Learning and Instruction xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Learning and Instruction

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

Effects of study intention and generating multiple choice questions on expository text retention

Vincent Hoogerheide[a,b,*], Justine Staal[b], Lydia Schaap[b,c], Tamara van Gog[a,b]

[a] Department of Education, Utrecht University, The Netherlands
[b] Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, The Netherlands
[c] Learning and Innovation Centre, Avans University of Applied Sciences, The Netherlands

ARTICLE INFO

Keywords:
Generative learning strategies
Self-generated questions
Memory
Retention

ABSTRACT

Teachers often recommend their students to generate test questions and answers as a means of preparing for an exam. There is a paucity of research on the effects of this instructional strategy. Two recent studies showed positive effects of generating test questions relative to restudy, but these studies did not control for time on task. Moreover, the scarce research available has been limited to the effects of generating open-ended questions. Therefore, the aim of this study was to investigate whether generating multiple-choice test questions would foster retention (as measured by a multiple-choice test) relative to restudy when time would be kept constant across conditions. Using a 2 × 2 design, university students (N = 143) studied a text with the intention of either generating test items or performing well on a test, and then either generated multiple-choice items or restudied the text. Retention was measured by means of a multiple-choice test, both immediately after learning and after a one-week delay. Results showed no effects of study intention. Generating multiple-choice items resulted in lower test performance than restudying the text for the same amount of time.

1. Introduction

It is quite common for teachers to recommend students to generate test questions and answers to these questions when preparing for an exam (Weinstein, McDermott, & Roediger, 2010). This seems to be a sensible recommendation, because generative learning strategies such as question generation tend to foster learners' engagement with the learning material, help them focus on main ideas in the material, and improve their long-term retention relative to merely (re)studying the material (King, 1992, 1994; Mayer, 2003; for a review of generative learning strategies: Fiorella & Mayer, 2016). Given the wealth of research on the benefits of taking a practice test on long-term retention (see Roediger, Putnam, & Smith, 2011; Rowland, 2014), the paucity of research on the efficacy of having (untrained) students generate test questions is surprising. In the absence of practice tests, instructing students to generate test questions themselves may be the next best option (Weinstein et al., 2010). It has been shown that students benefit greatly from elaborate training programs designed to help them acquire certain 'self-questioning' skills such as asking oneself 'why', 'what', and 'how' questions (e.g., García, García, Berbén, Pichardo, & Justicia, 2014; for reviews, see; Rosenshine et al., 1996; Wong, 1985). However, only two studies investigated the effects of *generating test items* (i.e., test questions and answers to those questions) as a means of preparing for an exam for students who had not received extensive training beforehand (Bugg & McDaniel, 2012; Weinstein et al., 2010).

Both studies showed beneficial effects. Weinstein et al. (2010) provided a group of adults –undergraduates and graduates– with three short texts in a within-subjects design. They were instructed to read each text once and to subsequently restudy the text once, answer experimenter-made short answer questions (i.e., test-*taking*), or generate short answer questions and answers themselves. Those who generated test questions and those who answered experimenter-made test questions consistently outperformed those who had restudied. No performance differences were found between those who generated and those who answered test questions. These findings were robust across different materials (i.e., different passages of 350 and 575 words), tests (i.e., a short answer test and a cued recall test), and test moments (i.e., immediately after learning and two days later). Time on task was not controlled for, however, and in the question generation condition participants typically took more than three times longer than they did in both other conditions. Therefore, the additional time on task might provide an alternative explanation for why question generation was more effective than restudy.

Bugg and McDaniel (2012) provided undergraduate students with

* Corresponding author. Department of Education, Utrecht University, P.O. Box 80140, 3508 TC, The Netherlands.
E-mail address: v.hoogerheide@uu.nl (V. Hoogerheide).

https://doi.org/10.1016/j.learninstruc.2017.12.006
Received 3 January 2017; Received in revised form 21 November 2017; Accepted 20 December 2017

several short texts in a between-subjects design. After studying a text once, students either restudied the text once, generated and answered three questions about a detail or fact that could be answered with a single sentence (detail condition), or generated and answered three questions that required multiple sentences to answer (conceptual condition). Afterwards, students received three 'detail' and three 'conceptual' cued-recall questions per passage on the posttest. There were no differences among conditions on the detail test items, but those who generated and answered conceptual questions performed better on the conceptual cued-recall posttest items than both other conditions. Time on task, however, was not reported. Given how complex, effortful, and time consuming it is to create questions that integrate different textual elements and answers to these questions, it is likely that students in the conceptual condition would have spent more time than students in the other conditions.

Because research on the effects of generating test questions has been scarce, there are various important open issues to address. Firstly, although recent findings suggest that generating test questions is an effective learning strategy (Bugg & McDaniel, 2012; Weinstein et al., 2010), it is an open question whether generating test questions is an *efficient* instructional strategy. Would it still be beneficial for learning when time on task is controlled for? Next to time on task, mental effort is another important aspect of efficiency. That is, when effortful learning activities, like generating test questions, result in better learning outcomes (when time is kept constant), asking learners to invest this additional effort pays off. However, when such activities result in the same or lower learning outcomes, they are less cognitively efficient (Hoffman & Schraw, 2010; Van Gog & Paas, 2008). Knowing whether generating test questions is an efficient instructional strategy is important for educational practice, as students typically only have limited time available for preparing for exams and because of limitations in working memory capacity, their cognitive resources are easily depleted, so their effort should be invested in processes that actually contribute to learning (Sweller, Ayres, Kalyuga, 2011).

Secondly, the research available thus far has focused on the effects of generating open-ended questions on a cued-recall or short answer final test. It is as yet unclear whether generating multiple-choice questions in preparation for a multiple-choice test would also be an effective and efficient learning activity. This question is both theoretically relevant, considering the different cognitive processes involved in generating open-ended questions and multiple-choice questions (which require more elaboration by requiring learners to think of alternative answers), and important for educational practice, where multiple-choice tests are commonly used.

A third open question is whether merely studying a text with the intention of generating test questions (i.e., without actually doing so) would already be conducive to students' learning. Weinstein et al. (2010) and Bugg and McDaniel (2012) controlled for the potential benefits that studying with a generation intention might have by only informing participants about the question generation or restudy activity *after* the text had already been studied. Yet the benefits of training programs designed to help students acquire self-questioning skills are presumed to not only arise from both the act of generating and answering questions, but also from studying the learning material in such a way that questions and answers can be generated later on (King, 1994; Palincsar & Brown, 1984; Rosenshine et al., 1996; Wong, 1985).

1.1. The present study

The main aim of the present study was to investigate whether generating multiple-choice test items would be more effective (i.e., lead to better performance on an immediate and delayed multiple-choice test) and efficient (i.e., in terms of mental effort expended during the learning phase in relation to test performance) than restudying the text when time for question generation and restudy is kept equal. On the one hand, generating multiple-choice questions can be expected to stimulate students to not only identify the main ideas in the text, but to also elaborate on those ideas through the process of generating plausible alternative but incorrect answers. This elaboration can be expected to help students develop a richer mental model of the text, for instance by creating more connections to their prior knowledge (Reder, Charney, & Morgan, 1986) and by increasing the distinctiveness of their memory representations (Stein, Littlefield, Bransford, & Persampieri, 1984). As such, generating questions can be expected to be more effortful, but also a qualitatively better way (i.e., resulting in better posttest performance) to spend the available time with the learning material than restudy. If so, this may become especially evident at a delayed posttest (Fiorella & Mayer, 2016). Indeed, benefits of other generative learning strategies are known to only emerge on a delayed posttest, or to become more pronounced after a one-week delay (e.g., taking practice tests: Roediger & Karpicke, 2006; teaching on video: Fiorella & Mayer, 2014; generating and reflecting upon alternative diagnoses: Mamede et al., 2012).

On the other hand, if the benefits of generating test questions found by Bugg and McDaniel (2012) and Weinstein et al. (2010) emerged only because students spent more time with the material, then question generation might no longer be more effective than restudy when time on task is controlled for. If generating questions would be more effortful than restudy, but would not result in better learning outcomes, then generating questions would be less cognitively efficient (Hoffman & Schraw, 2010; Van Gog & Paas, 2008).

The second aim of this study was to address the role of study intention, so whether merely studying a text with the intention of generating questions (i.e., without actually doing so) would already be conducive to students' retention of the material. Compared to studying with a test-*taking* intention, which is how students normally study, a generation (or test-*making*) intention may stimulate students to study with higher levels of engagement (King, 1994; Palincsar & Brown, 1984), to monitor their own comprehension (Wong, 1985), and focus on the central concepts and facts of the learning material (Rosenshine et al., 1996).

2. Method

2.1. Participants and design

Participants were 143 undergraduate students (38 male; *M* age = 20.27 years, *SD* = 2.21) who studied Psychology at a Dutch university and received course credits for their participation. The experiment consisted of five phases: 1) pretest, 2) learning phase I, 3) learning phase II, 4) immediate posttest, and 5) delayed posttest. The study had a 2 × 2 design with between-subject factors Study Intention (Test Intention vs. Generation Intention) and Question Generation (No: Restudy vs. Yes: Generate Questions). Study Intention was manipulated in learning phase I, and Question Generation in learning phase II. Participants were randomly allocated to one of the four conditions: Test Intention – Restudy (n = 36), Test Intention – Generate Questions (n = 34), Generation Intention – Restudy (n = 36), or Generation Intention – Generate Questions (n = 37). Participants were expected to have little knowledge of the topic of the learning materials and indeed, participants indicated that their prior knowledge of earthquakes was low (average score of 3.92 out of 9, see Table 1).

2.2. Materials

All the study and test materials were paper-based.

2.2.1. Pretest

The pretest emulated Fiorella and Mayer's (2013, 2014) procedure for assessing prior knowledge by asking participants to rate their knowledge of earthquakes on a scale ranging from 1 (very low) to 5 (very high), and to mark each of the following items that applied to them: "I had geography courses in my final two years of high school," "I

Table 1
Mean (SD) of test scores per condition.

	Test Intention		Question Intention	
	Restudy	Question Generation	Restudy	Question Generation
Pretest (range 1–9; $N = 139$)	3.49 (1.60)	3.84 (1.79)	4.22 (1.77)	4.11 (1.70)
Immediate Posttest - Percentage Correct Total Score (range 0–100%; $N = 132$)	83.33% (10.50)	73.57% (10.80)	80.95% (12.19)	69.84% (12.56)
Delayed Posttest - Percentage Correct Total Score (range 0–100%; $N = 132$)	78.10% (12.85)	68.81% (13.20)	76.19% (11.72)	65.09% (14.16)
Immediate Posttest - Percentage Correct – Factual Items (range 0–100%; $N = 132$)	82.00% (10.95)	68.67% (12.24)	76.94% (14.70)	65.28% (15.40)
Delayed Posttest - Percentage Correct – Factual Items (range 0–100%; $N = 132$)	75.00% (13.83)	62.00% (15.62)	72.22% (13.96)	58.33% (17.65)
Immediate Posttest - Percentage Correct – Conceptual Items (range 0–100%; $N = 132$)	86.67% (17.03)	85.83% (15.65)	90.97% (14.82)	81.25% (19.25)
Delayed Posttest - Percentage Correct – Conceptual Items (range 0–100%; $N = 132$)	85.83% (16.97)	85.83% (15.65)	86.11% (17.37)	81.94% (15.37)

know exactly what causes earthquakes," "I can accurately describe the focus of an earthquake," and "I know what a seismograph is." Note that we used a self-report measure (cf. Fiorella & Mayer, 2013, 2014), because a prior knowledge test might affect students' study and question generation behavior (i.e., the questions could serve as a cue for what information from the text is important).

2.2.2. Learning phase I

In learning phase I, participants studied a 993 words text on why, when, and where earthquakes occur. This text was the same as used in the study by Agarwal, Karpicke, Kang, Roediger, and McDermott (2008). A prompt was placed at the start and at the end of the text. For those in the Test Intention Conditions, these prompts asked students to study the text in such a way that they would be able to complete a multiple-choice test about the most important facts and concepts of the text. For those in the Generation Intention Conditions, the prompts instructed students to study the text in such a way that they would be able to create a multiple-choice test about the most important facts and concepts of the text as if they were a teacher who had to design a multiple-choice test.

2.2.3. Learning phase II

In the second learning phase, half of the participants restudied the text, while the other half generated multiple-choice test items with three answer alternatives (with the text still available to them). Those in the Restudy Conditions were instructed to restudy the text, either in such a way that they would be able to complete a multiple-choice test about the most important facts and concepts of the text later on (Test Intention – Restudy Condition) or in such a way that they would be able to create a multiple-choice test about the most important facts and concepts of the text as if they were a teacher who had to design a multiple-choice test (Generation Intention – Restudy Condition). Those in the Question Generation Conditions were instructed to design a multiple-choice test with three answer possibilities about the most important facts and concepts of the text.

2.2.4. Posttests

The immediate and delayed posttest consisted of 14 multiple-choice questions with three answer alternatives. The questions tested retention of the most important facts (10 items) and concepts (4 items) from the text. Example items are: "How often do earthquakes take place on average? A) once every 10 s, B) once every 30 s, C) once every 60 s" and "What causes most earthquakes? A) plates that move away from each other, B) plates that slip past each other, C) colliding plates". These two versions consisted of the same questions and answer alternatives, but both the questions and answer alternatives per question differed in order between the two versions.

2.2.5. Mental effort

Participants rated how much mental effort they invested in learning phase I, learning phase II, and each of the posttest items (immediately after completing that phase or item), on a scale ranging from (1) very, very low effort to (9) very, very high effort (Paas, 1992).

2.3. Procedure

The study was run in the university lab. Participants were seated in individual cubicles, with a maximum of eight students tested in parallel. The study consisted of two sessions. The first session lasted approximately 60 min. The experimenter first gave a general introduction to the experiment and handed each student an envelope containing four different booklets. Then, participants were instructed to take out the first booklet, which contained a short demographic questionnaire and the pretest, for which participants received two minutes. After placing the first booklet on the corner of their table, participants were instructed to take the second booklet out of the envelope. Those in the Test Intention Conditions were instructed to study the experimental text for 6 min in such a way that they would be able to complete a multiple-choice test about the most important facts and concepts of the text later on. Those in the Generation Intention Conditions were instructed to study the experimental text for 6 min in such a way that they would be able to create a multiple-choice test about the most important facts and concepts of the text as if they were a teacher who had to create a multiple-choice test. The study time duration was based on a pilot test, which indicated that 6 min was sufficient to study the whole text at least once. After the six minutes were up, participants rated how much mental effort they invested in the first learning phase, and placed the second booklet on the corner of their table. Half of the participants then restudied the text for 12 min. Those in the Test Intention – Restudy Condition were instructed to restudy the text in such a way that they would be able to complete a multiple-choice test about the most important facts and concepts of the text later on. Those in the Generate Intention – Restudy Condition were instructed to restudy so that they would be able to create a multiple-choice test about the most important facts and concepts of the text. The other students generated multiple-choice test items with three answer possibilities (with the text present). It was emphasized repeatedly that participants should spend the full time available on restudying/generating questions. When time was up, participants rated how much effort they invested in the second learning phase, after which they completed the immediate posttest (booklet 4, max. 20 min). The second session took place one week later and lasted maximally 20 min, during which participants completed the delayed posttest (booklet 5). Half of the participants within each condition received version A on the immediate posttest and version B on the delayed posttest, while for the other half the order was reversed.

2.4. Data analysis

Students' self-reported pretest scores could range from 1 to 9 points. One point could be earned for each marked item (range 0–4), plus 1 through 5 points depending on the level that was indicated on the knowledge rating scale. As for the posttest, each correctly answered item was worth one point, which means that a total of 14 points could be earned. Posttest scores were converted to percentage correct scores. For invested mental effort, averages were computed separately for effort invested in the immediate posttest and in the delayed posttest.

Four participants were removed from all analyses, three because they skipped a page on the posttest by accident which caused them to miss several questions (one participant from the Test Intention – Restudy Condition, one from the Test Intention – Generate Questions Condition, and one from the Generation Intention – Generate Questions Condition), and one because s/he did not comply with the instructions (i.e., did not generate any questions during learning phase II; Test Intention – Generate Questions Condition). An additional seven participants were removed from the posttest analyses (performance and effort) because they were absent during the delayed posttest (five participants from the Test Intention – Restudy Condition and two from the Test Intention – Generate Questions Condition). Three participants had one missing effort rating on the posttest, which was replaced with their series mean. In addition, two participants did not fill in their invested mental effort in learning phase I, and seven did not fill in their invested mental effort in learning phase II, so they were not included in the analysis involving the respective measure.

3. Results

Partial eta squared is provided as a measure of effect size, with values of 0.01, 0.06, and 0.14 representing a small, medium, and large effect respectively (Cohen, 1988). Because in a factorial design, partial eta squared can result in an overestimation of the true effect size of an effect (Levine & Hullett, 2002), we additionally report cohen's *d* for the between-subject comparisons, where values of 0.20, 0.50, and 0.80 are indicative of a small, medium, and large effect size respectively (Cohen, 1988).

Before addressing our hypotheses, ANOVAs were computed to check whether the four conditions were comparable in terms of age, gender, and self-reported prior knowledge on the topic. Results showed no significant difference among conditions in terms of age, $F < 1$, gender, $F(3, 135) = 1.09$, $p = .358$, $\eta_p^2 = 0.024$, $d = 0.311$, or self-reported prior knowledge, $F(3, 135) = 1.29$, $p = .282$, $\eta_p^2 = 0.028$, $d = 0.338$. We also checked whether there was a relationship between self-reported prior knowledge and posttest performance. Results showed no significant correlations between students' (self-reported) pretest performance and performance on the immediate posttest ($r = 0.065$, $p = .460$) or the delayed posttest ($r = 0.163$, $p = .061$).

3.1. Test performance

Performance scores on the immediate and delayed posttests are presented in Table 1. A 2 × 2 repeated measures ANOVA on posttest performance, with Study Intention (Test, Generation) and Question Generation (Restudy, Generate Questions) as between-subject factors and Test Moment (Immediate, Delayed) as within-subjects factor, showed a main effect of Test Moment, $F(1, 128) = 30.98$, $p < .001$, $\eta_p^2 = 0.195$, indicating that students' performance decreased from the Immediate ($M = 76.79\%$; $SD = 12.74$) to the Delayed Posttest ($M = 71.92\%$; $SD = 13.74$). There was no main effect of Study Intention, $F(1, 128) = 2.27$, $p = .134$, $\eta_p^2 = 0.017$, $d = 0.266$. There was a significant main effect of Question Generation, $F(1, 128) = 28.04$, $p < .001$, $\eta_p^2 = 0.180$, $d = 0.936$, showing that those who had restudied the learning material ($M = 79.64\%$; $SD = 11.19$) performed significantly better than those who had generated multiple-choice items ($M = 69.33\%$; $SD = 11.19$). There were no significant interaction effects, $ps > 0.684$.[1]

We explored whether the performance advantage of those who restudied over those who generated questions depended on the nature of the posttest questions, that is, on whether the questions assessed students' memory of facts (i.e., isolated pieces of information) or their memory of concepts (i.e., underlying principles and relationships). Note that the answers to both fact and concept questions could literally be found in the text. Two repeated measures ANOVAs were conducted with Study Intention (Test, Generation) and Question Generation (Restudy, Generate Questions) as between-subject factors and Test Moment (Immediate, Delayed) as within-subjects factor. Concerning performance on fact questions, results showed a main effect of Test Moment, $F(1, 128) = 32.07$, $p < .001$, $\eta_p^2 = 0.200$, indicating that students' memory for facts decreased from the Immediate ($M = 73.03\%$; $SD = 14.98$) to the Delayed Posttest ($M = 66.74\%$; $SD = 16.74$). There was no main effect of Study Intention, $F(1, 128) = 2.66$, $p = .105$, $\eta_p^2 = 0.020$, $d = 0.289$, but there was a main effect of Question Generation, $F(1, 128) = 32.37$, $p < .001$, $\eta_p^2 = 0.202$, $d = 1.006$, indicating that those who restudied ($M = 76.54\%$; $SD = 13.10$) outperformed those who had generated questions ($M = 63.57\%$; $SD = 13.10$) on fact questions. There was no significant interaction effect, $ps > 0.568$. Concerning performance on concept questions, results showed no main effect of Test Moment, $F(1, 128) = 1.11$, $p = .293$, $\eta_p^2 = 0.009$, no main effect of Study Intention, $F < 1$, no main effect of Question Generation, $F(1, 128) = 1.93$, $p = .167$, $\eta_p^2 = 0.015$, $d = 0.246$, and no significant interaction effects, $ps > 0.179$.

3.2. Mental effort

3.2.1. Learning phase

Mental effort data are presented in Table 2. In the first learning phase, only Study Intention could have affected effort investment. Therefore, an ANOVA was conducted on effort invested in the first learning phase with Study Intention (Test, Generation) as between-subjects factor, which showed no significant difference, $F < 1$. A 2 × 2 ANOVA on mental effort invested during the second learning phase, with Study Intention (Test, Generation) and Question Generation (Restudy, Generate Questions) as between-subject factors, did not show a main effect of Study Intention, $F(1, 128) = 3.72$, $p = .056$, $\eta_p^2 = 0.028$, $d = 0.341$. There was, however, a main effect of Question Generation, $F(1, 128) = 5.68$, $p = .019$, $\eta_p^2 = 0.042$, $d = 0.421$, indicating that Generating Questions ($M = 6.57$; $SD = 1.63$) was more effortful than Restudy ($M = 5.85$; $SD = 1.88$). There was no significant interaction between Study Intention and Question Generation, $F < 1$.

3.2.2. Test phase

A 2 × 2 repeated measures ANOVA was conducted on mental effort invested in the posttests, with Study Intention (Test, Generation) and Question Generation (Restudy, Generate Questions) as between-subject factors and Test Moment (Immediate, Delayed) as within-subjects factor. There was a main effect of Test Moment, $F(1, 128) = 16.63$, $p < .001$, $\eta_p^2 = 0.115$. Participants invested significantly more effort in answering questions on the Delayed Posttest ($M = 3.55$; $SD = 1.22$) than on the Immediate Posttest ($M = 3.20$; $SD = 1.13$). There was no

[1] While examining students' generated questions, it became apparent that some of the answer possibilities were not relevant to the topic of the text and obviously incorrect. We reran the main analyses to explore whether restudy would still lead to better retention than question generation if students that created one or multiple of such poor incorrect answer possibilities (n = 12) were excluded from the analysis. The same findings emerged, meaning that there was no main effect of Study intention (p = .282, $\eta_p^2 = 0.010$, d = 0.201), a main effect of Question Generation (p < .001, $\eta_p^2 = 0.181$, d = 0.939) with the Restudy Conditions outperforming the Question Generation Conditions, and no significant interaction effects (ps > 0.705).

Table 2
Mean (SD) of invested mental effort (range 1–9) per condition.

	Test Intention		Question Intention	
	Restudy	Question Generation	Restudy	Question Generation
Learning Phase I (Test vs. Question Intention; $N = 137$)	5.23 (1.72)	6.20 (1.10)	5.53 (1.38)	6.03 (1.50)
Learning Phase II (Restudy vs. Generate Questions; $N = 132$)	5.49 (2.01)	6.33 (1.94)	6.19 (1.70)	6.81 (1.25)
Immediate Posttest ($N = 132$)	2.77 (1.12)	3.62 (0.86)	2.82 (0.97)	3.57 (1.27)
Delayed Posttest ($N = 132$)	3.36 (1.49)	4.05 (1.05)	3.19 (1.15)	3.84 (1.05)

main effect of Study Intention, $F < 1$, but there was a significant main effect of Question Generation, $F(1, 128) = 15.50$, $p < .001$, $\eta_p^2 = 0.108$, $d = 0.696$. Participants who had restudied the learning materials indicated having invested significantly less effort on the posttests ($M = 3.04$; $SD = 1.01$) than those who generated questions ($M = 3.73$; $SD = 1.01$). None of the interaction effects were significant, $ps > 0.123$.

3.3. Explorative follow-up analyses: does generation relate to retention?

We conducted several additional analyses to explore whether and how the number of generated questions, the type of generated questions, and the topical match between generated and posttest questions related to posttest performance in the Question Generation Conditions. On average, students generated a total of 8.50 questions ($SD = 2.85$). The number of questions generated did not correlate significantly with performance on either the Immediate ($r = -0.025$, $p = .842$) or Delayed Posttest ($r = -0.102$, $p = .416$). Some of these questions were incomplete in the sense that they missed one or more answer alternatives. If only complete questions are considered, then the average number of generated questions is slightly lower ($M = 6.93$, $SD = 2.90$), and there is still no correlation with students' test performance (Immediate Posttest: $r = -0.133$, $p = .285$; Delayed Posttest: $r = 0.132$, $p = .289$).

Because the benefits of generating multiple-choice questions for retention may depend on the type of questions students generate, we scored for each generated question with three answer possibilities (i.e., complete questions only) whether the question covered a fact or a concept from the text.[2] To measure the reliability of the ratings, two raters first independently rated the questions generated by 14 participants (i.e., more than 10%). Because of the high intra-class correlation coefficient (1.00), the remainder was scored by one rater. The number of fact questions students generated ($M = 4.05$; $SD = 1.95$) did not correlate significantly with their test performance (Immediate Posttest: $r = -0.154$, $p = .224$; Delayed Posttest: $r = 0.021$, $p = .870$), nor was there a significant association between the number of concept questions generated ($M = 3.16$; $SD = 1.61$) and test performance (Immediate Posttest: $r = -0.019$, $p = .881$; Delayed Posttest: $r = 0.232$, $p = .065$). We further explored the association between the number of fact or concept questions students generated and their performance on the posttest fact and concept questions, but there were no significant correlations, $ps > 0.05$.

Because generating multiple-choice questions is likely most useful for improving retention when the topic of the generated questions corresponds with the topic of the posttest questions, we rated for each student whether the topics of the 14 posttest items was reflected in their generated questions (1) or not (0). We did this once for all generated questions (regardless of completeness) and once for the complete questions only (i.e., those that had three answer options). To measure the reliability of the ratings, two raters first independently rated whether the topics of the posttest items were reflected in participants' generated questions ($n = 14$, more than 10%). The intra-class correlation coefficient was 0.73 for all generated questions and 0.90 for complete questions only. Any discrepancies were discussed and because the inter-rater reliabilities were acceptable, the remainder was scored by one rater. Then, we calculated the percentage of correct answers on the posttest items that students had generated questions about, and the percentage of correct answers on the posttest items that students had not generated questions about (again, we did this once for all generated questions, and once for complete questions only), and compared performance with paired-samples *t*-tests (see Table 3). As Table 3 shows, students in the Question Generation Conditions performed significantly better on the posttest items that they had generated questions about than on the posttest items that they had not generated questions about. This was a large effect that was present on both the Immediate and Delayed Posttest, and irrespective of whether all generated questions were taken into account or only the complete ones.

Subsequently, we compared performance of the Question Generation Conditions on the items that students had and had not generated questions about, to (overall) performance of the Restudy Conditions, with independent-samples *t*-tests. As can be seen in Table 4, performance on the posttest items about which questions had been generated was at the same level as (overall) performance in the Restudy condition. However, when comparing the level of performance on the items about which no questions had been generated to the (overall) level of performance in the Restudy condition, the Restudy condition performed significantly better (see Table 5). Again, this applied on both the Immediate and Delayed Posttest, and irrespective of whether all generated questions were taken into account or only the complete ones. Thus, even though they should be interpreted with caution, these results seem to qualify the finding that the Restudy conditions outperformed the Question Generation conditions in the main analyses. When questions were generated on the topics covered in the test items, performance was as good as in the Restudy condition, though not better, as one might expect if generating multiple-choice questions was an effective generative learning strategy.

4. Discussion

The main question we investigated was whether generating multiple-choice test questions would be more effective and efficient than restudying the text when time available for restudy and question generation was kept equal. The second question addressed was whether studying a text with the intention of generating multiple-choice test items, so without actually doing so, would already foster students' retention compared to studying to complete a test.

Regarding study intention, we found no evidence that studying a text with the intention of generating multiple-choice test items affected students' effort investment or fostered their retention test performance compared to studying with the intention of completing a test. This seems in line with several studies on teaching expectancy showing null-findings. In these studies, students were instructed to study a text with

[2] Two participants failed to generate any *complete* questions that focused on the content of one of the posttest items. These two were omitted from the analyses focusing on only the complete questions.

Table 3
Results of paired samples *t*-test comparing the percentage correct scores on the posttest items that did vs. did not have a topical match with students' generated questions during the learning phase.

	Percentage correct score on the posttest items that had a topical match with students' generated questions	Percentage correct score on the posttest items that had *no* topical match with students' generated questions	*df*	*t*-value	*p*-value	Cohen's *d*
All generated questions						
Immediate Posttest	82.36% (16.54)	64.83% (17.98)	65	5.53	< .001	1.01
Delayed Posttest	76.29% (18.57)	60.67% (18.65)	65	5.20	< .001	0.84
Complete questions only						
Immediate Posttest	81.64% (20.09)	65.65% (17.07)	62	4.53	< .001	0.86
Delayed Posttest	77.62% (20.51)	60.87% (18.03)	62	5.23	< .001	0.87

Table 4
Results of independent samples t-tests comparing the percentage correct test scores of those in the Question Generation Conditions when the posttest items did have a topical match with students' generated questions to those in the Restudy Conditions.

	Restudy Conditions	Question Generation Conditions		*df*	*t*-value	*p*-value	Cohen's *d*
		All Generated Questions	Complete Questions Only				
Immediate Posttest	82.03% (11.43)	82.36% (16.54)		130	−0.13	.894	−0.02
Immediate Posttest	82.03% (11.43)		81.64% (20.09)	127	0.14	.891	0.02
Delayed Posttest	77.06% (11.72)	76.29% (18.57)		130	0.28	.777	0.05
Delayed Posttest	77.06% (11.72)		77.62% (20.51)	127	−0.19	.848	−0.03

Table 5
Results of independent samples t-tests comparing the percentage correct test scores of those in the Question Generation Conditions when the posttest items did not have a topical match with students' generated questions to those in the Restudy Conditions.

	Restudy Conditions	Question Generation Conditions		*df*	*t*-value	*p*-value	Cohen's *d*
		All Generated Questions	Complete Questions Only				
Immediate Posttest	82.03% (11.43)	64.83% (17.98)		130	6.56	< .001	1.14
Immediate Posttest	82.03% (11.43)		65.65% (17.07)	127	6.43	< .001	1.13
Delayed Posttest	77.06% (11.72)	60.67% (18.65)		130	6.04	< .001	1.05
Delayed Posttest	77.06% (11.72)		60.87% (18.03)	127	6.07	< .001	1.06

the intention of teaching the content later on or with the intention to complete a test (e.g., Hoogerheide, Deijkers, Loyens, Heijltjes, & Van Gog, 2016; Hoogerheide, Loyens, & Van Gog, 2014, Experiment 1; Renkl, 1995). Note that some other studies did show beneficial effects of a teaching expectancy on learning outcomes (e.g., Bargh & Schul, 1980; Hoogerheide et al., 2014; Experiment 2; Muis, Psaradellis, Chevrier, Di Leo, & Lajoie, 2016; Nestojko, Bui, Kornell, & Bjork, 2014). Fiorella and Mayer (2013, 2014) postulated that the mixed findings regarding teaching expectancy might have emerged because teaching expectancy only leads to short-term and not to long-term benefits, but in the present study we did not find any indication that the results differed across the immediate and delayed posttest.

As for the effects of generating multiple-choice items, we found that students who restudied invested less effort and attained better retention test performance than those who generated test questions. Exploratory follow-up analysis showed that the restudy conditions outperformed the question generation conditions on the fact questions but not the concept questions. Note though that on the concept questions, question generation was only as effective as restudy and not more effective as one would expect based on prior research (Bugg & McDaniel, 2012). However, these results have to be interpreted with caution, because the posttest contained only a limited number of concept questions.

Several exploratory analyses were conducted to further examine the relationship between question generation and retention. There was no association between the number of questions students had generated during the learning phase and retention. Students in the question generation conditions did perform better on the posttests if the posttest items had a topical match with their generated questions than when such a topical match was absent. This finding that generating items on topics corresponding to posttest item topics improves performance corresponds with findings of Bugg and McDaniel (2012). Moreover, in the context of test-*taking*, it is also well-established that after an initial study phase, recalled information is remembered better than unrecalled information (e.g., Pyc & Rawson, 2009). However, even when we look only at the items that topically matched the posttest, question generation was not more effective for retention than restudy.

Thus, restudy was both a more effective instructional strategy than generating multiple-choice questions and more efficient in the sense that greater test performance was attained with less effort investment (Van Gog & Paas, 2008). Importantly, while the effects of other generative learning strategies tend to differ across immediate and delayed tests (Fiorella & Mayer, 2016), we found the same pattern of results across both test moments. These findings do beg the question of why students benefitted so little from generating multiple-choice questions relative to restudy. Restudy is a notoriously ineffective study strategy, particularly for delayed posttests (Rowland, 2014). It is rather surprising that those who generated multiple-choice items performed worse than those who restudied, because other generative strategies such as taking a practice test (Roediger, Putnam, & Smith, 2011; Rowland, 2014) and teaching on video (e.g., Hoogerheide et al., 2016) have been shown to foster long-term retention compared to restudy.

One possible explanation is that generating test questions and answers is not a qualitatively better way to spend the available time with the learning material than restudy, that is, does not help students to process the learning materials more deeply or to generate a richer cognitive schema than restudy. In prior research with self-paced study

conditions, generating test items may 'only' have stimulated learners to spend more time on the learning phase, which can be expected to be beneficial for retention (cf. Bugg & McDaniel. 2012; Weinstein et al., 2010). We kept time on task equal and did not find any benefits. Another potential explanation is that generating multiple-choice questions, and especially the alternative but incorrect answers, may elicit extraneous processing (Mayer, 2014). That is, even though one might argue that generating multiple-choice questions leads to elaboration of the essential information (i.e., information related to the instructional goal), it might also induce cognitive processing that is extraneous to (i.e., irrelevant for) the instructional goal. Learners may benefit most from generative learning strategies when they do not impose too much extraneous processing (cf. findings on collaborative learning, which is most useful when the benefits of group discussions outweigh the transaction costs of group communication; Kirschner, Paas, & Kirschner, 2009). Future research could test this extraneous processing hypothesis by comparing the effectiveness and efficiency of different generative learning strategies, such as generating multiple-choice vs. open questions.

A related issue is students may need guidance or training before generating test questions becomes more effective than restudy. Those who have little to no experience with this instructional strategy might struggle to spend the available time in an efficient way and focus more, for example, on designing the questions rather than using the strategy as a tool to elaborate on the content of the learning material. Indeed, earlier studies in which students were provided with elaborate training programs to help them acquire 'self-questioning' skills did quite consistently show beneficial effects of asking questions and generating answers to those questions (Rosenshine et al., 1996; Wong, 1985). Notably, in the studies of Bugg and McDaniel and Weinstein and colleagues, students received more guidance than in the present study, for instance in the form of specific prompts, a practice opportunity to generate test questions and answers, or feedback on their generated questions.

It is also imaginable that the effectiveness of generating test questions and answers depends on the type of learning outcomes that is assessed. The elaboration that may be evoked by generating multiple-choice questions and particularly the plausible alternative but incorrect answers may not help students to remember factual or conceptual information better. Instead, elaborating on the learning material might particularly be beneficial for students' understanding, which can be expected to be more conducive to answering higher-order questions that measure comprehension and transfer. Findings of Bugg and McDaniel (2012) also suggest that the benefits may depend on the type of knowledge that is being assessed. They only found benefits of generating and answering test questions on conceptual posttest items that required multiple sentences to answer and covered information that was not literally in the text but not on detail questions that could be answered with a single sentence. A related issue is that generating test questions might be a more fruitful strategy when texts consist of many related information elements that learners need to process simultaneously in working memory than when texts consist of mostly isolated facts.

A potential limitation of our study is that learners were presented with the same posttest twice. We chose for this design because it allowed us to test whether differences among conditions would emerge only immediately after learning, only after a delay, or both immediately and after a one week delay. Given the benefits of test-taking on long-term retention (Rowland, 2014), taking the immediate posttest may have slowed the rate of forgetting. However, it is unlikely that this would explain our results, as the possible memory benefit of taking the same posttest twice would be present in all conditions. Another limitation is that, because the questions students generated were rather low-level in that they mainly covered isolated facts or concepts from the text, we were unable to explore whether there was a relationship between the quality of the generated questions and answer possibilities and retention. An interesting avenue for future research would be to test whether, as has been shown with taking practice tests (cf. Little & Bjork, 2015), generating competitive incorrect answer possibilities would lead to better retention than generating noncompetitive incorrect answers. In the present study, such a test was not possible because the materials were not specifically designed to present competitive incorrect answer possibilities on the posttest. Another important direction for future research is to investigate the robustness of our findings by examining whether the results would replicate using different learning materials and different student populations. This is particularly important for the finding that restudy leads to better recall of learning materials than generating multiple-choice questions. Future research should not only replicate but also extend these findings, given that other studies in which learners generated different kinds of questions and time was not controlled did find positive effects of generating test questions. For instance, by comparing whether the same finding (i.e., restudy better than question generating) would apply to generating other types of questions when time is kept equal, or by determining whether the restudy benefit would disappear under self-paced conditions.

Our findings can be important for educational practice. Students in educational practice are often recommended by their teacher to generate test questions and answers in preparation for an exam. Yet it seems that there is little benefit to studying materials with the intention of generating (multiple-choice) questions relative to the intention of completing a test. Moreover, actually generating (multiple-choice) test questions is not a viable alternative to taking a practice test if the aim is long-term retention – at least not for untrained students who are provided little guidance. Findings from other studies (see Rosenshine et al., 1996; Wong, 1985) do suggest that having students generate questions might entice them to spend more time on the study material, which can be beneficial for test performance. However, students who only have limited time available might be better off engaging in restudy, or other generative strategies when no practice test is available, such as teaching on video (Fiorella & Mayer, 2016; Hoogerheide et al., 2016).

Overall, this study contributes to our understanding of the effects of generating test questions, an instructional strategy that is commonly used in educational practice yet has received surprisingly little attention in research. Generating multiple-choice questions appears to result in remembering less information of learning materials than restudy when time is kept equal. We hope that this perhaps counterintuitive result sparks new research questions and provides a contribution to uncovering under which conditions generating test items is and is not an effective instructional strategy.

Acknowledgment

The authors would like to thank dr. Pooja K. Agarwal for allowing us to use the experimental text from Agarwal, Karpicke, Kang, Roediger, and McDermott (2008).

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. http://dx.doi.org/10.1002/acp.1391.

Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology, 72*, 593–604. http://dx.doi.org/10.1037/0022-0663.72.5.593.

Bugg, J. M., & McDaniel, M. A. (2012). Selective benefits of question self-generation and answering for remembering expository text. *Journal of Educational Psychology, 104*, 922–931. http://dx.doi.org/10.1037/a0028661.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale: Lawrence Erlbaum Associates.

Fiorella, L., & Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology, 38*, 281–288. http://dx.doi.org/10.1016/j.cedpsych.2013.06.001.

Fiorella, L., & Mayer, R. E. (2014). Role of expectations and explanations in learning by teaching. *Contemporary Educational Psychology, 39*, 75–85. http://dx.doi.org/10.

1016/j.cedpsych.2014.01.001.
Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28*, 717–741. http://dx.doi.org/10.1007/s10648-015-9348-9.
García, F., García, Á., Berbén, A. B., Pichardo, M. C., & Justicia, F. (2014). The effects of question-generation training on metacognitive knowledge, self regulation and learning approaches in science. *Psicothema, 26*, 385–390. http://dx.doi.org/10.7334/psicothema2013.252.
Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem-solving. *Educational Psychologist, 45*, 1–14. http://dx.doi.org/10.1080/00461520903213618.
Hoogerheide, V., Deijkers, L., Loyens, S. M. M., Heiltjes, A., & Van Gog, T. (2016). Gaining from explaining: Learning improves from explaining to fictitious others on video, not from writing to them. *Contemporary Educational Psychology, 44*, 95–106. http://dx.doi.org/10.1016/j.cedpsych.2016.02.005.
Hoogerheide, V., Loyens, S. M. M., & Van Gog, T. (2014). Effects of creating video-based modeling examples on learning and transfer. *Learning and Instruction, 33*, 108–119. http://dx.doi.org/10.1016/j.learninstruc.2014.04.005.
King, A. (1992). Comparison of self-questioning, summarizing and notetaking-review as strategies for learning from lectures. *American Educational Research Journal, 29*, 303–323. http://dx.doi.org/10.3102/00028312029002303.
King, A. (1994). Autonomy and question asking: The role of personal control in guided student-generated questioning. *Learning and Individual Differences, 6*, 163–185. http://dx.doi.org/10.1016/1041-6080(94)90008-6.
Kirschner, F., Paas, F., & Kirschner, P. A. (2009). A cognitive load approach to collaborative learning: United brains for complex tasks. *Educational Psychology Review, 21*, 31–42. http://dx.doi.org/10.1007/s10648-008-9095-2.
Levine, T., & Hullett, C. (2002). Eta squared, partial eta squared and misreporting of effect size in communication research. *Human Communication Research, 28*, 612–625. http://dx.doi.org/10.1111/j.1468-2958.2002.tb00828.x.
Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition, 43*, 14–26. http://dx.doi.org/10.3758/s13421-014-0452-8.
Mamede, S., Van Gog, T., Moura, A. S., De Faria, R. M. D., Peixoto, J. M., Rikers, R. M. K. P., et al. (2012). Reflection as a strategy to foster medical students' acquisition of diagnostic competence. *Medical Education, 46*, 464–472. http://dx.doi.org/10.1111/j.1365-2923.2012.04217.x.
Mayer, R. E. (2003). *Learning and instruction.* Upper Saddle River, NJ: Merrill Prentice Hall.
Mayer, R. E. (2014). Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In R. E. Mayer (Ed.). *The Cambridge handbook of multimedia learning* (pp. 345–368). New York: Cambridge University Press.
Muis, K. R., Psaradellis, C., Chevrier, M., Di Leo, I., & Lajoie, S. P. (2016). Learning by preparing to teach: Fostering self-regulatory processes and achievement during complex mathematics problem solving. *Journal of Educational Psychology, 108*, 474–492. http://dx.doi.org/10.1037/edu0000071.
Nestojko, J. F., Bui, D. C., Kornell, N., & Bjork, E. L. (2014). Expecting to teach enhances learning and organization of knowledge in free recall of text passages. *Memory & Cognition, 42*, 1038–1048. http://dx.doi.org/10.3758/s13421-014-0416-z.
Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology, 84*, 429–434. http://dx.doi.org/10.1037/0022-0663.84.4.429.
Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 2*, 117–175.
Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*, 437–447. http://dx.doi.org/10.1016/j.jml.2009.01.004.
Reder, L. M., Charney, D. H., & Morgan, K. I. (1986). The role of elaborations in learning a skill from an instructional text. *Memory & Cognition, 14*, 64–78. http://dx.doi.org/10.3758/BF03209230.
Renkl, A. (1995). Learning for later teaching: An exploration of mediational links between teaching expectancy and learning results. *Learning and Instruction, 5*, 21–36. http://dx.doi.org/10.1016/0959-4752(94)00015-H.
Roediger, H. L., & 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., 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.
Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research, 66*, 181–221. http://dx.doi.org/10.3102/00346543066002181.
Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 160*, 1432–1463. http://dx.doi.org/10.1037/a0037559.
Stein, B. S., Littlefield, J., Bransford, J. D., & Persampieri, M. (1984). Elaboration and knowledge acquisition. *Memory & Cognition, 12*, 522–529. http://dx.doi.org/10.3758/BF03198315.
Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory.* New York: Springer.
Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*, 16–26. http://dx.doi.org/10.1080/00461520701756248.
Weinstein, Y., McDermott, K. B., & Roediger, H. L. (2010). A comparison of study strategies for passages: Re-reading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied, 16*, 308–316. http://dx.doi.org/10.1037/a0020992.
Wong, B. Y. L. (1985). Self-questioning instructional research: A review. *Review of Educational Research, 55*, 227–268. http://dx.doi.org/10.3102/00346543055002227.