indicate that retrieval of one or two key terms from a fill-in-the-blank question typically yields a memory enhancement for those key terms only relative to restudy. Moreover, feedback consisting of tested key terms may aid learning in cases where they were not successfully retrieved but is unlikely to enhance transfer to other nontested key terms.

The results of Experiment 2, in which repeatedly viewed whole concept feedback was implemented, contrast with McDaniel et al.'s (2007) results involving similar training methods. Several design differences between the two experiments may account for the contrasting transfer results, although neither possibility can at present be clearly linked to a psychological mechanism. Specifically, the authentic educational context and longer learning interval (i.e., up to 3 weeks instead of 48 hr) in McDaniel, Anderson, et al. may have been pivotal. During that time, students likely took advantage of repeated and spaced feedback learning opportunities, plus studied relevant course materials. Postretrieval study of materials has been associated with successful transfer of different types than in the present work (e.g., McDaniel, Bugg, Liu, & Brick, 2015; see also Thomas, Weywadt, Anderson, Martinez-Papponi, & McDaniel, 2017). Accordingly, it appears that in the absence of the added design features of McDaniel, Anderson, et al., fill-in-the-blank questions with whole concept feedback will not yield transfer of learning to stimulus–response rearranged items. Overall, the results of Experiments 1 and 2 reinforce the conclusion that retrieving individual key terms from sentence-based facts or concepts, followed by immediate, delayed, repeated, and/or whole concept feedback, yields highly specific learning (Pan et al., 2015).

8.1.2 | Relational questions with whole concept feedback

The results of Experiments 3 and 4 reveal that specificity of learning following cued recall practice is not limited to term retrieval questions. Rather, questions which require learners to consider the order of steps in a process, examine cause-and-effect relationships, or make inferences about a concept—all of which presumably encourage at least some cognitive processing of the different elements of each concept—can yield similar results. For process step questions, there was a clear enhancement for the retrieved step (i.e., half of the concept), but that step only. The correct answers to the order and inference questions were similarly selectively enhanced. The implementation of whole concept feedback (i.e., explanatory feedback) in these experiments also did not improve transfer, just as in Experiment 2. Explanatory feedback has not always yielded strong positive transfer in prior work involving transfer to application and inference questions (e.g., it

was observed in Butler et al., 2013, but was not significant in McDaniel, Wildman, & Anderson, 2012, although there were differences in the amount of explanatory content provided). Overall, the results of Experiments 3 and 4 not only suggest limits of relational questions for transfer to stimulus–response rearranged items but also indicate that indirect transfer cannot be guaranteed using explanatory feedback.

8.1.3 | Retrieval–verification–scoring

When we did observe transfer, it required difficult fill-in-the-blank questions coupled with extensive postretrieval study-like activities. In Experiment 5, we interpreted the resulting transfer as evidence of the potentially unique efficacy of the retrieval–verification–scoring procedure, although that inference was tentative due to a potential time-on-task confound. That confound was eliminated in Experiment 6, and more complete, albeit smaller in magnitude, transfer was observed (the reduction in training repetitions may have been a key factor in the reduced effect size). Thus, based on those experiments it appears that transfer of test-based learning to stimulus–response rearranged items drawn from process-based concepts is attainable, but that such transfer may require any or all of the postretrieval activities used in the retrieval–verification–scoring procedure.

Retrieval–verification–scoring differed from the cued recall testing and feedback methods in the prior experiments in five major ways. First, subjects had to make a much more extensive retrieval attempt on each test trial by retrieving nearly the entire concept except for one essential key term. Second, the verification step directed subjects' attention to each of the essential terms per concept, including the term that was later assessed in the tested-different condition on the criterial test. Third, that step provided subjects with firsthand knowledge of the importance of fully and exactly recalling key components of each concept, including with correct spelling. Fourth, the scoring step furnished subjects with a numerical measure of their mastery of each concept. This may have corrected inaccurate judgments of learning and motivated greater learning on subsequent trials. Finally, all these feedback processing steps may have enhanced the efficacy of the final copying step. Thus, a prerequisite for transfer from tested to untested terms appears to be training procedures that (a) require more extensive retrieval of target materials (i.e., mental reconstruction of nearly the entire stimulus) and (b) deliberately focus learners' attention to tested and untested terms during feedback. This involves more than just instructional prompts, the provision of added explanatory details, or other types of elaborative information.

By our analysis, it appears that both direct (i.e., the difficult fill-in-the-blank questions) and indirect effects (i.e., the verification and scoring procedure) of retrieval practice contributed to the transfer results in Experiments 5 and 6. The two were intertwined: the extensive scoring and feedback processing procedure required first having attempted retrieval of nearly the entire concept. Moreover, it appears that neither effect is sufficient by itself to achieve transfer. For example, in Pan and Rickard (2017), practice test questions which involved retrieving whole definitions prompted by key term cues (e.g., "Consciousness is….?") coupled with simple correct answer feedback—a form of difficult retrieval similar to that used in Experiments 5 and 6, albeit with differences in test format—consistently yielded specific learning benefits. Further, elaborative forms of feedback did not improve transfer in Experiments 2–4. Ultimately, it may be that for transfer to occur, training must yield additional processing of not just directly retrieved materials but also other parts of a tested stimulus, and the combination of more extensive retrieval *and* feedback processing methods can foster that processing. It remains to be fully determined whether that additional processing is attributable to retrieval, more focused study, or both.

8.2 | Limitations, educational implications, and future work

The reader might suspect that the lack of transfer in Experiments 1 through 4 reflects the fact that subjects did not undergo deep conceptual learning or thorough processing of the different constituent elements of each concept during either the initial study or training phases. We are open to those possibilities, even in the case of extensive initial study in Experiment 2. If true, however, it does not necessarily constitute a weakness of this work from the applied perspective. We are probably not alone in our strong intuition that the onset of deep conceptual understanding in STEM fields lags the acquisition of facts and of more tenuous, and often piecewise, conceptual understanding. Further, progress through those levels may be necessary for achieving deeper understanding (Anderson & Krathwohl, 2001). In that light, the current insights regarding transfer may be especially applicable to classroom-based cued recall tests with feedback, particularly for novice and intermediate students. It is also important to recognize that the new manipulations that improved transfer in Experiments 5 and 6 may not have yielded a substantial leap in deep understanding relative to Experiments 2–4. Rather, that transfer likely reflects the greater efficacy relative to restudy of those manipulations in focusing subjects' attention on the entire stimulus presented on practice test trials compared with restudy. Future work that further explores the critical properties of activities that produce transfer relative to restudy and other non-retrieval tasks (cf. Pan, Rubin, & Rickard, 2015), and especially feedback methods such as the retrieval-verification-scoring method employed here, is warranted.

8.2.1 | Methodological considerations

We encouraged subjects to attempt responding even if they were unsure of correct spellings, but we only scored perfect spellings as correct. The reader might expect that the results would differ if minor misspellings were accepted; supplementary analyses doing just that yielded slightly higher mean performance across conditions but no discernable differences in the overall patterns of results.

A further consideration is that the samples obtained did not enable fully equated sample sizes for each counterbalanced list. To address this, we repeated the analysis sequence for each experiment but with the minimal number of subjects randomly dropped from some conditions to achieve parity across lists; the statistical outcome for the critical comparison (tested vs. not tested; tested-different vs. restudy) was unaffected in all cases.

8.2.2 | Differences with the retrieval-induced facilitation paradigm

Readers might draw comparisons between the present work and the *retrieval-induced facilitation* paradigm (Chan, McDermott, & Roediger, 2006). In that paradigm, subjects typically read a text passage and then train using cued recall without feedback or restudy on facts drawn from that passage. On a subsequent criterial test, there is usually a benefit of cued recall for questions that assess related but previously untested facts. That finding has been interpreted as evidence that retrieval practice yields improved processing of semantically related content, possibly due to a process of spreading activation (Chan et al., 2006). The critical difference between that paradigm and the present work is that we assessed transfer to terms that were present during initial test or restudy trials, rather than from facts that were only seen during an initial study phase. Yet other transfer paradigms may also yield different results.

8.2.3 | Further cued recall training methods

Given the limitations of the majority of cued recall and feedback methods observed in the present work, an alternative strategy may be to test on all possible responses where feasible (for prior discussion see Pan et al., 2015). That method stands to yield retrieval practice benefits across the entirety of tested materials. Alternatively, our demonstration of retrieval–verification–scoring represents a proof-of-concept for more complex (but potentially more effective, contingent on learning objective) forms of cued recall practice that may be useful not only for improving transfer across terms for process-based concepts and other stimuli (e.g., multiterm facts) but also for computer-based training programs more generally. Such methods could be refined via the use of automated scoring procedures, adaptive algorithms that tailor training schedules to performance levels, or even gamification that incorporates the goal of score improvement over training trials. As retrieval practice research increasingly addresses issues of practical application, more sophisticated implementations of cued recall may emerge as a viable alternative to conventional cued recall training methods.

ORCID

Steven C. Pan https://orcid.org/0000-0001-9080-5651
Sarah A. Hutter https://orcid.org/0000-0003-0661-2086

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.

Beck, I. L., McKeown, M. G., Sinatra, G. M., & Loxterman, J. A. (1991). Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly*, *26*, 251–276. https://doi.org/10.2307/747763

Blount, H. P., & Johnson, R. E. (1973). Grammatical structure and the recall of sentences in prose. *American Educational Research Journal*, *10*(2), 163–168. https://doi.org/10.3102/00028312010002163

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology*, *105*(2), 290–298. https://doi.org/10.1037/a0031026

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. https://doi.org/10.1002/acp.1507

Chan, J. C., McDermott, K. B., & Roediger, H. L. III (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553

Clark, H. H., & Clark, E. V. (1968). Semantic distinctions and memory for complex sentences. *Quarterly Journal of Experimental Psychology*, *20*(2), 129–138. https://doi.org/10.1080/14640746808400141

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

Eglington, L. G., & Kang, S. H. (2018). Retrieval practice benefits deductive inference. *Educational Psychology Review*, *30*(1), 215–228. https://doi.org/10.1007/s10648-016-9386-y

Fillenbaum, S. (1971). On coping with ordered and unordered conjunctive sentences. *Journal of Experimental Psychology*, *87*(1), 93–98. https://doi.org/10.1037/h0030177

Freeman, S., Quillin, K., & Allison, L. (2014). *Biological science*. San Francisco, CA: Benjamin-Cummings Publishing Company.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory*, *19*(3), 290–304. https://doi.org/10.1080/09658211.2011.560121

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. https://doi.org/10.1007/s10648-013-9248-9

Kang, S. H., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory and Cognition*, *3*(3), 183–188. https://doi.org/10.1016/j.jarmac.2014.05.006

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*(4), 279–308. https://doi.org/10.1007/BF01320096

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. https://doi.org/10.1177/0956797612443370

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490. https://doi.org/10.3758/BF03210951

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology*, *82*(4), 715–726. https://doi.org/10.1037/0022-0663.82.4.715

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. https://doi.org/10.1080/09541440701326154

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied*, *21*(4), 370–382. https://doi.org/10.1037/xap0000063

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

McDaniel, M. A., & Little, J. L. (in press). Multiple-choice and short-answer quizzing on equal footing in the classroom: potential indirect effects of testing. In J. Dunlosky, & K. Rawson (Eds.), *Handbook of cognition and education*. Cambridge University Press.

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, *1*(1), 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1–43. https://doi.org/10.1207/s1532690xci1401_1

Pan, S. C., Gopal, A., & Rickard, T. C. (2015). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology*, *107*(4).

Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language*, *83*, 53–61. https://doi.org/10.1016/j.jml.2015.04.001

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

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

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. https://doi.org/10.1037/xap0000062

Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., & Rickard, T. C. (2015). Does test-enhanced learning transfer for triple associates? *Memory & Cognition*, *44*(1).

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283–302. https://doi.org/10.1037/a0023956

Rickard, T. C., & Pan, S. C (2017). Test-enhanced learning of pairs, triplets, and facts: When and why does transfer occur? Unpublished manuscript.

Roediger, H. L. III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210.

Roediger, H. L. III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. P. Mestre & B. H. Ross (Eds.), *The psychology of learning and motivation: Vol. 55. The psychology of learning and motivation: Cognition in education* (pp. 1–36). San Diego, CA: Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

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

Thomas, R. C., Weywadt, C. R., Anderson, J. L., Martinez-Papponi, B., & McDaniel, M. A. (2017). Testing encourages transfer between factual and application questions in an online learning environment. *Journal of Applied Research in Memory and Cognition*.

How to cite this article: Pan SC, Hutter SA, D'Andrea D, Unwalla D, Rickard TC. In search of transfer following cued recall practice: The case of process-based biology concepts. *Appl Cognit Psychol*. 2019;1–17. https://doi.org/10.1002/acp.3506

APPENDIX A

LIST OF PROCESS-BASED BIOLOGY CONCEPTS

Concept no.	Concept name	Example essential terms
1	Active transport	antiporter, hydrogen, sodium
2	Audioception	cochlea, hair, pinna
3	Baleen filtration	filter, plankton, whale
4	Blood flow	capillaries, oxygen, vein
5	Endosymbiotic theory	endosymbiosis, mitochondria, protist
6	Enzyme catalysis	enzyme, product, substrate
7	Exocytosis	plasma, vesicle, waste
8	Gastrulation	embryo, gastrula, germ
9	Hormonal signaling	hypothalamus, pituitary, signals
10	Immune response	clone, infection, lymphocyte
11	Membrane formation	bilayer, phospholipid, tails
12	Movement initiation	CNS, effector, receptor
13	Muscle contraction	actin, discs, myosin
14	Neuronal signaling	axon, dendrite, soma
15	Opthalmoception	bipolar, ganglion, light
16	Paracrine signaling	cell, paracrine, target
17	Passive transport	concentration, diffuse, equilibrium
18	Photosynthesis	chloroplast, light, thylakoid
19	Phylogenetic representation	ancestor, branch, node
20	Protein synthesis	RNA, transcription, translation
21	Reflex arc	interneuron, motor, sensory
22	Sexual reproduction	gametes, meiosis, organism
23	Upwelling	coastline, current, wind
24	Viral reproduction	generation, host, virus
25	Allopatric speciation	evolution, isolation, species
26	Cellular respiration	ATP, phosphate, sugar
27	Condensation reaction	condensation, covalent, hydroxyl
28	Crossing over	chiasma, chromatid, genetic
29	Dehydration reaction	bond, dehydration, monomer
30	Gene flow	allele, flow, population
31	Hydrolysis	hydrolysis, polymer, water
32	Lipid digestion	bile, fat, lipase
33	Membrane potential	electrochemical, ion, membrane
34	Metabolism	bacteria, carbohydrate, pyruvate
35	Mitosis	centromere, chromosome, mitosis
36	Plasmogamy	fertilization, fungi, plasmogamy

Note. For Experiments 1, 3, and 4, concept numbers 25–36 were also used. As described in the text, some concept names and terms were modified (i.e., replaced with synonyms) where necessary for certain question types. No., number.

APPENDIX B

EXAMPLES OF TRAINING PHASE QUESTIONS, FEEDBACK, AND CRITERIAL TEST QUESTIONS

Training phase				
Exp.	Question subtype	Example	Feedback	Criterial test questions
Term retrieval				
1	Fill-in-the-blank	The hypothalamus releases hormones which then affect the _____ and the signals it sends.	pituitary	The WHAT releases hormones which then affect the pituitary and the signals it sends?
2	Fill-in-the-blank	The hypothalamus releases hormones which then affect the _____ and the signals it sends.	Correct answer: pituitary The hypothalamus releases hormones; those hormones affect the pituitary and the signals that it sends. Process: hypothalamus–hormones–pituitary–signals	The WHAT releases hormones which then affect the pituitary and the signals it sends? The hypothalamus releases hormones which then affect the WHAT and the signals it sends? The hypothalamus releases hormones which then affect the pituitary and the WHAT it sends?
Relational questions				
3, 4	Process step	Hormonal signaling involves two main steps. What occurs in the first step? Answer completely with all details.	First, the hypothalamus releases hormones. Second, those hormones affect the pituitary and the signals it sends.	First, the WHAT releases hormones? Second, those hormones affect the pituitary and the signals it sends.
3	Order	In hormonal signaling, what is involved first? The hypothalamus or the pituitary?	First, the hypothalamus releases hormones. Second, those hormones affect the pituitary and the signals it sends.	First, the hypothalamus releases hormones. Second, those hormones affect the WHAT and the signals it sends?
4	Inference	What serves as the ultimate "control center" that is in charge of the entire process?	First, the hypothalamus releases hormones. Second, those hormones affect the pituitary and the signals it sends.	First, the hypothalamus releases hormones. Second, those hormones affect the pituitary and the WHAT it sends?
Retrieval-verification-scoring				
5, 6	Difficult fill-in-the-blank	The process of hormonal regulation involves: The hypothalamus_____.	Did your answer include these exact words (Yes or No)? Hypothalamus? Pituitary? Signals? How many out of those three did you include? Concept: the hypothalamus releases hormones which then affect the pituitary and the signals it sends.	The WHAT releases hormones which then affect the pituitary and the signals it sends? The hypothalamus releases hormones which then affect the WHAT and the signals it sends? The hypothalamus releases hormones which then affect the pituitary and the WHAT it sends?

Note. Exp., experiment. Criterial test questions assessed previously retrieved terms (*tested-same* condition), previously unretrieved terms (*tested-different* condition), or *restudied* terms; there were three criterial test questions per concept in each experiment, each assessing a different term (due to space limitations, only some criterial test question examples are shown). The choice of cues and to-be-retrieved term(s) were counterbalanced over subjects in each experiment.

APPENDIX C

EXAMPLE CONCEPT PARAGRAPH USED DURING INITIAL STUDY IN EXPERIMENT 2

Protein synthesis is the process by which a cell makes protein. It involves DNA, a molecule that contains genetic information, and RNA, a modified version of DNA. The process is as follows: First, DNA is copied to RNA via transcription. Second, RNA is coded into protein via translation. The process therefore involves two steps: transcription and then translation. The order of the components is DNA to RNA to protein.