Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

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

The testing effect with authentic educational materials: A cautionary note

Cynthia L. Wooldridge[a,*], Julie M. Bugg[b], Mark A. McDaniel[b], Yiyi Liu[b]

[a] Lindenwood University, United States
[b] Washington University in Saint Louis, United States

ARTICLE INFO

Article history:
Received 4 December 2013
Received in revised form 30 June 2014
Accepted 1 July 2014
Available online 10 July 2014

Keywords:
Testing
Testing effect
Formative assessment

ABSTRACT

Despite considerable evidence that testing benefits subsequent retrieval of information, it remains uncertain whether this effect extends to topically related information with authentic classroom materials. In the current study we first profile the way in which quizzing is used in the classroom through a survey of introductory psychology instructors. The survey results indicate that, instructors frequently use related but different questions on quizzes and tests unlike many laboratory experiments that use identical questions. In two subsequent experiments, participants studied information from a college biology textbook, were quizzed twice, and given a final test. The items on the final test were either identical to or were related but different than the quiz items. Experiment 1 showed that testing produced the typical robust testing effect for repeated items, but there was no significant effect of testing for topically related items. In Experiment 2, participants could use their quizzes to guide restudy, and there was still no positive effect of testing for topically related information.

Recently, there has been a surge of research investigating testing as a learning tool for promoting retention. Ample evidence suggests that repeated testing improves memory, a phenomenon known as the testing effect (see Roediger & Karpicke, 2006b, for review). For example, Roediger and Karpicke (2006a) asked subjects to read two passages and then restudy one passage and recall the other. Subjects later recalled 14% more of the tested than the restudied passage, showing a testing benefit beyond restudying. The majority of studies demonstrating the testing effect have been controlled experiments, requiring recall of materials such as word lists (e.g., Tulving, 1967) or paired associates (e.g., McDaniel & Masson, 1985). However, some studies have employed educational materials and shown the robust nature of the testing effect (e.g., McDaniel & Fisher, 1991; Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008).

Based on these results, researchers have encouraged educators to consider using quizzing in classrooms (Pashler et al., 2007) and some textbooks are now accompanied by quizzing ancillaries so that educators can use testing in a more general way. For example, Worth Publishing devised an online quizzing system that accompanies an introductory biology textbook (Phelan, 2009). The quizzes are designed with the assumption that answering factual and application questions will promote a more integrated mental model that incorporates the target knowledge (e.g., see Karpicke, 2012). One anticipated outcome from these quizzes is that students will perform better on related questions on class exams. As a concrete example, answering the quiz item, “Convergent evolution can occur only when two species: ________.” (Answer: evolve under similar selective forces) should allow students to better answer the exam question, “Penguins and dolphins have flippers but do not have a common ancestor. Their flippers are: ________.” (Answer: homologous structures.) The questions ask about information that is topically related, but involve *different* concepts from the same section of the chapter.

These materials reflect a departure from classroom experiments in which the exam questions are identical or directed at the same concepts (or facts) as the quiz questions (e.g., see McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; McDermott, Agarwal, D’Antonio, Roediger, & McDaniel, 2014). For example, Butler (2010) designed materials such that a single concept was targeted by several different questions. Thus, answering one question prompted recall of the underlying concept which then led to improved performance on the subsequent related exam item.

[*] Corresponding author at: Department of Psychology, Lindenwood University, 209 South Kingshighway, Saint Charles, MO 63301, United States.
Tel.: +1 636 949 4732.
E-mail addresses: cynthia.wooldridge@gmail.com, cynthia.fadler@gmail.com (C.L. Wooldridge).

http://dx.doi.org/10.1016/j.jarmac.2014.07.001
2211-3681/© 2014 Society for Applied Research in Memory and Cognition. Published by Elsevier Inc. All rights reserved.

However, as illustrated in the above example, at least some materials that are currently being employed involve quiz questions that have topically related (instead of identical) underlying target concepts. The assumption is that quizzing on topically related material (as opposed to questions targeted at the same concept) will still produce later benefits. Note that the above example is taken from a currently used quizzing system designed to boost exam performance and represents a way in which retrieval practice is actually being employed within the classroom. Therefore this type of testing is being used without requisite empirical support.

It is clear that using testing (quizzing) to promote course achievement can be implemented with a range of overlap between quiz items and exam items. In the present study, we examine two related questions. How do instructors in fact implement testing in the classroom to assist student learning, and what are the effects of testing for promoting exam performance when the relation between quiz and exam items moves from identical overlap to a more topical overlap? To foreshadow, based on the survey reported below, it appears that the use of testing in “the wild” likely includes thematic/topical overlap as described above (see also Mayer et al., 2009, for a classroom example) and is often instantiated with publishers’ quizzing and test bank materials. We first report these survey results, and then we describe two laboratory experiments that use off the shelf materials to investigate whether this use of testing confers benefits on final exam performance.

1. Survey

It is uncertain how instructors who are using quizzing to improve their students’ achievement in a course prefer to implement the technique. They may use questions that are identical on quizzes and exams (as in laboratory experiments) or they may tend more toward the use of topically related items as described above. Accordingly, our survey was focused on the degree to which college-level instructors match quiz and later exam questions.

1.1. Method

1.1.1. Participants

Two hundred and fifty-two participants were recruited through the Society for Teaching of Psychology Discussion (PsychTeacher^{TM}) list serve. Forty-six participants were excluded due to missing data and four because they had not taught an introductory psychology course (we were interested in introductory courses because publishers typically provide quizzing ancillaries for instructor use). The responses analyzed here are from the 162 participants who reported giving quizzes in their classes.

1.1.2. Materials and procedure

The survey took five to ten minutes to complete and consisted of two main parts (see Appendix 1 for complete survey). Part I included demographic items and questions related to the source of quiz and exam questions, including the proportion of created versus test bank items. Critically, participants were asked how often they use identical versus similar but not identical quiz questions on quizzes and exams.

Part II explored participants’ understanding of the testing effect. We asked if they knew of the testing effect and then gave them a concrete quizzing scenario adapted from an item used in Mayer et al. (2009; see Appendix 1) which asked them to select the types of questions which should show benefits from testing.

1.2. Results

1.2.1. Course characteristics

Many participants (65.4%) reported quizzing students “weekly” with the majority implementing in-class quizzes (62.3%) and out-of-class (e.g., web-based) quizzes (58%). Of those using out-of-class quizzes, 87.2% reported that the out-of-class quizzes were required.

1.2.2. Quiz and exam question characteristics

A majority of participants made up their own quizzes (66.0%) and used a test bank that they did not create (80.9%). On average 66.7% (*SD* = 29.9%) of quiz questions were taken directly from a test bank, while exams were constructed of an average of 51.3% (*SD* = 31.7%) test bank items.

Only 25.3% of instructors gave students identical questions on quizzes and exams. On average, 36.5% (*SD* = 32.5%) of the exam items were identical to quiz items, with the most frequently reported percentages in the range of 1–25% (58.5% of participants; see Fig. 1 for complete distribution). Most participants (74.7%) reported that they gave students similar but not identical questions on quizzes and the later exam, with an average of 42.3% (*SD* = 24.3%) of the exam items constructed in this way (see Fig. 1). For those who used these “similar-matching” items, 91.7% believed that answering quiz questions (versus no quizzing) would result in higher performance on a later exam.

Overall, 6.2% of participants reported using identical-matching only, 55.6% reported using similar-matching only, 19.1% reported using both matching methods, and 19.1% reported not using any matching methods at all.

1.2.3. Knowledge of testing effect

Most participants (76.7%) reported being familiar with the testing effect. In an applied question we presented a scenario containing a quiz question with four options and asked which option(s) would show improved performance on a subsequent exam (see Appendix 1). The most frequently chosen option was “a similar but not identical question” (81.7% of participants); followed by “an identical question on the subsequent exam” (46.5%); “covering the same general topic but different concepts or theories” (34.7%); and “a question that is unrelated to the quiz question” (5.0%).

1.3. Discussion

Overall, we found that most instructors use similar items on quizzes and exams, with almost half of items on exams representing similar but not identical questions. Especially noteworthy, 34.7% of the instructors who used quizzing believed that quiz questions covering related but different concepts would improve exam performance. This belief may be due to an assumption that quizzing helps students to develop a better structured mental model for all material, thus producing benefits for related material through knowledge integration (Mayer et al., 2009; Karpicke, 2012). Yet, this quizzing situation is not representative of that evaluated in laboratory and even classroom research on the testing effect (see McDaniel et al., 2013, for review). In most of the extant literature, quiz items are typically identical (or closely similar) from quiz to exam. Therefore, the objective of the two following experiments was to determine whether a testing effect would emerge when exam questions reflected the topical information targeted in the quiz, but the questions were not identical. In both experiments, we used authentic classroom text and quizzing materials in a controlled laboratory setting.

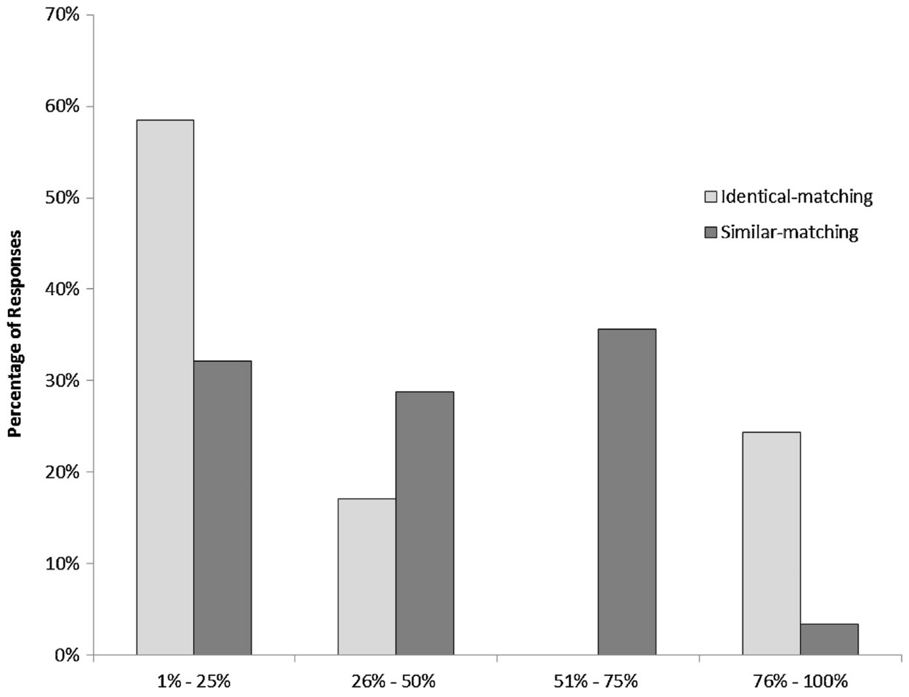

Fig. 1. Percentage of responses as a function of proportion of quiz questions given in each matching type. Identical-matching refers to questions where the same question appears on the initial and final test. Similar-matching refers to questions where a similar, but not identical, question appears on the initial and final test.

2. Experiment 1

In our survey, a large majority of instructors indicated using conceptually similar questions across their quizzes and exams. The survey did not determine the nature of the similarity between quiz and exam questions. At least three kinds of similarity can be found in the recent testing-effect literature. One is that the identical associative information is tested in one direction on the initial test (quiz) (A–?) and in the reverse direction (B–?) on the final test (for an example, see Part II of Appendix 1, answer option B for Question 2). In both laboratory experiments (Rohrer, Taylor, & Sholar, 2010, with map location-name pairings) and classroom experiments (McDaniel et al., 2013, Experiment 1, with term-definition pairings) initial testing produces enhanced performance on these similar exam items. A second is that a particular concept or construct is initially tested (quizzed) on its application in a particular context, and then the same construct is tested in a different application on the final exam. Again, laboratory (Butler, 2010; Carpenter, 2012) and classroom experiments (McDaniel et al., 2013) have shown test-enhanced performance when quiz and exam items are related in this fashion.

A third kind of similarity is that identified earlier, in which the quiz and exam items are related topically but do not focus on the same particular concept or theory. An initial quasi-experimental classroom study reported no testing effects in this situation (Mayer et al., 2009). The instructors in our survey provided mixed opinions, however, with a third of the instructors in our survey suggesting that testing effects would be likely in this situation. Experiment 1 evaluated this expectation in a controlled laboratory setting. Based on recent findings in classroom contexts (Glass, 2009; McDaniel, Wildman, & Anderson, 2012), we assumed that repeated quizzing would increase the likelihood of observing testing effects. Accordingly, participants received two quizzes. We implemented conditions that reflected the more typical paradigm in which both quiz and exam items were identical, and we also implemented conditions in which the three questions (two quiz, one exam) were all different but targeted the same general topic (this in part reflects a situation akin to quizzing ancillaries and test banks, which were the source of our materials). Finally, quizzes and the final test included both factual and applied questions, again reflecting common questions types found in the quizzing ancillaries and test banks that we adopted for our materials.

2.1. Method

2.1.1. Subjects and design

One hundred and forty undergraduates from Washington University participated for $5 per half hour of participation or partial fulfillment of a course requirement.

A 5 (Condition: Repeated Facts, Repeated Application, Related Facts, Related Application, Highlight) x 2 (Final question type: Fact, Application) mixed design was used such that initial question type was manipulated between-subjects and final question type was manipulated within-subjects. The main dependent variable was proportion correct on the final test as a function of question type. A second dependent variable targeted the accuracy of participants' predicted performance.

2.1.2. Materials

The materials consisted of a textbook chapter, "Evolution and Natural Selection", taken from Phelan (2009), *What is Life? A Guide to Biology*. The book is accompanied by a database of questions called "Prep-U" and the targeted level of Bloom's taxonomy for each question (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956; Anderson et al., 2001). Questions that represented the lower two levels of Bloom's taxonomy (remember, understand) were designated as *fact* questions while the upper two levels (evaluate, create) were designated *application* questions.

For 19 sections of the textbook chapter, 6 unique questions (3 fact, 3 application) were used for a total of 114 questions. Of these 114 questions, the experimenters created 52 questions because there were not enough questions provided in the database. Blind ratings across two experimenters indicated high inter-rater agreement (85%) as to question type and all discrepancies were resolved.

Table 1
Predicted and observed performance on the initial short answer quizzes as a function of condition in each experiment.

	Quiz 1 Predicted	Quiz 1 Observed	Quiz 2 Predicted	Quiz 2 Observed
Experiment 1				
Related fact (N = 30)	.74	.44	.50	.40
Related application (N = 30)	.66	.51	.46	.52
Repeated fact (N = 25)	.74	.54	.60	.69
Repeated application (N = 26)	.74	.43	.48	.68
Experiment 2				
Quiz-restudy-quiz-restudy (N = 24)		.41		.50

All quiz items were short-answer in order to maximize the likelihood of obtaining testing effects (e.g., Butler & Roediger, 2007; Glover, 1989; Kang, McDermott, & Roediger, 2007) and final test items were multiple choice. For the full set of questions, please refer to the supplemental electronic materials.

The initial quizzes contained 19 questions, one from each section of the chapter, and each quiz contained either all fact or all application questions (depending on condition). In the two "repeated" quizzing conditions, subjects received the same quiz questions on each of the initial quizzes, which appeared again on the final test. In the two "related" conditions, the two quizzes were composed of questions that were topically related but different than those on the final test. The final test was identical for all subjects and was composed of 19 fact and 19 application items. As an example, for the "repeated fact" condition, subjects received a fact quiz, a second fact quiz with identical items (from quiz 1), and a final test with the same fact questions that appeared on the quizzes and new related application questions. For the "related fact" condition, subjects received a fact quiz, a second fact quiz with topically related items, and a final test with new related fact items and new related application items. In the related conditions, the questions on each initial quiz and the final test were pre-determined and the order of quiz questions was identical for all subjects.

2.1.3. *Procedure*

Subjects were first given 45 min to read through the chapter without marking on it and without going back. If they finished early, they were allowed to move on. After reading, subjects in the quizzed conditions were asked to predict what percent of the upcoming questions they would correctly answer. Then received unlimited time to answer the 19 short answer items and process feedback, which consisted of a re-presentation of the question and correct answer after each item. All quiz questions were presented using E-Prime software (Schneider, Eschman, & Zuccolotto, 2007). All subjects in a given condition received the same order of quizzes, but the questions within each quiz were randomly ordered. After the first quiz, participants played the game Tetris for 10 min before taking another 19-item quiz, which followed the same procedure described above. Pilot testing indicated that the quizzes took approximately 20 min to complete. Therefore, instead of quizzing, the Highlight group was given 20 min to highlight any relevant parts of the chapter in order to prepare for the final test.

Subjects returned after 48 h and were again asked to predict what percent of the upcoming test questions they would correctly answer. Participants then took a multiple-choice test that consisted of one fact and application question for each section of the chapter (total of 38 items). The test was presented in a single random order to all participants and was self-paced, forced report, and included correct answer feedback after each question.

2.2. *Results*

All results, unless otherwise stated, were significant at the .05 level. Table 1 shows the predicted and observed performance levels for each quizzed condition. As we were primarily interested in final test performance, we will focus on these data herein. Fig. 2 shows the proportion of correct responses on the final test as a function of question type and group. In the repeated question conditions, there was a robust testing effect such that accuracy was higher on items that had been previously tested (Fact questions for the Repeated Fact group; Application questions for the Repeated Application group). However, a testing effect did not emerge in the related conditions, which showed no benefit of prior testing compared to the Highlight condition. These observations were confirmed with a 2 (Question type: Fact, Application) × 5 (Condition: Repeated Fact, Repeated Application, Related Fact, Related Application, Highlight) mixed analysis of variance (ANOVA), which revealed a significant interaction, $F(1,135) = 20.61$, $p < .001$, $MSE = .008$, $\eta_p{}^2 = .38$. Planned pair-wise comparisons showed that for fact final-test questions, the Repeated Fact quizzing group scored significantly and substantially higher than the Highlight condition (.84 vs. .65), $F(1, 135) = 58.65$, $p < .001$, $MSE = .008$, $\eta^2 = .30$, but no other quizzing group produced significant improvement (relative to Highlight). For the application final-test questions, the Repeated Application quizzing group produced a significant and substantial benefit relative to the Highlight group (.86 vs. .66), $F(1,135) = 69.16$, $p < .001$, $MSE = .008$, $\eta^2 = .34$,but no other quizzing group showed a benefit.

Table 2 shows the predicted and actual performance of each condition. Subjects in each tested condition were underconfident when predicting test performance (observed > predicted). However, the Highlight group was overconfident (predicted > observed). These observations were confirmed with a 2 (Performance: Predicted, Observed) × 5 (Condition) repeated measures ANOVA, which revealed a significant interaction, $F(1,135) = 8.67$, $p < .001$, $MSE = .015$, $\eta^2 = .06$. Planned pair-wise comparisons showed that each tested group was significantly underconfident, while the Highlight group was significantly overconfident, all Fs $(1,135) > 4$, ps $< .05$, $MSE = .015$. Because only a global prediction was made, we could not examine whether these patterns varied as a function of question type.

2.3. *Discussion*

In line with previous laboratory and classroom experiments, we found a robust testing effect when quiz questions were identical

Table 2
Predicted and observed performance on the final multiple choice test as a function of condition in each experiment.

	Predicted	Observed	F-value
Experiment 1			
Related fact (N = 30)	.54	.69	23.01**
Related application (N = 30)	.52	.62	9.22**
Repeated fact (N = 25)	.65	.76	10.23**
Repeated application (N = 26)	.61	.78	25.28**
Highlight (N = 29)	.72	.65	4.37**
Experiment 2			
Quiz-restudy-quiz-restudy (N = 24)	.61	.68	5.66**
Highlight (N = 24)	.79	.67	15.71**

** Indicates significance at the .05 level.

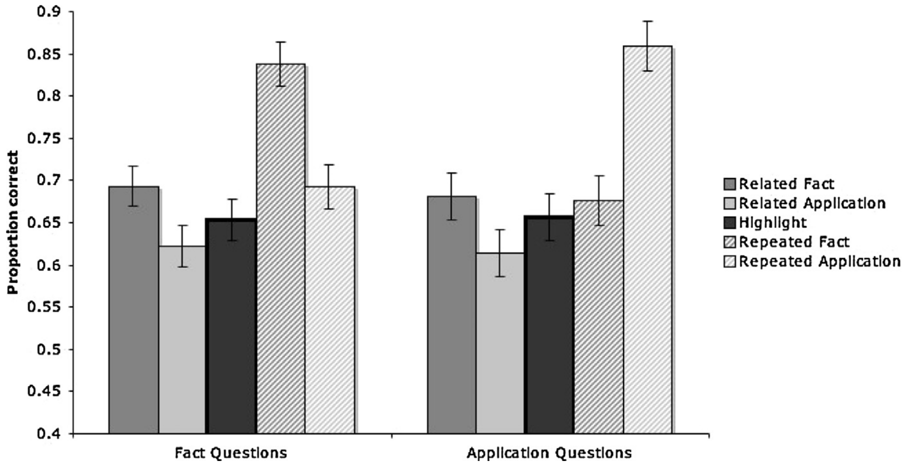

Fig. 2. Proportion correct in Experiment 1 as a function of condition and final test question type. Error bars represent the standard error.

to the test questions. By contrast, no significant benefits of prior quizzing were observed (relative to a restudy group) when the quiz items were topically related to the exam items. This finding signals a possible limit to the extent to which testing can promote transfer to related final test questions (as reported in Butler, 2010; Chan, McDermott, & Roediger, 2006; McDaniel et al., 2013). As noted earlier, in previous experiments that found transfer, the related quiz and final test questions were explicitly developed to target the same associative information (McDaniel et al., 2013; Rohrer et al., 2010), or the same target concepts but in different applications (Butler, 2010; Glass, 2009; Carpenter & Kelly, 2012; see Carpenter, 2012 for review).

Diverging from these previous studies, the current quiz and exam materials were not aligned in terms of identical associative content or in terms of testing the same constructs and concepts. Rather, the present materials were inspired by classroom practice in which instructors use the quizzing ancillaries and test banks provided by textbook publishers (as indicated by our survey results). In this case, quiz items and exam items are not necessarily so closely aligned. Typically, the quizzes and test banks sample items from similar sub-sections in the textbook but not necessarily the same information. For example, in the current experiment subjects in the related fact quiz condition were asked (on the quiz): "Estimates of evolutionary relatedness based on the 'molecular clock' are supported by what?" (Answer: the fossil record). On the exam they were asked: "The longer two species have been evolving on their own, the greater the number of ____ that accumulates between them." (Answer: genetic differences). These questions are related in terms of the general topic (evolution) and are from the same section of the text, but recalling the first answer apparently was not a strong prompt to recall information related to the latter (exam) question (as was the case in the laboratory materials of Chan et al., 2006).

The current results thus suggest that in applied settings and with "off-the-shelf" quizzing materials supplied by publishers, direct benefits of testing do not readily extend to a relatively common situation where items tested on quizzes and exams are related only at a topical level (see also, Mayer et al., 2009, for absence of testing effects with quizzing used in this fashion in a classroom).

3. Experiment 2

In a classroom context, testing (quizzing) may provide broader benefits than just direct effects from the retrieval processes stimulated by testing (Roediger, Putnam, & Smith, 2011). Specifically, the metacognitive results of Experiment 1 indicate that the quiz may demonstrate to students that they know less than they originally thought. That is, testing serves as a formative assessment that could guide students' subsequent study not only of quizzed material but related material as well; this is in fact a pervasive assumption of many researchers and educators (Angelo & Cross, 1993; Black & William, 1998; Clark, 2012; Stiggins, Arter, Chappius, & Chappius, 2006). To explore this possible benefit of topically related quiz and exam items, in Experiment 2 quizzes were immediately scored and participants were able to use the quiz feedback to guide a restudy opportunity, which occurred after each of the quizzes. The question of primary interest was whether quizzing on items that were topically related to the exam items would stimulate more effective restudy than no quizzing (highlighting during restudy).

3.1. Method

3.1.1. Subjects and design

Forty-eight undergraduates from Washington University participated for $5 per half hour of participation or partial fulfillment of a course requirement.

We used a 2 (Condition: Quiz, Highlight) × 2 (Final question type: Fact, Application) mixed design, with condition manipulated between subjects. The main dependent variable was the proportion correct on the final test as a function of question type and subjects were again asked to predict their performance.

3.1.2. Procedure

The procedure used in Experiment 2 was similar to Experiment 1 with the following exceptions. Only two conditions were used: the Related Fact and the Highlight condition. Subjects were given 15 min to complete each quiz on paper (as determined through pilot testing). Incorrect answers were then marked by the experimenter while the subject played Tetris for 10 min. The quiz was returned and the subject was given 20 min to go back and determine why they missed each item. This procedure was repeated for the second quiz. During this period (15 min per quiz, 10 min per Tetris session, 20 min per review session for a total of 90 min) the Highlight group was highlighting any sections of the chapter they identified as important. Session 2 was identical to that of Experiment 1.

3.2. Results and discussion

Fig. 3 shows the proportion of correct responses as a function of condition and final test question type. There was no difference between the tested and highlight groups for either fact

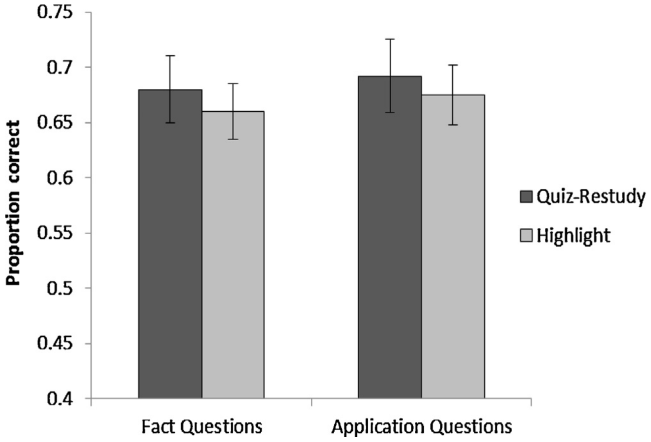

Fig. 3. Proportion correct in Experiment 2 as a function of condition and final test question type. Error bars represent the standard error.

or application questions. This observation was confirmed with a 2 (Condition: Quiz-Restudy-Quiz-Restudy, Highlight-Restudy) × 2 (Final test question: Fact, Application) mixed ANOVA, which yielded no significant results (all *F's* < 1).

Table 2 includes the predicted and observed values for each group. As in Experiment 1, the tested group appeared underconfident while the restudy group appeared overconfident. This observation was confirmed by a 2 (Performance: Predicted, Observed) × 2 (Condition: Quiz-Restudy-Quiz-Restudy, Highlight-Restudy) mixed ANOVA which yielded a significant interaction, F (1, 46) = 21.01, $p < .001$, $MSE = .01$, $\eta^2 = .31$. Planned pair-wise comparisons showed that the difference between the predicted and observed performance was significant in the tested group (underconfident), $F(1,46) = 5.66$, $p < .05$, $MSE = .01$, $\eta^2 = .11$, and the restudy group (overconfident), F (1, 46) = 15.71, $p < .001$, $MSE = .01$, $\eta^2 = .25$. Because the tested group showed no facilitative effect of restudying relative to the highlighting group, it appears that testing does not foster restudy benefits when the quizzed items are only topically related to the exam items. We were initially surprised at this finding in light of the widely held assumption that quizzes serve a useful formative assessment function, thereby fostering more effective study policies. In retrospect, however, this finding may not be so unexpected. To the extent that the subjects in the quizzing group did guide their restudy according to their quiz performances, it is likely that such restudy would focus specifically on the particular concept/fact highlighted in the quiz, not topically related content. In sum, the findings of Experiment 1 and 2 are not encouraging with regard to quizzing benefits, either direct or indirect, for topically related exam items.

4. General discussion

While testing has been convincingly demonstrated to aid retrieval of repeated information, the benefits did not emerge for topically related information with the current authentic text, quizzing, and final test materials. These results diverge from previous laboratory work demonstrating that repeated testing leads to greater transfer of knowledge (Butler, 2010; Carpenter & Kelly, 2012; Chan et al., 2006). We suggest that a possible reason for the divergent results is the difference in materials used across studies. The materials used in typical laboratory studies are strictly controlled and involve transfer to closely related items, whereas the current study examined materials that are used in classrooms in which items can be related at a general topic level. In Barnett and Ceci's (2002) framework, transfer is a multi-part process, such that individuals must recognize that prior information can be used, recall the previous information correctly, and map that information onto the current situation. The information on the "related quizzes" in the present study was arguably not closely related to the information on the final test. Accordingly, activation and recall of the information on the final test was apparently not prompted by the "related" quiz items.

Applying Barnett and Ceci's (2002) framework to the testing effect literature may thus explain why transfer only sometimes occurs. As an example, there are two primary differences between the materials used in Butler's (2010) study and the current study. First, Butler created materials that were intended to induce transfer. That is, the target material that was initially quizzed was the same target material used to create the final test questions. In the current study, the related questions were topically related, but did not necessarily require the same underlying target concept. Thus (as described above) activation and recall of "related" information might not result in benefits on the final test. Second, Butler showed a benefit for related questions when the questions were very closely related. However, when farther transfer was used in his Experiment 3, the participants were given a hint as to which prior information might be useful in the current context. Butler explained that without the hint, subjects would be unable to recognize the potential for knowledge transfer. Therefore, a major contributing factor for the discrepant results may be that subjects in the current study did not recognize the utility of previously learned information while taking the final test. According to Barnett and Ceci (2002), this may represent a failure at the first stage – recognition of a transfer situation – and/or a failure at the second stage in boosting recall for target information.

This reasoning also offers a potential reason why the review opportunities in Experiment 2 did not benefit participants. Just as quizzing may limit recall to relevant information, participants were likely only reviewing information that was directly quizzed. Though we did not find quizzing to function effectively as a formative assessment, this type of quizzing procedure (quizzing followed by student-driven review) is frequently encouraged in the classroom (Angelo & Cross, 1993). Unfortunately, there is little empirical work that evaluates the utility of formative assessment for the student (broadly defined; Dunn & Mulvenon, 2009), despite the theoretical claims that it produces enhanced metacognitive awareness, self-directed learning, and therefore general understanding in students (Clark, 2012). The data presented here suggest that students' review may be narrowly limited to the particular information targeted in the quiz, and therefore the use of quizzing to guide student review needs more careful consideration.

5. Practical applications

It appears that the testing effect, as demonstrated in a plethora of laboratory settings, may have important boundary conditions. Specifically, the standard procedure of quizzing may not benefit performance unless the summative-exam questions are closely tied to the content targeted in the quizzes. Our results suggest that current published quizzing ancillaries may not substantially improve exam performance if the exam questions are haphazardly related to the quiz materials. Such limits suggest that we need to be cautious and precise in our advice to educators. It may not be sufficient to encourage testing without specifying the situations in which testing has proven to be helpful. Along these lines, in at least introductory psychology classes (those for which quizzing ancillaries are widely available), our survey results indicate that some instructors are commonly using related exam and quiz items (with degree of relatedness unknown) or are not attempting to match exam and quiz items at all.

Still, a number of instructors indicate that they construct exam questions that are identical or very closely related to quiz items, and in those situations testing will likely be beneficial (e.g., Carpenter, Pashler, & Cepeda, 2009; Lyle & Crawford, 2011; McDaniel, Anderson, Derbish, & Morissette, 2007; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDaniel et al., 2012, Roediger, Agarwal, McDaniel, & McDermott, 2011). In addition, in one condition Mayer et al. (2009) found a testing effect for topically related items when a thorough discussion followed each quiz question; the discussion focused on how to arrive at the correct answer and general critical thinking skills. This finding suggests it may be possible to obtain benefits of testing for topically related information, but considerable explanatory feedback must accompany testing.

Across experiments, including conditions that did and did not produce a testing effect, students were generally underconfident after testing. The hope is that underconfidence propels students to devote additional efforts toward learning the information (e.g., Thomas & McDaniel, 2007), but the parameters for successful metacognitively motivated study are still uncertain, as Experiment 2 demonstrates. Additional research is needed to delineate the facilitative effects of testing, including metacognitive benefits, and the boundary conditions (e.g., see Mayer et al., 2009). The benefits of testing for identical items have been clearly demonstrated here and in laboratory and classroom experiments. However, based on the present results with materials that more closely approximate authentic classroom situations, we suggest that recommendations to educators regarding testing benefits be carefully formulated.

Conflicts of interest statement

The authors declare that they have no conflicts of interest.

Acknowledgements

This research was supported by Grant 22020166, Applying Cognitive Psychology to Enhance Educational Practice, from the James S. McDonnell Foundation. Mark McDaniel's effort in preparing the article was supported by Grant R305A110550 from the Institute of Education Sciences. We thank Victor Benassi for suggestions regarding survey construction and survey distribution, the instructors who participated in the survey, and Jessye Brick for assisting in collecting data.

Appendix 1.

Part I

1. Have you taught an introductory psychology course?

1b. Approximately how many students are enrolled in your typical introductory psychology section?

2. Do you give students quizzes in your introductory psychology course?

3. Do you implement quizzing in your introductory psychology course by giving:

In-class quizzes Yes_No_
Out-of-class (e.g., web-based) quizzes Yes_No_
[If yes] Are they required? Yes_No _

4. How often do you quiz your students in the introductory psychology course? (Please choose the option that best represents your situation)
   a. Every class, or nearly every class
   b. Weekly
   c. Monthly
   d. Less than Monthly

5. Do you make up your own quiz questions?
   a. Yes
   b. No
   c. I make up some of the questions, and others I do not

6a. Do you use questions from a test bank that you did not create (e.g., from the textbook publisher) for quizzes?

6b. Approximately what percentage of the quiz questions do you take directly from a test bank that you did not create?

7. How many EXAMS do you give your students in the introductory psychology course?

8. Approximately what percentage of your own EXAM questions do you personally create?

9. Approximately what percentage of the EXAM questions do you take directly from a test bank that you did not create?

10a. Do you try to match exam questions to previous quiz questions by giving students IDENTICAL questions on both the quiz and the exam?

10b. Approximately what percentage of the items on the exam are IDENTICAL to quiz items?

11a. Do you try to match exam questions to previous quiz questions by giving students SIMILAR BUT NOT IDENTICAL questions on both the quizzes and the later exam?

11b. Approximately what percentage of the items on the exam are SIMILAR BUT NOT INDENTICAL to quiz items?

11c. When giving SIMILAR BUT NOT IDENTICAL questions, do you believe that answering the quiz questions will improve students' performance on a later exam?

12. When you give quizzes, what is your primary goal?
   a. Assess learning
   b. Improve learning
   c. Both

Part II

1. Are you familiar with the "Testing Effect", also called "Test-Enhanced Learning"?

2. Imagine that you asked the question below on an initial QUIZ:

Thorndike asked a group of students who had learned Latin and a group of students who had not taken Latin to learn a new subject such as bookkeeping. According to Thorndike's theory of transfer by identical elements, which group should learn the new subject better?

(i) Students who knew Latin will learn better because Latin fosters proper habits of mind.
(ii) Students who had not taken Latin will learn better because the components in Latin conflict with the components in bookkeeping.
(iii) Both will learn the same.
(iv) The theory of transfer by identical elements does not make a prediction.

Based on your understanding of the testing effect (test-enhanced learning), after answering the above QUIZ question, which of the following EXAM questions would show improved performance (relative to not taking the QUIZ) on a subsequent EXAM? [Check all that apply]

a. An IDENTICAL question on the subsequent EXAM.
b. A SIMILAR (to the quiz question) BUT NOT IDENTICAL question on the subsequent EXAM. An example could be:

A researcher asked a group of students who had learned Latin and a group of students who had not taken Latin to learn a new subject such as bookkeeping. Which theory would predict that both groups will learn the new subject the same?

(i) Thorndike's theory of transfer by identical elements.
(ii) Doctrine of formal discipline
(iii) Both (i) and (ii) make this prediction.
(iv) Neither (i) or (ii) makes this prediction.

c. A question covering the SAME GENERAL TOPIC BUT DIFFERENT CONCEPTS OR THEORIES on the subsequent EXAM. An example could be:

Thorndike asked a group of students who had learned Latin and a group of students who had not taken Latin to learn a new subject such as bookkeeping. According to doctrine of formal discipline, which group should learn the new subject better?

(i) Students who knew Latin will learn better because Latin fosters proper habits of mind.
(ii) Students who had not taken Latin will learn better because the components in Latin conflict with the components in bookkeeping.
(iii) Both will learn the same.
(iv) The doctrine of formal discipline does not make a prediction.

d. A question that is UNRELATED to the quiz question on the subsequent EXAM. An example could be:

Jenny is given a word problem in math class, and she immediately classifies the problem as a right triangle problem. Which type of knowledge is Jenny using to classify this problem as a right triangle problem?

(i) Procedural
(ii) Situational
(iii) Strategic
(iv) Schematic

3. Have you analyzed data from your course(s) regarding potential effects of quizzes, or conducted studies on the testing effect (test-enhanced learning)?

Appendix B. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:10.1016/j.jarmac.2014.07.001.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861–876.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives (abridged ed.)*. New York, NY: Addison Wesley Longman, Inc.

Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers* (2nd ed.). San Francisco: Jossey-Bass.

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

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*, 7–74.

Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). In B. S. Bloom (Ed.), *The taxonomy of educational objectives: The classification of educational goals (Handbook 1: Cognitive domain)*. New York, NY: David McKay Company, Inc.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1118–1133.

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*, 514–527.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*, 279–283.

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review*, *19*, 443–448.

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.

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

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*, 205–249.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research, & Evaluation*, *14*. http://pareonline.net/getvn.asp?v=14&n=7

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

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.

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*, 528–558.

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*, 157–163.

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, *38*, 94–99.

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., Bulger, M., Campbell, J., Knight, A., & Zhang, H. (2009). Clickers in the classroom: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, *34*, 51–57.

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

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *11*, 371–385.

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

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*, 299–414.

McDaniel, M. A., Wildman, K., & 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*, 18–26.

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*, 360–372.

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*, 3–21.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning (NCER 2007–2004)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Phelan, J. (2009). *What is life? A guide to biology*. New York: W H Freeman.

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

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.

Roediger, H. L., 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*, *11*, 382–395.

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre, & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (pp. 1–36). Oxford: Elsevier.

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

Schneider, W., Eschman, A., & Zuccolotto, A. (2007). *E-Prime 2 user's guide*. Pittsburgh, PA: Psychology Software Tools.

Stiggins, R. J., Arter, J. A., Chappius, J., & Chappius, S. (2006). *Classroom assessment for student learning: Doing it right-using it well*. Portland: Educational Testing Service.

Thomas, A. K., & McDaniel, M. A. (2007). The negative cascade of incongruent generative study-test processing in memory and metacomprehension. *Memory & Cognition*, *35*, 668–678.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*, 175–184.