Comparative effects of test-enhanced learning and self-explanation on long-term retention

Douglas P Larsen,[1] Andrew C Butler[2] & Henry L Roediger III[3]

CONTEXT Educators often encourage students to engage in active learning by generating explanations for the material being learned, a method called self-explanation. Studies have also demonstrated that repeated testing improves retention. However, no studies have directly compared the two learning methods.

METHODS Forty-seven Year 1 medical students completed the study. All students participated in a teaching session that covered four clinical topics and was followed by four weekly learning sessions. In the learning sessions, students were randomised to perform one of four learning activities for each topic: testing with self-generated explanations (TE); testing without explanations (T); studying a review sheet with self-generated explanations (SE), and studying a review sheet without explanations (S). Students repeated the same activity for each topic in all four sessions. Six months later, they took a free-recall clinical application test on all four topics.

RESULTS Repeated testing led to better long-term retention and application than repeatedly studying the material ($p < 0.0001$, $\eta^2 = 0.33$). Repeated generation of self-explanations also improved long-term retention and application, but the effect was smaller ($p < 0.0001$, $\eta^2 = 0.08$). When data were collapsed across topics, both testing conditions produced better final test performance than studying with self-explanation (TE = 40% > SE = 29% [$p = 0.001$, $d = 0.70$]; T = 36% > SE = 29% [$p = 0.02$, $d = 0.48$]). Studying with self-explanation led to better retention and application than studying without self-explanation (SE = 29% > S = 20%; $p = 0.001$, $d = 0.68$). Our analyses showed significant interaction by topic ($p = 0.001$, $\eta^2 = 0.06$), indicating some variation in the effectiveness of the interventions among topics.

CONCLUSIONS Testing and generating self-explanations are both learning activities that can be used to produce superior long-term retention and application of knowledge, but testing is generally more effective than self-explanation alone.

Medical Education 2013: ***47:*** *674–682*
doi: 10.1111/medu.12141

Discuss ideas arising from the article at www.meduedu.com 'discuss'

[1]Department of Neurology, Washington University in St. Louis School of Medicine, St. Louis, Missouri, USA
[2]Department of Pscyhology & Neuroscience, Duke Univeristy, Durham, North Carolina, USA
[3]Department of Psychology, Washington University in St. Louis, St. Louis, Missouri, USA

Correspondence: Dr Douglas Larsen, Department of Neurology, Washington University in St. Louis, 660 South Euclid Avenue, Campus Box 8111, St. Louis, Missouri 63110, USA.
Tel: 00 1 314 454 6120; E-mail: larsend@neuro.wustl.edu

INTRODUCTION

One of the primary goals of medical education is to help students acquire a large body of knowledge in the hope that they will retain it for long periods until they need to apply it in a future clinical setting. Despite the immense challenges presented by this goal, relatively little research has investigated the direct effects on long-term retention of particular educational interventions. One potentially beneficial intervention that has recently garnered attention is a method called 'test-enhanced learning'. Test-enhanced learning is a process that involves learners repeatedly retrieving information through tests (or other opportunities to practise retrieval), which results in superior long-term retention of that information.[1,2] In test-enhanced learning, tests are conceptualised as learning tools, rather than as assessment tools. Although testing may have indirect effects on student motivation and study habits, a large body of research has shown that the direct act of retrieving information from memory enhances learning and retention better than simply increasing the quantity or quality of study.[2] Research in cognitive psychology laboratories has shown that this method for learning is superior to repeated study of the same information.[2–4] Recent studies in actual educational settings have also demonstrated that test-enhanced learning can increase the retention of knowledge for extended intervals of 6–9 months.[5–7]

Despite the established benefits of test-enhanced learning, its implementation in educational practice raises two important questions. Does test-enhanced learning simply represent rote memorisation of facts with little ability to transfer or apply that knowledge to new settings? How well does test-enhanced learning compare with other methods of active learning? A major strand of research in educational psychology has focused on constructivism, which emphasises the construction by learners of their own personal systems of understanding.[8] In constructivism, the recall of facts is only important to the degree that it represents the framework of why information is important and how items of knowledge work together.[9] In this light, the testing of factual information is often seen as a misdirection of educational priorities.[8] In order to help learners accomplish the goal of generating their own systems of understanding, several elaborative techniques have been developed.

One technique that has shown promise involves having students generate explanations about why a particular piece of information is important and how it relates to their existing knowledge. In this method, often referred to as 'self-explanation', learners are thought to benefit from generating their own personal understanding of the to-be-learned information because this helps them to integrate the new information with existing knowledge.[10] Studies have shown that this technique can be beneficial over a wide variety of materials and learners. However, the durability of the benefits obtained by engaging in self-explanation is unclear because most studies have measured the effects after very brief intervals of time (e.g. anywhere from minutes to a couple of weeks).[11–14] Clearly, more research is needed to ascertain whether these benefits last over educationally meaningful intervals of time.

In the present study, we directly compared the efficacy of, respectively, testing and generating self-explanations in terms of promoting the long-term retention and application of clinical material. Importantly, the two techniques are not mutually exclusive and thus we also investigated the efficacy of combining testing and self-explanation. By manipulating the presence or absence of each technique as an independent variable, we were able to isolate the individual and combined contributions of these learning methods to long-term retention. In our design, the two independent variables were fully crossed to result in four different learning activities: testing with self-generated explanations (TE); testing without explanations (T); studying a review sheet with self-generated explanations (SE), and studying a review sheet without explanations (S). Students were given an initial teaching session that introduced four neurology topics, after which they engaged in four testing/study sessions at weekly intervals in which they performed one of the four learning activities for each topic. Long-term retention and transfer of knowledge were assessed in a free-recall clinical application test given 6 months after the initial teaching session.

METHODS

Participants

Forty-nine Year 1 medical students were recruited to participate in the study. Year 1 medical students were selected because they had no prior exposure to the topics taught in the study, which ensured a more uniform level of baseline knowledge across

the cohort. All students completed the initial testing or study sessions; however, two students did not complete the final test and therefore their data were excluded from the statistical analyses. Participation was voluntary; all students provided consent. All activities related to the study were performed outside standard class time. The study was conducted independently of any course taught in the medical school. Students were reimbursed for the time they spent participating in the study at a rate of approximately US$15/hour. Students spent a total of about 9 hours in study-related activities. The study was approved by the institution's Human Research Protection Office.

Materials

Materials were developed to cover four clinical neurology topics: seizures; optic neuritis; myasthenia gravis, and migraine. The materials covered the information a student would be expected to draw upon in interviewing, diagnosing and treating patients with these conditions. Four types of learning materials were developed for each topic: (i) a short-answer written test (T); (ii) a written test that also required students to write out their explanations of their answers (TE); (iii) a review sheet that restated all of the information covered (S, for study), and (iv) a review sheet that contained space for students to write out their explanations for the statements on the review sheet (SE). The content covered in each learning activity was identical. The content required students to learn 26–30 items for each topic. An essay application test was developed to assess long-term retention and transfer. This test consisted of a short clinical scenario: students were given a patient's demographic information and chief complaint, and were then asked to use the information they had learned about the topic to outline their approach to managing the patient's condition. They were told to include all of the information that had been covered in the initial teaching session and further practised through the various learning activities. The test sheet presented the brief clinical scenario at the top of the page and the remainder was left blank for students to write out their responses.

Questionnaires were developed to ascertain from students how the various activities influenced their learning. These were administered both after the initial learning sessions were completed and after the final test at the end of the study. The questionnaires used open-ended questions in order to allow students to fully describe their thoughts and observations. For example, students were asked: 'Describe how the review sheet helped (or did not help) you to learn' and 'Describe how the explanations on the tests or review sheets helped (or did not help) you to learn.'

Procedures

Students participated in two teaching sessions held on consecutive days. Two topics were covered on each day; 1 hour was dedicated to each topic. The topics were taught in an interactive didactic format. At the end of each teaching session, each student was randomised in a counterbalanced fashion to engage in one of the four learning activities (i.e. T, TE, S or SE) for each of the topics covered in that session. In this way, each student performed all four learning activities, but each activity was paired with a different topic. For example, Student 1 may have been randomised to testing alone on seizures, testing with self-explanations on optic neuritis, studying the review sheet alone on myasthenia gravis, and studying with self-explanations on migraine. However, Student 2 may have been randomised to testing alone on optic neuritis, testing with self-explanations on myasthenia gravis, studying the review sheet alone on migraine, and studying the review sheet with self-explanations on seizures. The counterbalancing ensured that the pairing of topic and learning activity occurred equally often across the student sample.

Once a week for the next 3 weeks students returned for additional test/study sessions in which they performed the same learning activities they had done previously for each of the four topics. The students' pairings of topics and learning activities did not change between test/study sessions (i.e. Student 1 would perform the same learning activities for each topic as in the initial test/study session and Student 2 would perform the same learning activities for his or her assigned topics as in the initial learning/study session). Including the test/study session immediately after the teaching session, students completed four test/study sessions in total. In this way, each student gained equal practice for each learning activity over four repetitions. No time limits were placed on any of the activities, but generally students completed all four activities in about an hour per test/study session.

During each test/study session, once students had completed the tests and studied the review sheets, they were asked to score their tests and compare their explanations (for those activities requiring

explanations) with an answer key. Explanations were scored only according to their completion because the purpose of generating self-explanations was to have students use their own unique explanations, which they believed would help them to understand and retain the information. The example explanations on the answer key were provided to help them if they were unable to come up with an explanation on their own. Students were also asked to rate their overall effort in providing the explanations. After the fourth test/study session, students were asked to fill out a questionnaire describing their perceptions of their learning.

Students were explicitly instructed not to quiz themselves with the review sheets in order to avoid confounding the study activities with self-testing. They were allowed to read and re-read the review sheets as many times as they felt necessary to learn the material. Students were also explicitly instructed not to study the material outside the test/study sessions. They were also told not to review the material with one another.

About 6 months after the initial teaching sessions, students returned to take an application test that covered all four topics. In the application test, students were given a clinical scenario and asked to use all of the information they had learned on the appropriate topic to describe how they would deal with the scenario. After taking the final test, they filled out a questionnaire. Responses to the interim and final questionnaires were compiled and analysed.

Scoring and statistical analyses

Two raters scored each final test in a blinded fashion. Because students were asked to include all information from the initial teaching session in their free-recall tests, raters used the answer sheets from the initial tests as checklists with which to score the presence or absence of the required information in the students' final tests. Kappa (κ) statistics were calculated to measure inter-rater reliability in the scoring of the final tests. Inter-rater reliability for the application test was excellent ($\kappa = 0.93$) and thus the initial tests were scored by a single rater. As part of our analysis of the initial tests, we also directly measured students' compliance with writing explanations in the SE and TE conditions by scoring the presence or absence of explanations for items in each condition. For the TE condition, we scored only items that students answered correctly because it is difficult to interpret the reason for the absence of an explanation for an item that was incorrect; students could have failed to provide an explanation for an incorrect item because they failed to retrieve that item (i.e. the response was omitted), because they did not know the material well enough to generate an explanation or because they simply chose not to write an explanation. Students were also asked to rate their effort in generating explanations on the review sheets and tests in each of the four learning sessions using a 10-point scale on which a score of 10 represented maximum effort ('Full effort, I did my best to make each explanation useful for me to remember') and a score of 1 represented minimal effort ('No effort, I only wrote explanations to fill the space').

All analyses were performed using SPSS Version 18 (SPSS, Inc., Chicago, IL, USA) and Microsoft Excel for Mac. The results for the final application test were analysed using a repeated-measures analysis of variance (ANOVA) that included testing (i.e. test versus re-study), self-explanation (i.e. explanation versus no explanation), and neurological topic as independent variables. The proportion of correct responses on the final tests, a measure of long-term retention and application, served as the dependent variable. Planned post hoc *t*-test comparisons were made between the various learning activities. Eta-squared (η^2) and Cohen's *d* are the effect sizes reported for the ANOVA and *t*-test analyses, respectively.

RESULTS

Initial learning

Performance on the initial tests showed that the students had good retention of the material immediately after the teaching sessions. As expected, because no manipulation had yet been implemented, students showed roughly equivalent performance on Test 1 for the TE (0.74) and T (0.72) activities. After an initial decline in performance from Test 1 to Test 2, performance steadily improved in both testing conditions from Test 2 until Test 4 (Table 1). However, the rate of improvement was greater when self-explanation was included with testing. On Test 4, which occurred 3 weeks after the initial teaching sessions, the TE activity (0.86) resulted in better performance than the T activity (0.78). A 4 (test number) $\times$ 2 (testing condition) ANOVA on the initial learning phase performance confirmed these observations by revealing a significant main effect of test number ($F_{3,138} = 69.91$, mean squared error [MSE] = 0.01, $p < 0.0001$, $\eta^2 = 0.26$), as well as a significant interaction ($F_{3,138} = 3.91$, MSE = 0.01, $p = 0.005$, $\eta^2 = 0.01$), indicating that the rate of change dif-

Table 1 Mean proportions of correct answers on the initial learning tests in the testing with self-explanation (TE) and testing without self-explanation (T) conditions

Learning activity	Test 1	Test 2	Test 3	Test 4
TE	0.74	0.63	0.77	0.86
T	0.72	0.61	0.68	0.78

fered significantly between conditions. The main effect of testing condition was marginally significant ($F_{1,46}$ = 3.91, MSE = 0.06, p = 0.054, η^2 = 0.03). Overall, these findings indicate that performance at the start of the two testing conditions was similar, but the testing with self-explanation (TE) activity produced a slightly greater increase in performance across the four initial tests relative to the testing without self-explanation (T) activity.

In addition to test performance, a few other measures of student performance during initial learning were collected. After each study activity, students were asked how many times they had read through the review sheets. On average, they reported having read through the review sheet 1.7 times per session for the study with explanations (SE) activity and 2.0 times per session for the study without explanations (S) activity (p = 0.02, d = 0.29). In our analysis of student compliance with the request to write self-explanations, students were found to have produced

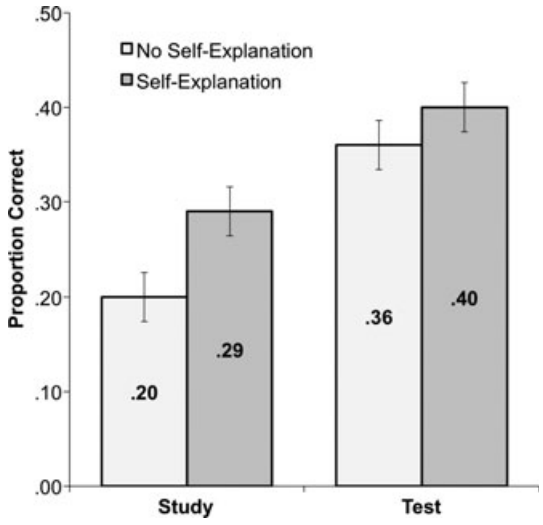

Figure 1 Mean proportions of correct answers on the final application test as a function of testing (test versus study) and self-explanation (self-explanation versus no self-explanation). Error bars represent 95% confidence intervals

an explanation for 96% of the items in the study with self-explanation condition; however, they generated an explanation for only 71% of the items they successfully retrieved in the test with self-explanation condition (p < 0.0001, d = 0.96). Finally, students were asked to rate on a 10-point scale their effort in generating explanations on the review sheets and tests; no significant differences were identified in ratings of effort. When students' reports of effort were collapsed across the four testing/study sessions, mean scores of 6.7 and 6.4 in the study with self-explanation condition and test with self-explanation condition, respectively (p = 0.11) were identified.

Final test of retention and application

The final test was administered approximately 6 months after the initial teaching sessions. Figure 1 depicts the proportion of correct responses on the final free-recall application test. Overall, the TE activity produced the best performance followed, respectively, by the T, SE and S activities. A 2 (testing) × 2 (self-explanation) × 4 (topic) repeated-measures ANOVA showed a significant main effect of testing ($F_{1,43}$ = 122.59, MSE = 0.01, p < 0.0001, η^2 = 0.33) and a significant main effect of self-explanation ($F_{1,43}$ = 20.90, MSE = 0.01, p < 0.0001, η^2 = 0.08). Although both testing and generating self-explanations increased long-term retention and application, it is important to note that the magnitude of the effect size for testing is much larger. In addition to the two main effects, there was a significant interaction between the learning conditions ($F_{1,43}$ = 4.12, MSE = 0.01, p = 0.049, η^2 = 0.01), indicating that generating self-explanations helped more when combined with studying a review sheet (i.e. SE versus S) than when used with testing (i.e. TE versus T). Follow-up pairwise comparisons showed that both testing conditions produced significantly better retention and application than studying with self-explanation (TE = 0.40 > SE = 0.29 [p = 0.001, d = 0.70]; T = 0.36 > SE = 0.29 [p = 0.02, d = 0.48]). However, there was no significant difference between the TE and T activities (p = 0.077, d = 0.28). The SE activity produced significantly better performance than the S activity (SE = 0.29 > S = 0.20; p = 0.001, d = 0.68).

The primary findings were qualified by interactions between the learning activities and the neurological topics. The repeated-measures ANOVA revealed a significant interaction between testing and topic ($F_{3,43}$ = 4.03, MSE = 0.01, p = 0.013, η^2 = 0.03), as well as between self-explanation and topic

($F_{3,43}$ = 4.31, MSE = 0.01, p = 0.01, η^2 = 0.05). In addition, there was significant three-way interaction among testing, self-explanation and topic ($F_{1.43}$ = 6.45, MSE = 0.01, p = 0.001, η^2 = 0.06). Overall, these interactions indicate some variation in the relative effectiveness of various learning activities among the topics. Table 2 shows the final test performance by topic. The variation among topics appears to be driven in part by differences in the effectiveness of generating self-explanations. Recall in the SE activity condition was equivalent to recall in the two testing activity conditions for some topics, but not for others. Similarly, the TE and T activities led to equivalent performance on some topics, but not others. Despite this variation by topic, one key result remained consistent across topics: both testing activities yielded superior performance relative to studying without self-explanation.

Questionnaire responses

After completing the four testing/study sessions, students were asked to describe how taking the tests and studying the review sheets had or had not helped in their learning. They were also asked to explain how writing out the explanations had affected their learning. Students overwhelmingly felt that repeated testing was beneficial: 90% of students reported testing to be helpful in their learning. Conversely, 42% of students reported that studying the review sheet was helpful to varying degrees. A total of 67% of students reported that writing explanations was helpful in some form and 50% commented that the explanations helped them to understand why material was important, to make connections between the study material and other facts, or to group information into overarching themes. However, 31% of students regarded the explanations as redundant or unnecessary. Many of these students commented that the ability to answer the question on a test indicated that they had sufficient knowledge of the material and giving an explanation did not add significantly to their understanding.

At the end of the study, the students completed a second questionnaire asking whether they had reviewed material outside of the study and whether they were willing to take repeated tests as part of future courses. None of the students indicated that they had studied the material outside the testing or study sessions, but a few of them mentioned that some of the topics had arisen in their regular medical school courses. Despite the fact that students were instructed not to quiz themselves with the review sheets, 17% reported that they had engaged in some form of self-quizzing with the review sheet. Overall, 60% of students stated that they would be willing to take repeated tests as part of future courses. Interestingly, 11% of the students spontaneously commented that they would only want to take repeated tests if the tests were to be used for learning purposes only and were not to contribute to their grades.

DISCUSSION

Our experiment demonstrates that both repeated testing and the repeated generation of self-explanations produced superior long-term retention and transfer of knowledge on a free-recall clinical application test compared with repeated studying without self-explanation. Our findings have both theoretical and practical implications.

One goal of the present study was to compare the relative efficacies of retrieval practice (i.e. testing)

Table 2 Mean proportions of correct answers on the final application test as a function of initial learning activity and topic in the testing with self-explanation (TE), testing without self-explanation (T), studying with self-explanation (SE) and studying without self-explanation (S) conditions

	Topic				
Learning activity	**Seizures**	**Migraine**	**Myasthenia**	**Optic neuritis**	**Grand mean**
TE	0.40	0.39	0.49	0.32	0.40
T	0.44	0.32	0.36	0.33	0.36
SE	0.39	0.32	0.21	0.25	0.29
S	0.24	0.26	0.19	0.12	0.20

and self-explanation as methods for facilitating learning. Although both interventions significantly improved long-term retention and transfer, their effects in isolation were not equal in magnitude. Testing had a much larger effect on final test performance ($\eta^2 = 0.33$) relative to self-explanation ($\eta^2 = 0.08$). In planned pairwise comparisons, both testing with self-explanation and testing without self-explanation produced moderate to large effects compared with repeated study with self-explanation ($d = 0.70$ and $d = 0.48$, respectively). Thus, our study suggests that repeated testing produces more robust long-term retention and transfer of knowledge relative to repeated self-explanation. This finding is concordant with the results of other research in which retrieval practice has been compared with other active learning techniques. For example, Karpicke and Blunt found that repeated testing over a 1-week interval resulted in superior retention compared with concept mapping, an active study technique that is considered by many in education to be an excellent method of integrative learning.[15]

It should be noted that 17% of students reported that they had engaged in some form of self-quizzing while studying the review sheets despite explicit instructions not to do so. Because of the uncontrolled nature of this activity and the relatively small percentage of students who had engaged in it, it is difficult to measure its effect on final test performance. However, given the benefits of testing observed in the present study and others, it would be expected that the presence of some self-testing would increase final test performance in the study condition, thereby reducing the differences observed among the four learning activities. Thus, the presence of self-quizzing in the study condition makes it all the more remarkable that repeated testing produced such a robust effect on retention.

The comparison between testing and self-explanation yields some insights into why they may have different effects on retention. Completing a test may produce better retention because it is an inherently more difficult task than generating explanations for information. Research has shown that the difficulty inherent in encoding and retrieving knowledge leads to more durable learning.[16,17] The work involved in the actual act of retrieval has a direct effect on the learner's ability to recall that information later. Agrawal *et al.* have suggested that testing serves as a self-monitoring intervention in which learners track and adjust their strategies of knowing and understanding information based on their ability to retrieve that knowledge.[18] Self-explanation may also be thought of as a self-monitoring intervention as students identify their own systems of knowing information. However, our results would indicate that testing might lead to more robust retention because it is more effective than self-explanation at causing students to develop successful systems of organising and retrieving information because it forces them to more fully assess the efficacy of their current systems when they are confronted with a question. The findings of Pyc and Rawson would support this conclusion.[19] These authors demonstrated that testing does lead to a change in knowledge organisation compared with simply re-studying when testing is coupled with feedback.[19] This effect is caused by students' reassessing of their structures of knowledge and discarding of ineffective or inaccurate systems for more successful structures. Zaromb and Roediger also showed that the retrieval effort in free-recall testing improves a learner's mental organisation of information more effectively than does study of the same information.[20]

Interestingly, combining testing with self-explanation significantly improved performance during the initial learning phase relative to testing without explanation. Although this significant advantage did not endure to the final assessment 5 months later, there was still a small numerical difference in favour of testing with self-explanation. One reason why this additional benefit may not have been as durable as might have been hoped is that the rate of generating explanations in the TE activity is lower (71% of correct items versus 96% in the SE activity). Alternatively, the lower rate of generating explanations may have been driven by the students' perception that the explanations were unnecessary on the tests and were redundant. Many students commented that their answering of questions on the test indicated that they already knew the information from the explanation and had no need to make it explicit. The benefit of adding self-explanation to testing may depend on the type of information being learned. Understanding how to identify topics that most benefit from the combination of self-explanation and testing may be a fruitful direction for future research.

Although our findings suggest that self-explanation is not as robust as testing, it is important to note that generating explanations had a significant effect on long-term retention and transfer of knowledge. The finding that studying with self-explanation produced superior performance on the final test relative to studying without self-explanation ($d = 0.68$) is important because it demonstrates the durability of the benefits of engaging in self-explanation dur-

ing learning on a test given 6 months after initial teaching. Previous studies on self-explanation have only assessed performance after relatively short retention intervals ranging from a few minutes to a few weeks.[11–14] The present study extends this finding to show that generating self-explanations increases long-term retention over an educationally meaningful retention interval of 6 months from initial learning.

Another goal of the present study was to investigate whether test-enhanced learning would support the transfer of knowledge to new settings (i.e. relative to simply promoting the rote memorisation of facts). To this end, we used a final assessment that required students to apply their knowledge to a clinical scenario in which they had to outline an approach to determining a hypothetical patient's condition but were given only the patient's demographic data and information on his or her chief complaint. Our findings show that repeated retrieval practice during learning significantly improved students' ability to apply their knowledge in this new setting. Although the overall level of performance was lower than that hoped for in medical education, it is important to consider the nature of the final application test. Six months after learning the material, students were asked to retrieve and apply as much of their knowledge as possible (from 26 to 30 items) with very little prompting or structure; this is similar to the requirements of an essay test. Given that the students had minimal exposure to the material in the intervening months, and in view of the difficulty of a final essay test, we think it is impressive that repeated testing produced 36–40% correct performance (almost double the amount derived by studying without self-explanations).

Our finding that repeated testing promotes better long-term retention and transfer fits well with the findings of recent studies which have also reported that retrieval practice improves learners' application of knowledge to new situations.[21–25] Whereas many studies show that testing improves the retention of information that is retrieved from memory, these new studies suggest that repeated retrieval practice may also improve understanding of that information. By contrast, generating self-explanations may improve understanding, but may not be as powerful a tool for promoting long-term retention.

Our study does have some limitations. Our study design does not fully distinguish the effect of writing self-explanations from that of reading an answer sheet because sample explanations were provided on the answer sheet. The sample explanations were given in an effort to avoid having to penalise students who might have forgotten so much information that they could not generate an explanation or who were unsure how to write an explanation. It should be noted that even with this additional help, testing had a more robust effect than making the explanations available. Future studies of self-explanations will need to provide additional comparisons to further tease out the effects of providing sample explanations.

In terms of other limitations, we also observed an interaction indicating some variation in the effectiveness of the learning techniques among topics. In interpreting this result, it is important to consider that our materials consisted of four heterogeneous topics. Studies in laboratory settings typically use carefully constructed materials, such as word pairs and short texts, which are matched on many different dimensions. Our materials are more representative of the uncontrolled variety of information found in a typical medical school course. Although we were careful to have each topic cover roughly the same amount of information, the nature of the information in each topic is inherently different. Some types of information may be learned through self-explanation just as well as through testing. Similarly, self-explanation may be more effective when learners have significant prior knowledge to help generate an explanation or when students have received explicit training in generating explanations.[10,14,26] Our study showed that repeated testing with and without self-explanation led to performance equivalent to that derived by studying with self-explanation on certain topics, but to superior performance for others. However, self-explanation never produced significantly better retention and transfer of knowledge than testing. Thus, our findings suggest that for educators who must choose between the two learning activities, repeated testing appears to be the more robust and reliable technique.

CONCLUSIONS

Our study has important implications for educators. We demonstrated that repeated testing and self-explanation both lead to superior long-term retention and transfer compared with repeated studying of the same information. Although for some topics these two learning interventions produce equivalent performance, our findings show that the act of repeated retrieval generally produces superior reten-

tion than the act of self-explanation. Educators should seek to incorporate both of these powerful learning techniques into their curricula in order to enhance long-term retention and the application of learned material to clinical settings.

Contributors: DPL developed the study concept, contributed to its design, conducted the study, collected and analysed the data and drafted the paper. ACB contributed to the study design and the analysis of data. HLR contributed to the study design. All authors contributed to the critical revision of the paper and approved the final manuscript for publication.
Acknowledgements: we would like to thank Jessye Brick and Laura Najjar for their assistance in scoring tests and compiling data.
Funding: this study was funded by an Education Research Grant from the American Academy of Neurology and by the McDonnell Center for Systems Neuroscience at the Washington University in St. Louis School of Medicine.
Conflicts of interest: none.
Ethical approval: this study was approved by the Washington University in St Louis School of Medicine's Human Research Protection Office.

REFERENCES

1 Larsen DP, Butler AC, Roediger HL III. Test-enhanced learning in medical education. *Med Educ* 2008;**42**:959–66.
2 Roediger HL III, Karpicke JD. The power of testing memory: basic research and implications for educational practice. *Perspect Psychol Sci* 2006;**1**:181–210.
3 Karpicke JD, Roediger HL III. The critical importance of retrieval for learning. *Science* 2008;**319** (5865):966–8.
4 Roediger HL III, Karpicke JD. Test-enhanced learning: taking memory improves long-term retention. *Psychol Sci* 2006;**17**:249–55.
5 Larsen DP, Butler AC, Roediger HL III. Repeated testing improves long-term retention relative to repeated study: a randomised, controlled trial. *Med Educ* 2009;**43**:1174–81.
6 McDaniel MA, Agarwal PK, Huelser BJ, McDermott KB, Roediger HL III. Test-enhanced learning in a middle school science classroom: the effects of quiz frequency and placement. *J Educ Psychol* 2011;**103**:399–414.
7 Carpenter SK, Pashler H, Cepeda NJ. Using tests to enhance 8th grade students' retention of US history facts. *Appl Cogn Psychol* 2009;**23**:760–71.
8 Duffy TM, Cunningham DJ. Constructivism: implications for the design and delivery of instruction. In: Jonassen DJ, ed. *Handbook of Research for Educational Communication and Technology*. New York, NY: McMillan 1996;170–98.
9 Biggs J. Enhancing teaching through constructive alignment. *High Educ* 1996;**32**:347–64.
10 Dunlosky J, Rawson KA, Marsh EJ, Nathan MJ, Willingham DT. Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol Sci Public Interest* 2013;**14**:4–58.
11 Berry DC. Metacognitive experience and transfer of logical reasoning. *Q J Exp Psychol* 1983;**35A**:39–49.
12 Chi MTH, de Leeuw N, Chiu M-H, LaVancher C. Eliciting self-explanations improves understanding. *Cogn Sci* 1994;**18**:439–77.
13 Rittle-Johnson B. Promoting transfer: effects of self-explanation and direct instruction. *Child Dev* 2006;**77**:1–15.
14 Wong RMF, Lawson MJ, Keeves J. The effects of self-explanation training on students' problem solving in high-school mathematics. *Learn Instr* 2002;**12**:233–62.
15 Karpicke JD, Blunt JR. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science* 2011;**331**:772–5.
16 Pyc , MA , Rawson KA. Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? *J Mem Lang* 2009;**60**:437–47.
17 Bjork RA. Memory and metamemory considerations in the training of human beings. In: Metcalfe J, Shimamura A, eds. *Metacognition: Knowing about Knowing*. Cambridge, MA: MIT Press 1994;185–205.
18 Agrawal S, Norman GR, Eva KW. Influences on medical students' self-regulated learning after test completion. *Med Educ* 2012;**46**:326–35.
19 Pyc MA, Rawson KA. Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *J Exp Psychol Learn Mem Cogn* 2012;**38**:737–46.
20 Zaromb FM, Roediger HL III. The testing effect in free recall is associated with enhanced organisation processes. *Mem Cognit* 2010;**38**:995–1008.
21 Butler AC. Repeated testing produces superior transfer of learning relative to repeated studying. *J Exp Psychol Learn Mem Cogn* 2010;**36**:1118–33.
22 Johnson CI, Mayer RE. A testing effect with multimedia learning. *J Educ Psychol* 2009;**101**:621–9.
23 Rohrer D, Taylor K, Sholar B. Tests enhance the transfer of learning. *J Exp Psychol Learn Mem Cogn* 2010;**36**:233–9.
24 Kang SHK, McDaniel MA, Pashler H. Effects of testing on learning of functions. *Psychon Bull Rev* 2011;**18**:998–1005.
25 Larsen DP, Butler AC, Roediger HL III. The importance of seeing the patient: test-enhanced learning with standardised patients and written tests improves clinical application of knowledge. *Adv Health Sci Educ* 2012. doi: 10.1007/s10459-012-9379-7 (Epub ahead of print).
26 Woloshyn VE, Pressley M, Schneider W. Elaborative-interrogation and prior knowledge effects on learning of facts. *J Educ Psychol* 1992;**84**:115–24.

Received 11 June 2012; editorial comments to author 4 September 2012, accepted for publication 18 December 2012