Journal of Applied Research in Memory and Cognition 3 (2014) 171–176

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Does testing with feedback help grade-school children learn key concepts in science?

Amanda Lipko-Speed[a,*], John Dunlosky[b,**], Katherine A. Rawson[b]

[a] The College at Brockport, State University of New York, United States
[b] Kent State University, United States

ARTICLE INFO

Article history:
Received 27 November 2013
Received in revised form 2 April 2014
Accepted 6 April 2014
Available online 18 April 2014

Keywords:
Testing effect
Retrieval practice
Science concepts
Grade-school children

ABSTRACT

Testing effects have been well-established across a variety of studies involving school-age children. Specifically, children's test performance improves when they are given the opportunity to practice retrieval prior to the final test as compared to when practice involves only study. The current investigation focused on the influence of testing with feedback on fifth graders' learning of science concepts. Across two experiments, 65 students studied twenty key concept definitions. After study, the definitions were randomly assigned to one of four conditions: control, test only, study only, and test plus feedback. Concepts assigned to the latter three conditions were presented again (either for test only, study only, or test plus feedback) during two separate sessions, and a final test was administered during a separate session. At the final test, students were asked to provide the definition for each of the twenty key concepts. Final test performance was best for definitions that had been tested with feedback, followed by recall for those that had only been tested or only restudied, with lowest performance for those that did not receive additional practice. Despite multiple sessions spaced across several days, however, performance for all conditions was low, which may have resulted from the ineffective use of feedback. In summary, testing followed by feedback can boost younger students learning of science concepts, but mastery will require the use of other strategies as well.

© 2014 Society for Applied Research in Memory and Cognition. Published by Elsevier Inc. All rights reserved.

1. Introduction

Tests over to-be-learned materials do more than assess what students have learned, because testing itself requires retrieval practice that has been shown to improve learning, comprehension, and retention (for reviews, see Carpenter, 2012; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Karpicke, 2012; Roediger & Butler, 2011). Although the promise of practice tests for improving student learning has been demonstrated across numerous domains over the past 100 years, the degree to which testing can help younger students learn more complex science concepts has not yet been investigated. In the present research, we first consider the available evidence about the benefits of testing for young children, and we then report two experiments investigating the degree to which testing (with feedback) can help grade-school children learn science concepts.

Whereas the vast majority of testing effect research has involved adult learners, the number of studies involving upper-elementary school students is relatively small but growing (see Table 1 for published studies involving fifth and sixth graders). In this research, students usually began by studying target materials. After initial study, students typically either took tests followed by feedback (Test + FB in the far-right column of Table 1) or restudied the material (Study Only). For instance, Rohrer, Taylor, and Sholar (2010) had 4th and 5th graders learn the names of regions on fictional maps. Each student learned two maps, one via test-plus-feedback practice and the other via study only. In the test-plus-feedback condition, students were tested on their memory of region names and were then immediately shown the correct answer to restudy. In the study-only condition, students were presented with the region names for restudy. One day later, students took a final test in which they had to label the regions on a blank map. Final test performance was better for region names that had been tested with feedback than for region names that had only been restudied. Across the published studies listed in Table 1 (with one exception, Metcalfe

* Corresponding author at: The College at Brockport, Department of Psychology, 350 New Campus Drive, Brockport, NY 14420, United States. Tel.: +1 585 395 5082.
** Corresponding author at: Kent State University, Department of Psychology, Kent, OH 44241, United States.
E-mail address: alipko@brockport.edu (A. Lipko-Speed).

http://dx.doi.org/10.1016/j.jarmac.2014.04.002
2211-3681/© 2014 Society for Applied Research in Memory and Cognition. Published by Elsevier Inc. All rights reserved.

Table 1
Summary of current testing effect literature with fifth and sixth grade students.

Authors	Age	Materials	Conditions (after initial study)
Spitzer (1939)	6th graders	577-word text	Test only > no study
Metcalfe and Kornell (2007)	6th graders	Definition-Word Pairs	Test + FB = study only > test only
Metcalfe, Kornell, and Son (2007)	6th–7th graders	Definition-Word Pairs, Spanish Vocabulary	Test + FB > study > no study
Metcalfe, Kornell, and Finn (2009)	6th graders	Definition-Word Pairs	Test + FB > study only
Roediger, Agarwal, McDaniel, and McDermott (2011)	6th graders	Social study facts	Test + FB > study only
Rohrer, D., Taylor, K., & Sholar, B. (2010)	4th–5th graders	Map Location-Names	Test + FB > study only
Bourwmeester and Verkoeijen (2011)	Ages 8–12	DRM Lists	Test only > study only[a]

Note. FB, feedback
[a] Not all subgroups of participants showed this effect.

& Kornell, 2007), criterion test performance was greater when the students practiced retrieval followed by feedback than when they only restudied. Moreover, even younger children demonstrate testing effects (e.g., Fritz, Morris, Nolan, & Singleton, 2007; Lipowski, Pyc, Dunlosky, & Rawson, in press; Marsh, Fazio, & Goswick, 2012) as well as older middle-school students (e.g., Carpenter, Pashler, & Cepeda, 2009; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Nungester & Duchastel, 1982). Thus, the fifth-grade students who participated in the present experiments were expected to demonstrate a testing effect.

In the present experiments, we contribute to this small but growing literature in two key ways: (a) the to-be-recalled material—full definitions of key science concepts—are more complex than used in prior research with this age group, and (b) we included multiple conditions to assess the degree to which the effects of testing are direct or indirect (Roediger & Karpicke, 2006). These conditions were defined by whether the initial study phase was followed by practice tests and/or study: (a) testing only, (b) testing with feedback,[1] (c) study only, or (d) neither (control). The direct effects of testing occur when testing itself boosts performance, which largely arises from correctly recalling targets during practice (but see Kornell, Hays, & Bjork, 2009; Vaughn & Rawson, 2012). These direct effects are evident when criterion performance is greater after testing only than in the control condition.

Indirect effects of testing can occur when learners use the outcome from testing to enhance the effectiveness of subsequent study when feedback (i.e., the correct definition) is presented. In the present context in which students can study feedback as long as they want, these indirect effects could arise from effective self-regulated learning. According to general theories of self-regulated study (for a review, see Dunlosky & Ariel, 2011), accurate monitoring of one's learning can allow students to effectively allocate their study time. Namely, during test trials, students may evaluate how well they can recall each definition, which may help them to use the self-paced feedback trials to focus on materials that are not yet well learned. In this manner, testing could potentiate learning during restudy. These indirect effects of testing would result in superadditive effects exceeding the combined direct effects of testing alone and studying alone during practice. Considering that prior work with these age groups has typically compared test plus feedback to study alone (Table 1), whether the benefits of testing were due to direct or indirect effects remains an open question, because both may have contributed to the observed effects (although evidence from Bourwmeester & Verkoeijen, 2011, indicates that testing can directly influence performance for learning word lists). Thus, we included all four conditions in the present experiments.

Concerning the materials used in the current research, we had students learn key concept definitions. The materials were provided by fifth-grade teachers and represent science concepts that the students would be expected to learn in units on light and sound (Experiment 1) and geography (Experiment 2). For instance, one concept for light and sound is "What is amplitude? Height of a wave from its resting position to its highest or lowest point" (for other examples, see Appendix). Practice tests in the present experiments involved presenting a concept term (e.g., "amplitude") as the cue and having students attempt to retrieve the correct definition from memory. By contrast, other research using definitions to explore testing effects in children (see Table 1) has prompted them to retrieve key terms when presented with the definitions as a cue. For example, Metcalfe and colleagues (Table 1) presented children with definitions of vocabulary terms (e.g., "words or letters written, printed, or engraved on a surface") and were asked to type the corresponding target word (e.g., "inscription"). Retrieving a single key term may lead to fewer commission errors than attempting to retrieve a target definition, and the latter may undermine the benefits of testing. For instance, when presented with a definition to a newly learned concept, the key term often will either be correctly recalled or nothing will be recalled (i.e., an omission), and if a commission error does occur, then the feedback (i.e., presenting the correct key term) will allow the students to easily identify the error. However, when the task is to recall the correct definition (as in the present experiments), students often make commission errors, and they also often have difficulties evaluating the quality of their recall even when they receive the correct definition for feedback (Lipko et al., 2009). Thus, the feedback trials may not be as useful for helping them to correct errors, and if so, one might expect that the testing effect will arise more from its direct effects on memory than from its indirect effects.

To evaluate such possibilities, we conducted two experiments that used identical procedures but different stimuli, to evaluate the generalizability of the effects across different materials. In Experiment 1, we used definitions about light and sound, and in Experiment 2, we used geography definitions. In both experiments, children began by studying a set of twenty definitions. These definitions were then randomly assigned to one of four conditions: test only, study only, test plus feedback, or control. On the final day, children were tested on all twenty definitions. Given the similarity of the two experiments, we present them together for brevity.

2. Experiments 1 and 2

2.1. Method

2.1.1. Participants

Thirty fifth graders participated in Experiment 1 and 35 fifth graders participated in Experiment 2. Due to class absences, 8 participants did not complete all sessions (3 and 5 in Experiments 1 and 2, respectively), and their data were excluded from analyses. Thus, 27 and 30 participants contributed to analyses in the two experiments. Students were recruited from northeast Ohio middle schools by the Research Center for Educational Technology at Kent

[1] Functionally, this condition involved testing followed by restudy, which we refer to here as "feedback" in keeping with terminology commonly used in the testing effect literature.

Table 2
Mean proportion of correct recall performance on the final criterion test.

	Condition			
	Test + feedback	Test only	Study only	Control
Experiment 1 (light and sound)	.33 (.03)	.16 (.02)	.26 (.02)	.10 (.03)
Experiment 2 (geography)	.27 (.03)	.16 (.01)	.16 (.02)	.10 (.02)

Note. Values in parentheses are within-participant standard errors of each mean (see Loftus & Masson, 1994).

State University. At the time of data collection, their classes were being taught at the AT&T Classroom at Kent State University, where they were receiving daily instruction for several weeks.

2.1.2. Materials and design

Two sets of key concept-definition pairs, light and sound (Experiment 1) and geography (Experiment 2), were used as stimulus materials (see Appendix for examples). These definitions represented the kinds of material that the students were expected to learn in their classes. For each set, 20 definitions were developed with the assistance of fifth grade teachers.

The design was a 2 (Test) × 2 (Study) design with the two factors crossed within-participant. For each participant, five concepts were randomly assigned to each of the four conditions. The order of concepts presented during each session (regardless of condition) was randomized for each participant.

2.1.3. Procedure

Students participated in four sessions, with Sessions 1 and 2 separated by two days (e.g., Monday and Wednesday) and Sessions 3 and 4 occurring the next week and also separated by two days (e.g., Monday and Wednesday of the following week). During each session, students worked individually at computers. They were instructed to take their time and read all directions carefully. Each session began with task instructions that were presented on the computer screen.

The first session began with presentation of the 20 key term definitions one at a time for self-paced study. Students clicked a button when they were finished studying each definition. After all 20 definitions had been studied, the five items slated for the control condition were not presented again. The remaining 15 items were presented again in one of three possible conditions: study only, test only, or test plus feedback. Each of the five definitions in the study-only condition was presented again for one self-paced study trial. For definitions in the test-only condition, students were presented with the concept term (e.g., "amplitude") and attempted to type the correct definition into a response field. In the test-plus-feedback condition, students were asked to type the definition (as above) and were then immediately presented with the correct definition for self-paced study. For recall attempts, participants did not have to recall the definitions verbatim and were simply instructed to recall as much of each definition as possible.

The second and third sessions were identical to the first session except that (a) the initial study phase did not occur and (b) the 15 practiced concepts were presented for two blocks of trials. Each concept was presented in the same condition (e.g., study only) as during the first session. During the final session, students were tested on all twenty concepts. The concept terms were presented on a computer screen one at a time, and students were instructed to type the definition into a response field. Participants did not have to recall the definitions verbatim.

2.1.4. Scoring

Recall responses were scored by determining the proportion of ideas correctly recalled from each concept. Recall responses did not have to be verbatim to receive credit; instead, an idea was scored as present if the participant's response captured the correct meaning of the idea. Trained assistants had high agreement when scoring a subset of the protocols (87% agreement, with disagreements typically differing by only a single idea unit), so a single assistant scored the remainder of the protocols.

2.2. Results

2.2.1. Test performance

The critical outcomes pertaining to final cued-recall performance are presented in Table 2, which includes mean values for each condition for both experiments. For Experiment 1, a 2 (Test) × 2 (Study) repeated measures analysis of variance (ANOVA) revealed a main effect for study condition, $F(1,24) = 20.37$, $MSE = .032$, $\eta_p^2 = .46$, and for test condition, $F(1,24) = 5.89$, $MSE = .019$, $\eta_p^2 = .20$; the interaction was not significant, $F(1,24) = .19$. For Experiment 2, main effects also occurred for study condition, $F(1,28) = 7.42$, $MSE = .024$, $\eta_p^2 = .21$, and for test condition, $F(1,28) = 11.92$, $MSE = .017$, $\eta_p^2 = .30$; the interaction was not significant, $F(1,28) = 1.74$, $p = .20$. These results highlight the direct effect of testing (and study) on final performance, and the lack of interaction (i.e., no superadditive effects) suggests that any indirect effects were minimal.

Two conditions (test plus feedback and test only) included practice tests during the first three sessions, and performance on these practice tests is presented in Table 3. In both experiments, performance in the test-only condition did not increase from the first to second trial within a session, all $ts < 1.0$, $ps > .78$. For the test-plus-feedback conditions of Experiment 1, performance increased

Table 3
Mean correct recall on the practice tests.

Trial	Session 1[a]	Session 2		Session 3	
	1	1	2	1	2
Experiment 1					
Test + feedback	.23 (.02)	.16 (.02)	.27 (.02)	.25 (.01)	.32 (.02)
Test only	.18 (.02)	.16 (.01)	.15 (.03)	.16 (.01)	.17 (.02)
Experiment 2					
Test + feedback	.21 (.01)	.22 (.01)	.23 (.02)	.27 (.01)	.23 (.03)
Test only	.16 (.01)	.16 (.01)	.12 (.01)	.16 (.01)	.13 (.01)

Note. Values are proportions, and values in parentheses are within-participant standard errors of each mean (see Loftus & Masson, 1994).

[a] Only a single practice test was administered during Session 1.

Table 4
Mean study times for the test-plus-feedback and study only conditions.

Trial	Session 1		Session 2		Session 3	
	1	2	1	2	1	2
Experiment 1						
Test + feedback	28.4 (3.8)	7.9 (.7)	7.3 (.6)	3.8 (.4)	6.6 (.8)	3.8 (.5)
Study only	27.9 (3.2)	6.9 (.6)	9.6 (.8)	3.6 (.5)	6.5 (.6)	3.5 (.6)
Test only	28.3 (5.4)	–[a]				
Control	22.8 (2.2)	–[a]				
Experiment 2						
Test + feedback	25.0 (1.2)	7.4 (.9)	8.2 (.60)	5.1 (.7)	10.4 (1.7)	3.8 (.7)
Study only	25.9 (1.8)	7.7 (1.1)	14.6 (2.0)	4.4 (.7)	12.0 (1.4)	2.7 (.9)
Test only	26.5 (2.0)	–[a]				
Control	22.0 (1.6)	–[a]				

Note. Values are mean study time per concept (in seconds), and values in parentheses are within-participant standard errors of each mean (see Loftus & Masson, 1994) computed separately for the initial trial and for the subsequent trials.

[a] Participants did not study definitions in these conditions after the initial study trial.

from the first to second trial in Sessions 2 and 3, $ts > 3.0$, $ps = .005$. The same boost for the test-plus-feedback conditions was not evident in Experiment 2 ($ts < 1.7$, and note that in Session 3, the values trend in the opposite direction). These data suggest one reason why the testing effect was relatively modest in the present experiments; namely, students had difficulties learning the concepts even after receiving tests with feedback.[2]

2.2.2. Study time

Study times are presented in Table 4. To revisit, participants studied all concepts during the first trial of Session 1 before the concepts were randomly assigned to the four conditions. Thus, differences in study times among the conditions on the first trial of Session 1 reflect noise. Also, concepts that were assigned to the test only and control conditions were not restudied at any point after the initial study trial in Session 1.

What is evident from inspecting Table 4 is that for both experiments, participants spent more time studying concepts on the initial trial and spent relatively minimal time on subsequent trials. This trend is evident during all three sessions and was supported by a 2 (Practice condition: test plus feedback vs. study only) × 3 (Session: 1, 2 or 3) × 2 (Trial: 1 or 2) ANOVA, which revealed a significant main effect for trial for Experiment 1, $F(1,25) = 22.78$, $MSE = 314.0$, $p < .001$, $\eta_p^2 = .48$, and for Experiment 2, $F(1,25) = 30.7$, $MSE = 304.3$, $p < .001$, $\eta_p^2 = .55$. This significant trend also interacted with Session in both experiments, indicating that the difference in study times for the first and second trial diminished across sessions. This Session × Trial interaction is largely due to the fact that participants spent a great deal of time studying each concept during the initial study trial, but they used much less time restudying on the first trial of the subsequent sessions.

Concerning the effect of practice condition, the main effect was not significant either in Experiment 1, $F(1,25) = 0$, $MSE = 56.5$, or in Experiment 2, $F(1,25) = 3.06$, $MSE = 77.4$, $p = .09$, $\eta_p^2 = .11$. Thus, differences in final criterion performance (Table 2) cannot be attributed to students using more time for studying under the test-plus-feedback than study-only condition. A significant Practice condition × Trial interaction occurred in Experiment 2, $F(1,25) = 6.0$, $MSE = 59.9$, $p = .02$, $\eta_p^2 = .19$. Note, however, this interaction (which indicates that students studied longer on Trial 1 vs. Trial 2 in the study only condition than in the test-plus-feedback condition) does not qualify the conclusion above, and it was not evident in Experiment 1.

Finally, although students used about the same amount of time restudying under the test-plus-feedback and study alone conditions, they may have used the outcome from each test to regulate their restudy, such as by spending more time restudying after they recalled less (vs. more) of a definition during the practice test. To evaluate this possibility, we examined study time (for each session) as a function of recall outcome. Given that only five definitions per participant were assigned to the test-plus-feedback condition, relatively few observations would be available to compute study time conditionalized on recall performance for each trial. Accordingly, recall outcome was scored as either entirely incorrect (i.e., omission or commission error) vs. not entirely incorrect (i.e., partially or entirely correct response). Also, during Sessions 2 and 3, concepts in this condition received two test trials, each of which was followed by feedback; for these sessions, study times for a given trial were based on the immediately preceding recall attempt within that trial.

Means across individual values are presented in Table 5. In Experiment 1, the expected trend is evident across some trials and appears most evident in the first trial of each session. Even so, potentially due to noise, this relationship between recall and outcome was not statistically significant in a 2 (Session: 2 and 3) × 2 (Trial within Session: 1 or 2) × (Recall outcome: incorrect vs. not fully incorrect) ANOVA, all $Fs < 1$, $ps > .60$. For Experiment 2, the same trends were not as apparent, and the main effect of recall outcome (and all interactions involving it) were not significant, $Fs < 1.0$, $ps > .30$. Although these data suggest that practice recall did not help students to focus on the least well-learned material, we should emphasize that the conditional values were based on few observations and the levels of learning overall were low, suggesting that students needed to restudy most of the concepts regardless of the outcome of the practice recall attempts. The use of recall outcome to guide restudy in this age group will require further investigation before one can draw firm conclusions.

3. General discussion

3.1. The benefits of testing for learning key-concept definitions

Most critically, final test performance for difficult-to-learn science concepts was highest when students used testing plus feedback than when they used testing alone or studying alone. The overall pattern of data (Table 2) suggests that the benefit of testing in this context is at least partly due to the independent effects of testing and feedback—both appear to benefit performance. Nevertheless, it is also evident that testing had a small effect in these experiments, which we consider further under "Practical Applications" below. The lack of an interaction also suggests that the

[2] Across all trials and both experiments, the rate of omission errors on test trials was low (median = .2, mode = 0) and hence too few observations were available to support conditional analysis.

Table 5
Mean study time on feedback trials as a function of recall outcome on preceding test trial for the test-plus-feedback condition.

Trial	Session 1[a]	Session 2		Session 3	
	1	1	2	1	2
Experiment 1					
Incorrect[b]	7.7 (.08)	7.6 (.08)	3.7 (.04)	7.1 (1.3)	4.3 (4.4)
Partially correct[c]	7.8 (1.1)	5.0 (.04)	4.2 (.05)	5.8 (1.2)	3.4 (.06)
Experiment 2					
Incorrect	7.7 (1.1)	8.5 (1.2)	4.5 (.6)	7.9 (.8)	3.8 (.5)
Partially correct	6.9 (1.1)	7.3 (.7)	6.1 (1.1)	9.6 (2.1)	4.1 (1.3)

Note. Study time is seconds per concept, and values in parentheses are within-participant standard errors of each mean (see Loftus & Masson, 1994).
[a] Session 1 included only one test-plus-feedback trial.
[b] Few responses were omissions; thus, to increase the number of observations per cell, we collapsed across all incorrect responses.
[c] Few responses were entirely correct; thus, to increase the number of observations per cell, we collapsed across all responses that were at least partially correct.

combination of testing and feedback did not have a superadditive effect that would have been expected if testing was (partly) boosting performance by potentiating restudy.

Why didn't testing potentiate restudy? One way that testing could have potentiated restudy was by allowing students to focus on restudying those concepts that they could not correctly retrieve. To obtain these benefits, students would need to accurately judge the quality of their recall responses and then differentially allocate their study time to those definitions that they could not correctly recall. We did not measure judgment accuracy in the present experiments, but grade-school students and even college students show poor judgment accuracy (and overconfidence) when they are asked to evaluate response accuracy when recalling key-term definitions (Lipko et al., 2009). Thus, poor judgment accuracy may partly explain why testing did not potentiate restudy.

Even if their judgments were highly accurate, however, students may not properly allocate their study time during feedback. Relevant evidence comes from Metcalfe and Finn (2013), who had third and fifth graders study items, make delayed judgments of learning, and then choose which items to restudy. In all of their experiments, students' judgments were highly accurate. Their allocation of study time, however, was not optimal: third graders' choices of which items to restudy were random, and although fifth graders tended to choose items that were judged as least well learned, their control did not appear to improve their performance. Consistent with their results, the fifth grade students in the present experiments did not consistently allocate more time to feedback after test responses that were entirely incorrect vs. partially correct. Moreover, students did not use much time at all restudying the definitions during feedback, which suggests that they may have merely reread each definition and missed the opportunity to process them more deeply.

In summary, all this evidence suggests that students' poor regulation—due to inaccurate judgments and ineffective allocation of time—may partly explain why testing did not potentiate restudy. Poor regulation also may have contributed to their overall low levels of performance. Most important, despite the potential limits of poorly regulated learning, results from both studies establish that testing plus feedback can benefit grade-school children when they are learning science concepts.

3.2. Practical applications

The low levels of performance (see Table 2) indicate that even with the aid of testing with feedback, upper-elementary school students were not retaining the majority of the concepts, at least as measured by the challenging task of definition recall. Although recall performance was poor, perhaps the students had mastered some of the concepts well enough that they would have demonstrated better retention and understanding on a different kind of test, such as recognition tests or measures that had students apply their knowledge but did not require them to verbalize it. We leave evaluation of this possibility for future research and turn to reasons for the students' low levels of criterion recall. One reason pertains to their low rate of success during the practice sessions (Table 3), which may have been low for any number of reasons (e.g., difficult materials; inappropriate motivation levels). More important, test practice is most beneficial for retention when it results in correctly recalling the target information, but in the present case, the students did not practice enough to obtain high levels of correct retrieval during the first three sessions.

Accordingly, to obtain better long-term success, students should continue practicing the concepts until they can correctly recall each one. Such *successive relearning* in which students continue testing with feedback until they succeed across multiple sessions has been shown to yield long-term retention (across many months and even years) of foreign-language vocabulary and science concepts (Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Rawson & Dunlosky, 2011, 2012), at least for adults. We suspect successive relearning will be just as powerful when used by younger students, but currently few studies have been conducted to demonstrate the power of successive relearning. Thus, one important avenue for future research will be to systematically explore students' use of successive relearning and the degree to which it can support efficient and durable learning that improves in-class performance.

One potential problem of using testing with feedback to learn science concepts is highlighted by some of the present outcomes. In particular, after two trials of practice (Table 3), students obtained only about 27–30% correct recall, and this lack of learning in part can be traced to the relatively brief amount of time (Table 4) that the students used during the feedback trials. In fact, during the second trial of each session, it appears that students were doing little more than reading the correct definition. Unfortunately, rereading is not a very effective learning strategy (for a review, see Dunlosky et al., 2013), and in the present case, students may have difficulties mastering the concepts in a single session if they rely solely on rereading while restudying. A potential solution to this problem is to supplement testing and feedback with other material-appropriate strategies that would enhance the efficacy of the feedback trial. For instance, self-explanation and elaborative interrogation are strategies in which students describe how a new concept is related to other information that they know or generate explanations for why a particular concept is true. As compared to rereading, these strategies would encourage students to more actively process the target information and elaborate on it and often boost students' learning and comprehension of concepts (for reviews, see Dunlosky et al., 2013; Roediger & Pyc, 2012). Note, however, that such elaborative processing of feedback did not benefit performance in one investigation on the testing effect (McDaniel & Fisher, 1991), but the materials were trivial facts (e.g., Who invented the light bulb?) that may have limited the usefulness of elaboration. Given the benefits of elaborative techniques

for improving retention and comprehension in general and the success of testing plus feedback, evaluating the degree to which these (and other) strategies can further improve student success is a promising area for future cognitive-education research.

Acknowledgments

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H050038 and Grant R305A080316 to KSU. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. It was also partially supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior Collaborative Award.

Appendix.

Sample definitions

Experiment 1

What makes sound echo and/or not echo? Sound can be absorbed by many materials that make sound not echo; sound echoes when there is nothing around to absorb it.

What is a convex lens? It is thin at the edges and thick near the center; it brings parallel light rays together at a focal point.

What is refraction of light? Bending of light as it passes from one material into another.

What is sound? Form of energy that you can hear that travels through matter as waves.

What is amplitude? Height of a wave from its resting position to its highest or lowest point.

Experiment 2

What is absolute location? The exact location of a place on a map by noting its latitude and longitude.

What is a map legend? The part of a map that explains the symbols used on a map.

What does it mean to be interdependent? Depending on one another for resources and products.

What is a megalopolis? A huge urban region formed when two or more metropolitan areas grow together.

What is the climate? The typical temperature and precipitation of a particular place over time.

References

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*, 316–321. http://dx.doi.org/10.1111/j.1467-9280.1993.tb00571.x

Bourwmeester, S., & Verkoeijen, P. P. J. L. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language*, *65*, 32–41. http://dx.doi.org/10.1016/j.jml.2011.02.005

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*, 760–771. http://dx.doi.org/10.1002/acp.1507

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

Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In B. Ross (Ed.). *Psychology of Learning and Motivation*, *54*, 103–140.

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*, 4–58. http://dx.doi.org/10.1177/1529100612453266

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *The Quarterly Journal of Experimental Psychology*, *60*, 991–1004. http://dx.doi.org/10.1080/17470210600823595

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*, 157–163. http://dx.doi.org/10.1177/0963721412443552

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

Lipko, A. R., Dunlosky, J., Hartwig, M., Rawson, K. A., Swan, K., & Cook, D. L. (2009). Using standards to improve middle-school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied*, *15*, 307–318. http://dx.doi.org/10.1037/a0017599

Loftus, G. R., & Masson, M. E. J. (1994). *Psychonomic Bulletin & Review*, *1*(4), 476–490.

Lipowski, S., Pyc, M., Dunlosky, J., & Rawson, K. A. (2014). Establishing and explaining the testing effect in free recall for young children. *Developmental Psychology*, *50*(4), 994–1000.

Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. *Memory*, *20*, 899–906. http://dx.doi.org/10.1080/09658211.2012.708757

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201.

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L., III. (2013). Quizzing in middle school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*, 30–372. http://dx.doi.org/10.1002/acp.2914

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition & Learning*, *8*, 19–46.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review*, *14*, 225–229. http://dx.doi.org/10.3758/BF03194056

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive science based programme to enhance study efficacy in a high and low risk setting. *European Journal of Cognitive Psychology*, *19*, 743–768. http://dx.doi.org/10.1080/09541440701326063

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, *37*, 1077–1087. http://dx.doi.org/10.3758/MC.37.8.1077

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, *74*, 18–22. http://dx.doi.org/10.1037/0022-0663.74.1.18

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*, 283–302. http://dx.doi.org/10.1037/a0023956

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review*, *24*, 419–435. http://dx.doi.org/10.1007/s10648-012-9203-1

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

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

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

Roediger, H. L., III, & Butler, A. C. (2011). The critical old or retrieval practice in long-term retention. *Trends in Cognitive Science*, *15*, 20–27. http://dx.doi.org/10.1016/j.tics.2010.09.003

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

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656. http://dx.doi.org/10.1037/h0063404

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review*, *19*(5), 899–905. http://dx.doi.org/10.3758/s13423-012-0276-0