Adv in Health Sci Educ
DOI 10.1007/s10459-012-9379-7

The importance of seeing the patient: test-enhanced learning with standardized patients and written tests improves clinical application of knowledge

Douglas P. Larsen · Andrew C. Butler · Amy L. Lawson · Henry L. Roediger III

Received: 20 January 2012 / Accepted: 11 May 2012

Abstract Previous research has shown that repeated retrieval with written tests produces superior long-term retention compared to repeated study. However, the degree to which this increased retention transfers to clinical application has not been investigated. In addition, increased retention obtained through written testing has not been compared to other forms of testing, such as simulation testing with a standardized patient (SP). In our study, 41 medical students learned three clinical topics through three different learning activities: testing with SPs, testing using written tests, and studying a review sheet. Students were randomized in a counter-balanced fashion to engage in one learning activity per topic. They participated in four weekly testing/studying sessions to learn the material, engaging in the same activity for a given topic in each session. Six months after initial learning, they returned to take an SP test on each topic, followed by a written test on each topic 1 week later. On both forms of final testing, we found that learning through SP testing and written testing generally produced superior long-term retention compared to studying a review sheet. SP testing led to significantly better performance on the final SP test relative to written testing, but there was no significant difference between the two testing conditions on the final written test. Overall, our study shows that repeated retrieval practice with both SPs and written testing enhances long-term retention and transfer of knowledge to a simulated clinical application.

D. P. Larsen (✉)
Department of Neurology, Washington University School of Medicine, 660 South Euclid Avenue, Campus Box 8111, St. Louis, MO 63110, USA
e-mail: larsend@neuro.wustl.edu

A. C. Butler
Department of Psychology and Neuroscience, Duke University, Durham, NC, USA

A. L. Lawson
Department of Pediatrics, Washington University School of Medicine, St. Louis, MO, USA

H. L. Roediger III
Department of Psychology, Washington University, St. Louis, MO, USA

Published online: 23 May 2012

Keywords Long-term retention · Simulation · Standardized patients · Tests · Test-enhanced learning

Introduction

Students, residents, and practicing physicians are required to learn a large body of information that they will need to use when treating patients months and even years later. Much research in medical education has focused on developing effective methods of instruction and creating assessment tools that are valid and reliable. Despite these efforts, less attention has been devoted to identifying educational interventions and strategies that can bridge the separation between initial learning and subsequent assessment and application. Many times information is forgotten until it is needed in the future, and then clinicians pursue efforts to re-learn the information. Techniques that improve clinicians' long-term retention of information and their ability to apply that information would greatly improve the efficiency and effectiveness of medical education; it would also save valuable time and resources by reducing the need for future re-learning. Both simulation and written tests have been used to improve assessment in medical education. However, the theoretical framework of test-enhanced learning suggests that both simulation and written testing can be used as learning tools to increase retention and transfer of knowledge to clinical settings.

Test-enhanced learning is an educational method developed in cognitive psychology to increase long-term retention (see Larsen et al. 2008; Roediger and Butler 2011; Roediger and Karpicke 2006a). Many educators consider tests to be assessment tools that influence learning by motivating students to study harder and providing feedback about their current state of knowledge that can be used to focus future learning activities (i.e. formative assessment). In contrast, a substantial body of research has developed in cognitive psychology demonstrating that repeated practice retrieving information from memory (typically in the form of tests) produces better long-term retention than repeated study of the material (e.g., Butler and Roediger 2007; Karpicke and Roediger 2008; Roediger and Karpicke 2006b). This finding, called the testing effect, is highly robust, replicable, and generalizes across materials and populations of learners. The testing effect even occurs when retrieval practice is compared to highly effective study activities such as concept mapping (Karpicke and Blunt 2011). In addition, there is growing evidence to suggest that practicing retrieval can improve people's understanding of the material, thereby facilitating subsequent application of that information to new contexts (e.g., Butler 2010).

Over the past few years, there has been an increased focus on investigating whether the benefits of test-enhanced learning can be applied in various educational settings outside of the cognitive psychology laboratory. For example, studies have shown that test-enhanced learning improves retention in middle school history and science classes (Carpenter et al. 2009; McDaniel et al. 2011). In the realm of medical education, our group demonstrated that having resident physicians repeatedly take written tests improved their retention of information taught in didactic conferences after a six-month interval (Larsen et al. 2009).

Almost all of the studies on test-enhanced learning have used written tests to improve retention, which prompts two critical questions. First, does the increased retention of knowledge produced by taking written tests transfer to real-life application with patient care? Test-enhanced learning represents a effective tool for promoting long-term retention on subsequent written tests, but a primary goal of medical education is the acquisition of

knowledge that will transfer to clinical settings. Second, how effective is retrieval practice when it is conceptualized in other forms besides a written test? Simulation provides a potentially powerful opportunity to practice retrieving and using knowledge in a way that may transfer more easily to actual patient care than written testing. Indeed, Kromann et al. (2010) used a simulation test to provide medical students with practice retrieving resuscitation techniques that they had just learned. Their study demonstrated that a single simulation test using mannequin simulators at the end of a resuscitation training class improved medical students' retention of skills over a six-month interval.

The concept of repeated retrieval practice nicely complements the principle of deliberate practice that has emerged in the simulation literature. McGaghie (2008) identified some of the key features of deliberate practice to be repeated practice with outcome measurements that yield informative feedback and that practice should continue until mastery is reached. In their systematic review of the simulation literature in medical education, the Best Evidence Medical Education (BEME) group reported that 39 % of the 108 studies that were reviewed found repetitive practice to be a significant element of effective high-fidelity simulation programs. The BEME group summarized their findings by stating that repetitive practice was a "primary factor in studies showing skills transferring to real patients" (Issenberg et al. 2005). A recent meta-analysis showed that simulation programs that use deliberate practice were consistently superior to traditional medical education practices (McGaghie et al. 2011). Despite these conclusions, repeated practice is not always used in simulation programs. Price et al. (2010) found that 15 out of the 16 anesthesia residency programs in Canada use high-fidelity simulation to train residents; however, 81 % of residents reported not being allowed to repeat the simulation scenarios.

Almost all of the research on deliberate practice has focused on simulation with mannequin simulation programs. However, standardized patients (SPs) offer another important form of simulation in medical education. Standardized patients are actors who use a pre-established script to portray the symptoms of a condition in a clinical setting. SPs provide the means to simulate many common clinical interactions and tasks. However, to our knowledge, simulation with SPs has not been studied with regard to the benefits of deliberate practice. Despite this lack of research, a few studies have found that even single exposures to SPs can improve retention of knowledge (Fallucco et al. 2010; Feddock et al. 2009; Stevens et al. 2009; Safdieh et al. 2011). However, no prior studies have investigated how repeated retrieval practice with SPs might benefit long-term retention and application of knowledge. In addition, no studies have directly compared retrieval practice with SPs, or any other form of simulation, to retrieval practice through written testing.

Our study represents a convergence of test-enhanced learning using written tests and deliberate practice in simulation and addresses gaps in both lines of investigation. First, how does repeated retrieval practice with standardized patients affect long-term knowledge retention compared with repeated retrieval practice with written tests and repeated studying? Second, how well does knowledge gained from repeated written testing transfer to clinical application as approximated by simulation with standardized patients? In order to answer these questions, we designed a study in which medical students learned three clinical topics over a month-long period through repeated testing with SPs, repeated testing with written tests, or repeated study of a review sheet. After 6 months, retention and application of their knowledge was evaluated through a final SP test and then a final written test.

Methods

Ethical approval

The Washington University School of Medicine Institutional Review Board approved the study. All students and standardized patients provided written informed consent.

Subjects

Forty-one first-year medical students participated in the study and each student received $120 in compensation. All of the students completed the study; however, two students missed one of the four initial learning sessions. During the initial learning phase of the study, the medical students were finishing their first year in medical school. The final assessment phase occurred 6 months after the initiation of the study, which fell within a few months of the beginning of the students' second year of medical school. Participation in the study occurred outside of standard class time.

Materials

The materials for the study were based on three clinical neurology topics: migraine, seizures, and myasthenia gravis. We determined the critical information necessary to perform basic clinical tasks regarding these topics: taking a topic-specific history, performing a targeted physical exam in the case of myasthenia gravis, and counseling the patient regarding the diagnosis and treatment. We constructed three alternate learning activities to cover each topic based on this content: written tests, SP scripts and checklists, and review sheets. The critical information covered in these three learning activities was identical.

Written tests used an open-ended, short-answer format (e.g., "*What symptoms would be indicative of a migraine with aura?"*). Each test asked students to retrieve 27–29 critical pieces of information, depending on the topic. Some questions required more than one piece of information, but each piece of information was scored individually (e.g., each of the symptoms that could be part of a migraine with aura would each be counted as one piece of information). All tests were identical for each administration.

The SP tests covered the same information as the written tests (i.e. 27–29 pieces of information per topic). During an encounter, the student needed to ask the SP about each piece of information (e.g., whether or not the patient was experiencing each of symptoms associated with migraine with aura). The SP followed a pre-determined script in answering the questions posed by the student. Although students were required to ask questions about the same information in each SP encounter (i.e. for a given topic), the demographic and historical details of each SP script were distinct so that students did not see the same patient twice (however, problems with SP rotations lead to six students seeing the same SP on two of their four SP tests). The SP encounters were videotaped so that independent coders could score student performance.

Review sheets covered the same information as the written and SP tests (e.g., a list of symptoms associated with migraine with aura). However, instead of requiring students to retrieve the information, the review sheets presented it in the form of factual statements so that they could study it. Students were instructed to study the review sheet however they would typically study that type of material.

During the final assessment phase, all of the students took an SP test and a written test for the three topics. The SP checklists used for the final assessment were the same those used during the initial learning phase. However, the demographic data and history for the final SP test was distinct from the scripts used in the initial SP tests so that all students would see a new clinical situation regardless of the type of initial learning activity that they had used in the first phase of the study. In this way we could approximate the experience of students applying the information in a clinical setting months after initial learning. The final written test was the same as the written tests used during the initial learning phase (i.e. open-ended, short-answer questions).

After completing the final assessment, students filled out a questionnaire regarding their experiences during the study. The questionnaire probed student perceptions about how the various learning activities affected learning, preferences regarding the various learning activities, and willingness to use repeated testing in the future. Questions were open-ended to allow students to fully explain their thoughts and feelings (e.g., "*Please describe how the SP encounters influenced your learning of the topic*," and "*Please describe how taking the written tests influenced your learning of the topic*.").

Standardized patients

Fourteen SPs were used during the initial phase of the study. A new group of 18 SPs were recruited for the final assessment phase so that students did not encounter the same SP in the final assessment that they had seen during the initial learning phase. All SPs were community members with no medical background. However, all SPs came from the pool of SPs used routinely in the SP program for our institution. As such, they all had extensive experience as SPs. SPs underwent at least 1 h of course-based, topic-specific training prior to participating in the study. They also had undergone at least 3 h of general training on performance, script design, checklists, and feedback.

Procedures

The initial learning phase of the study consisted of a teaching session followed by four additional learning sessions. First, students participated in a 2-h, interactive teaching session covering the three topics used in the study: migraine, seizures, and myasthenia gravis. The teaching session covered all of the information that would be used in the learning activities. The following day, students participated in another session in which they engaged in each of the three learning activities described above. The students had been randomized to participate in one of the three activities for each of the three topics. The assignment of activities to topic was counter-balanced across students so that each of the nine possible combinations of topic and activity occurred equally often. Each student first saw an SP, then took a written test on a different topic, and then studied a review sheet on the third topic. Students received feedback for both the SP test and the written test. After the SP encounter, students were asked to score their performance using the checklist. For the written tests, students graded their tests using an answer sheet in order to encourage careful consideration of each of the critical pieces of information. For the review sheet, students were instructed to study the review sheet as they normally would study this type of information. They were allowed to read and re-read the sheet as many times as they wanted. Once a week for the next 3 weeks, students returned for additional learning sessions in which they engaged in the same activities for each topic. Altogether, each student performed the assigned learning activity for each given topic four times.

Approximately 6 months after the initial teaching session, all students returned for the first session of the final assessment phase. During that first session, the students took a final SP test for each of the three topics. No feedback was given after these final SP tests. One week later, all students participated in another session in which they took written tests on all three topics. After completing the written test, students filled out the questionnaire regarding their experiences in the study.

Scoring and statistical analyses

Two coders (DPL and ALL) scored all of the SP tests and the written tests from the final assessment phase. Kappa statistics measuring inter-rater reliability were calculated separately for the two types of test. Reliability was high for both the SP tests ($\kappa = .92$) and the written tests ($\kappa = .96$). Differences were resolved by discussion. Given the high inter-rater reliability, the SP tests and written tests from the initial learning sessions were scored by individual coders. For two students, technical difficulties caused a malfunction of the video recording of one of their initial SP tests. For these two individual tests, the students' self-scores were used for analyses. In scoring the SP tests, if the SP revealed the information that the student was to elicit without the student asking, the item was removed from scoring for that student in that encounter and the total number of items was adjusted accordingly. In this way, students were neither penalized nor rewarded for SPs accidently giving away information. During coding of the final SP tests, one item from the migraine topic was discarded because it was discovered that the script for the scenario was written in such a way that almost all standardized patients accidently gave the information away before the student could ask about it.

For the questionnaire at the end of the study, the responses to the open ended questions were collated and coded for common themes. These themes were then compiled and summarized.

Results were collapsed across topics and analyzed by learning activity unless otherwise specified. All results were considered significant at the .05 alpha level. Eta-squared and Cohen's d are the measures of effect size reported for all significant effects in the ANOVA and *t* test analyses, respectively. Statistical analyses were performed with SPSS software.

Results

Initial learning

During the initial learning phase, the proportion of correct responses increased steadily across the four tests in both the SP and written testing conditions (see Table 1), which was expected given that students received feedback after each test. A 4 (initial test number) × 2 (learning condition) repeated measures ANOVA confirmed this observation by showing a significant linear trend of initial test number [$F(1, 38) = 251.05$, $MSE = 4.13$, $p < .0001$, $\eta^2 = .61$]. There was also a main effect of learning condition [$F(1, 114) = 9.39$, $MSE = .35$, $p = .004$, $\eta^2 = .05$] in which performance was significantly higher on the initial written tests relative to the SP tests. The interaction between these two variables was not significant [$F(1, 38) = 1.53$, $MSE = .01$, $p = .22$].

Table 1 Mean proportion correct on the written and standardized patient tests during the initial learning phase

Learning activity	Test 1	Test 2	Test 3	Test 4
Written test	.63	.73	.82	.89
Standardized patient test	.54	.67	.80	.85

Final standardized patient test

The left panel of Fig. 1 shows the mean proportion of correct responses on the final SP test as a function of initial learning activity. Students retrieved and applied more of the critical information during the SP test when they had learned the information through repeated SP tests relative to repeated written tests, which in turn produced better performance than repeated review of the material. A 3 (learning activity) × 3 (topic) repeated measures ANOVA was conducted to analyze performance. There was a significant main effect of learning activity [$F(2, 76) = 15.86$, $MSE = .2$, $p < .0001$, $\eta^2 = .26$]. Follow-up pair-wise comparisons showed that SP testing produced significantly better performance than both written testing [.59 vs. .49; $t(40) = 3.30$, $p = .002$, $d = .55$] and studying [.59 vs. .43; $t(40) = 5.11$, $p < .0001$, $d = .84$]. In addition, written testing produced significantly better performance than studying [.49 vs. .43; $t(40) = 2.14$, $p = .04$, $d = .33$]. There was no main effect of topic [$F(2, 38) = 2.54$, $MSE = .05$, $p = .09$). However, there was a

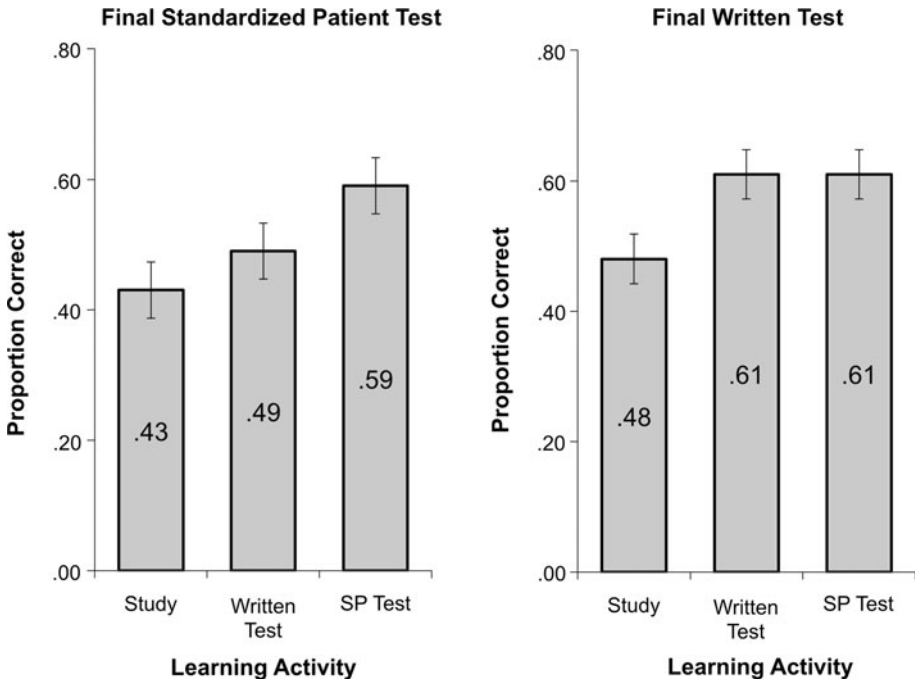

Fig. 1 Mean proportion correct on the final standardized patient test (*left panel*) and the final written test (*right panel*) as a function of initial learning activity. *Error bars* represent 95 % confidence intervals

significant interaction between learning activity and topic [$F(4, 76) = 3.67$, $MSE = .01$, $p = .009$, $\eta^2 = .12$], indicating that the pattern of results varied among the topics.

To follow up on this interaction, the final SP test performance was broken down by topic. Table 2 contains the mean proportion of correct responses as a function of learning activity and topic. In both the seizures and migraine topics, engaging in the SP and written test activities during initial learning led to better performance on the final SP test relative to repeatedly studying the material. However, this pattern did not hold for the myasthenia gravis topic—although the SP test condition produced the best performance (.69), the written test condition led to slightly worse performance than the study condition (.46 vs. .49). An item analysis indicated that this alternate pattern of results was driven by the subset of items that pertained to the physical exam portion of the myasthenia gravis topic. Whereas the written testing condition performed better on the final SP test than the study condition on items pertaining to history and treatment counseling (.54 vs. .47), the written testing condition performed worse than the study condition on the physical exam items (.41 vs. .50). Interestingly, this subset of items represents the only material across all three topics in which the written testing condition performed worse than the study condition.

Final written test

The right panel of Fig. 1 shows the mean proportion of correct responses on the final written test as a function of initial learning activity. The pattern of results on the final written test was different than that for the final SP test. In particular, the written testing group made substantial gains on the final written test compared to the final SP test. Both the written testing and SP testing conditions were better than the study condition. A 3 (learning activity) × 3 (topic) repeated measures ANOVA revealed a significant main effect of learning activity [$F(2, 76) = 16.69$, $MSE = .01$, $p < .0001$, $\eta^2 = .25$]. Follow-up pairwise comparisons showed that both the SP testing and written testing conditions produced a significantly greater proportion of correct responses relative to the study condition [.61 vs. .48; $t(40) = 4.44$, $SEM = .03$, $p = .0002$, $d = .73$; and .61 vs. .48; $t(40) = 4.62$, $SEM = .03$, $p = .0001$, $d = .70$; respectively]. However, there was no difference between the two testing groups ($t < 1$). Again, there was no main effect of topic [$F(2, 38) = 1.24$, $MSE = .05$, $p = .30$]. However, as with the previous analyses, there was a significant interaction between learning activity and topic [$F(2, 76) = 6.09$, $MSE = .01$, $p = .0003$, $\eta^2 = .18$].

The results of the final written test were broken down by topic to follow up on the interaction (see Table 3). The source of the interaction seems to be due to the variation between the two testing conditions, although both the SP and written testing conditions performed better than the study condition for all three topics. The difference between the final SP and written tests was most dramatic for the myasthenia gravis topic. For the final written test on myasthenia gravis the written testing group performed better than the study

Table 2 Mean proportion correct on the final standardized patient test as a function of initial learning activity and topic

Learning activity	Seizures	Migraine	Myasthenia	Grand mean
Study	.39	.40	.49	.43
Written test	.56	.46	.46	.49
Standardized patient test	.66	.46	.64	.59

Table 3 Mean proportion correct on the final written test as a function of initial learning activity and topic

Learning activity	Seizures	Migraine	Myasthenia	Grand mean
Study	.45	.45	.55	.48
Written test	.65	.55	.62	.61
Standardized patient test	.64	.49	.69	.61

group (.62 vs. .55, respectively), which was different than for the SP final test (.46 vs. .49, respectively). One potential explanation for this difference is the structure and questions of the final written test provided a cue to retrieve the physical exam items for the myasthenia gravis topic, whereas the final SP test did not. When students engaged in SP testing during initial learning, they learned to retrieve the physical exam items without the specific cues of questions and written structure, and thus they successfully retrieved this information on both the final SP test and the final written test. In contrast, when students engaged in written testing during initial learning, they learned to retrieve the physical exam items via familiarity with the specific cues of the questions and structure. When they did not receive this cue on the final SP test, they failed to retrieve those items. However, once they received the appropriate cue on the final written test, they were able to retrieve the physical exam information.

Conditional analyses

The critical mechanism that produces the mnemonic benefit of testing is the successful retrieval of information from memory. In order to investigate how retrieval during initial learning influenced performance on the final tests, we conducted linear regressions in which the number of successful retrievals on the initial tests was used to predict the proportion of correct responses on the final tests. Table 4 shows the mean proportion correct on the final SP test and the final written test as a function of initial learning activity and the number of times the critical piece of information was successfully retrieved during initial learning. When subjects engaged in SP testing during initial learning, the number of successful retrievals predicted performance on both the final SP test [$\beta = .34$, $t(1084) = 11.82$, $p < .0001$; $\Delta R^2 = .11$, $F(1, 1{,}085) = 139.73$, $p = .0001$] and the written test [$\beta = .31$, $t(1130) = 10.91$, $p < .0001$; $\Delta R^2 = .10$, $F(1, 1{,}131) = 119.13$, $p < .0001$]. Similarly, when subjects engaged in written testing during initial learning, the number of

Table 4 Mean proportion correct by item on the final SP test (top panel) and the final written test (bottom panel) as a function of learning activity and the number of times the item of information was successfully retrieved during initial four learning sessions

Learning activity	Number of successful retrievals				
	0	1	2	3	4
Final SP test					
Written test	.26	.21	.41	.53	.56
Standardized patient test	.20	.42	.49	.64	.75
Final written test					
Written test	.13	.24	.44	.62	.75
Standardized patient test	.22	.45	.57	.67	.73

successful retrievals also predicted performance on both the final SP test [$\beta = .21$, $t(1084) = 7.14$, $p < .0001$; $\Delta R^2 = .05$, $F(1, 1{,}085) = 50.99$, $p < .0001$] and the final written test [$\beta = .38$, $t(1130) = 13.80$, $p < .0001$; $\Delta R^2 = .14$, $F(1, 1{,}131) = 190.35$, $p < .0001$]. Thus, the greater the number of successful retrievals that a piece of information received through either SP or written testing, the more likely that piece of information was to be successfully retrieved again on the final SP and written tests.

Questionnaire responses

Our questionnaire data provide important insights into factors that may have influenced our results. For instance, on the questionnaire at the end of the study, students were asked how they used the review sheet for the repeated study portion. Seventy-one percent of the students reported that they had used the review sheet to quiz themselves on the material. Most said that they quizzed themselves multiple times. This finding suggests that many students our control group actually engaged in self-testing, which may have reduced the magnitude of the effects that we found.

Students were also asked if they studied the content outside of study activities. Only one student reported having studied additional outside information other than what was covered in regular medical school classes. Students were asked what additional teaching that they had received on the topics covered in the study. Students reported that myasthenia gravis and its treatment had been discussed in pharmacology classes during the first part of the second year of medical school. Blood pressure and pain medications that had been introduced in the migraine topic were briefly covered in their pharmacology course, but not in the context of migraine headaches. The additional teaching on myasthenia gravis provided a review that might have helped the study group perform better on this topic than other topics. However, students reported that the vast majority of the information covered in our study had not been included in any of their courses.

When students were asked if any topic was more difficult than others, the migraine topic was identified by 66 % of students. Students felt this topic was more difficult because of the larger number of medications that they had to learn. Fifty-one percent of students identified myasthenia gravis as the easiest of the topics. They felt that the ease of this topic was due to the physical exam components and the clear mechanisms of pathophysiology.

Our questionnaire data also provide insight into how repeated testing influenced student learning. With open-ended questions, students were asked to describe how repeated SP encounters and repeated written tests affected their learning. In response to the open-ended question, many students (20 %) remarked that learning from their mistakes and applying the feedback that they received from scoring their own checklist was a key component of learning with the SPs. Several students (17 %) also identified the emotional stress of facing a simulated patient as a major factor in their learning. As one student said, "I learn very well under pressure, especially from mistakes. Thus mistakes or lack of information while dealing with 'real' patients drove me to master the material, as the consequences would not just affect me, but my patients as well." Students (20 %) also felt that the SP encounters created experiences that they were able to more easily recall for future use (e.g., "[I] was able to better recall material since we could go through the motions in our head").

Some similarities emerged regarding the effects of written testing on students' learning. Even more students (29 %) felt that learning from feedback was critical to the effect of the tests on learning (e.g., "The written test allowed me to see what I knew. Going through and correcting the tests let me see what areas I had difficulty with and I could focus on these"). Students (22 %) also felt that the written tests helped them to organize the information. For

example, one student wrote: "It helped categorize (and group) the responses to questions so I could remember them better." It is notable that 15 % of students cited the direct effect of practicing recall of information as a major factor in their learning. As one student said, "The testing method works well for me. I am forced to recall [information] on my own before seeing what the answer is."

The final SP encounters and written tests allowed students the opportunity to compare both testing modalities directly for the same topics. The students' responses provide some insight into why different pattern of results were obtained on the two types of final test. In response to an open-ended question comparing these two final test formats, several themes emerged. As would be anticipated, many students (20 %) commented that they found the final test modality that corresponded to the method that they had used to learn the topic to be the easiest. Students (37 %) felt that the organization and structure of the written test also helped them to recall information and ultimately made the written test easier than the SP test. However, the flow of conversations with SPs also provided structure and help for some students (15 %). One student captured both concepts, "It was easier to remember details when completing the written test and the questions provided cues to recall. There may have been some information gained in the SP and not written in the exam by just completing a full normal H&P [history and physical]." Some students (17 %) found the task of translating their own idiosyncratic information structures from their previous learning with SPs into the format of written test questions to be difficult. "I think the SP [test] for my actual SP case was easier by far than the written test. The written test had groupings that I wasn't necessarily utilizing in my interviewing so it took a while to sort that out." Many students (17 %) also felt that the final SP encounter produced a more stressful experience (e.g., "The written test[s] were easier because there was no pressure to have something to say, I could come back after remembering something more easily"). While many students felt that there were some advantages to the final SP test, the final written test seemed to be easier.

When students were asked whether they were willing to participate in repeated written testing as a learning technique in future courses, 80 % responded in the affirmative. Interestingly, 93 % of students reported that they would be willing to engage in repeated SP encounters as part of their courses. When asked to rank the three learning activities in their order of preference to use in learning 59 % gave SPs as their preferred method, with 17 % listing written tests as their preferred method of learning. 46 % listed written tests as their second choice with 20 % listing SPs as their second choice.

Discussion

Effects of testing and test format on retention and application

Our study demonstrates that repeated testing with both SPs and written tests leads to superior long-term retention on both clinical application tests with SPs and written tests compared to repeatedly studying a review sheet. Importantly, the final SP test provided an approximation of how these various learning activities would lead to application of knowledge months after initial learning. When the topics were learned through written tests and studying, the final SP exam represented an entirely new context that required the application of knowledge. Even when students had learned the initial information through SP testing, they encountered a new patient on the final SP test who had a distinct history from any previous SP that they had seen during the initial learning tests.

We found that learning a clinical topic through repeated testing with a standardized patient is a powerful method of promoting long-term retention. Even after 6 months, retention on the final SP test for the students who learned their topic through repeated SP tests was close to or above the level of initial learning on the first SP tests. In our final SP test, repeated testing with SPs produced a large effect size of .84 standard deviations (SD) above repeatedly studying a review sheet and a moderate effect size of .55 SD above repeated written testing. Our findings have direct implications for clinical training in that they suggest that students who have repeated practice with patients (or at least simulated patients) will be able to retain and apply more of their knowledge in a clinical setting. Many training programs seek to have students see a wide variety of patients. However, our study suggests that trainees would also benefit from the practice of seeing the same types of patients repeatedly in order to engage in repeated retrieval practice. SPs can be used when repeated patient exposures are not possible or practical.

The above considerations fit well with the recent efforts to emphasize the importance of deliberate practice in medical education. The Best Evidence in Medical Education (BEME) review of simulation identified deliberate practice as a key element of learning from simulation (Issenberg et al. 2005). Repeated practice with mannequin simulators has been shown to improve retention of skills in several settings. However, the concept of deliberate practice has been most often discussed and applied to improve retention in the arena of procedural skills taught with mannequin simulators (McGaghie et al. 2011). Despite the emphasis on deliberate practice in mannequin simulation, very little work has been done to demonstrate that repeated deliberate practice in simulation with SPs can produce similar benefits. Typically SPs have been used to teach communication and physical examination skills (e.g., O'Sullivan et al. 2008, and Safdieh et al. 2011). Our study demonstrates how the concept of deliberate practice can be expanded to simulated patient encounters with SPs to generate long-term retention of knowledge-based content.

Our study also demonstrates that repeated written testing generally produced superior long-term retention compared to repeatedly studying a review sheet, replicating past research (Larsen et al. 2009). On the final SP test 6 months after initial learning, there was a significant difference in performance between the written testing and study groups, but this effect was small to moderate in size (equivalent to .33 SD); thus, the difference in retention between the written testing and study groups was not as apparent in a clinical setting when compared to the larger difference between the SP testing and study groups. Interestingly, the written test group made large gains when the final test was in the format of a written test. On the final written test both the SP and written test groups demonstrated moderate to large effect sizes of superior retention compared to the study group (.70 SD for the written testing group and .73 SD for the SP group).

Analysis of our questionnaire results provides further insight into why this result may have occurred. Many students felt that the final written test was easier than the final SP test because the written test had more structure and provided a greater number of retrieval cues compared to the SP encounter. Some students noted that the questions on the written test would cue them to remember an item, whereas they had no such cues during the SP encounter. In essence, the SP test was a free recall test in which the students had only the broad and generic structure of a typical patient encounter to guide their retrieval. Thus, when students learned through the SP testing, they were required to generate their own organization of the information. Indeed, Zaromb and Roediger (2010) demonstrated in a laboratory setting that free recall tests do enhance students' abilities to organize information relative to repeated study of the same information. The ability of students in our study to create their own mental organization of the material may have helped them on the

final SP test. In contrast, when students had learned through written testing, the questions on the written test provided the organization for them so that they never needed to generate it themselves. Once this structure was removed on the final SP test, these students may have struggled to recall some of the information that they had learned.

Students also mentioned the emotional challenge of trying to recall information in front of a patient. The emotional stress involved with performing in front of a live person may have led to better performance in that type of context. Similar results are seen in a study conducted by DeMaria et al. (2010). They found that students who trained in cardiac resuscitation using simulation with emotional stressors performed better in a simulated cardiac arrest 6 months later compared to students who trained without the emotional stressors. Interestingly, both groups performed equivalently on a written test of knowledge. These findings correspond nicely with our results and suggest that the emotional component of simulation may play an important role in preparing students to effectively retrieve and apply their knowledge in clinical settings. Presumably, the same emotions play less of a role in recall during written testing.

Another important point to consider is the degree of mental effort. Both the SP testing and written testing required greater mental effort in retrieving answers than studying the review sheet. This fact may also explain why both simulation and written testing produced better retention than the strategy of self-testing and reading that many students reported using to study the review sheet. The self-testing presumably did not require the same degree of mental effort as the instructor-generated tests. This interpretation is consistent with studies in cognitive psychology, which demonstrate that activities requiring greater mental effort are associated with greater retention. This principle has been termed "desirable difficulties" (Bjork 1994). Increased retrieval effort has been implicated in laboratory studies as underlying the mnemonic effects of retrieval practice (Pyc and Rawson 2009).

Implications for medical education

Despite the apparent advantages and benefits of repeated SP encounters, SPs are expensive, time-consuming, and impractical in many situations. Our study shows that substantial benefit can be obtained from repeated written testing, even though the benefit may not be as robust with respect to application in a clinical setting. Future studies must investigate whether written testing would be more effective at promoting transfer to clinical settings if it can be modified in ways that better approximate the experience of an SP encounter. For instance, would a written test that involves free recall in which students are asked to imagine interacting with an SP better simulate the lack of structure and retrieval cues that occurs in a clinical encounter? Could timed written tests simulate the emotional tension of the clinical setting? Another possible solution would be to provide students with a mix of SP encounters and written testing. Such a hybrid approach may give students the benefit of exposure to the person-to-person interaction of the clinic, while providing additional retrieval practice in a less costly and more flexible format.

Using both SP and written testing to help students learn seems reasonable given that both involve the process of repeated retrieval that promotes superior retention and application of knowledge. When we examined the relationship between performance during initial learning and on the final test, we found that the number of successful retrievals of an item correlated with that item being successfully recalled on the final SP test as well as the final written test. This relationship held regardless of whether students learned through initial SP or written testing. It should be noted that our finding is correlational and therefore

does not prove causation. However, laboratory studies in cognitive psychology in which the number of retrievals are manipulated have shown that increased retrieval does cause greater retention. (Wheeler and Roediger 1992; Karpicke and Roediger 2007). Although educators may be tempted to view a single test as an effective means of obtaining the benefits of retrieval practice, these findings indicate that *repeated* retrieval practice is important for promoting long-term retention of knowledge.

Although students in our study achieved a relatively high level of performance by the fourth test (averages of the 89 and 85 % on the fourth written and SP tests, respectively), their performance on the final retention tests may have been higher if we had provided additional retrieval practice (e.g., one or two extra tests such that they successfully retrieved 100 % of the information on the final learning test). Cognitive psychology studies have shown that engaging in additional retrieval practice after the information has been correctly recalled once produces superior retention (e.g., Karpicke and Roediger 2008). Such additional practice could be considered overlearning (i.e. continuing study or practice once material has been mastered), but it is likely to be beneficial if it is spaced out over time. Of course, any additional practice takes time, whether it occurs inside or outside the classroom. Educators will need to find a balance between the goal of providing enough opportunities to practice retrieval and the practical constraints of the time available for such activities.

In terms of implementing repeated testing in medical education, a few other points are important to consider. First, in our study students received feedback after each initial learning test, which produced steady improvement in performance across the tests. Indeed, when students in our study were asked to identify how repeated testing (through SPs or written tests) helped them to learn, they frequently referred to the feedback that they received that allowed them to identify and correct mistakes. Students also graded their written tests and SP encounters in order to force them to consider feedback for each item. Although we highly recommend providing feedback after retrieval practice, it is important to note that testing usually promotes long-term retention even if feedback is not given (e.g., Butler and Roediger 2008; Roediger and Karpicke 2006a, b; Karpicke and Roediger 2008), so long as performance on the tests is reasonably high. Second, students were tested on their knowledge every week, which provided some difficulty in terms of the mental effort needed to retrieve the information but also minimized the forgetting between testing events. Waiting too long between tests can increase inter-test forgetting, which may reduce the benefits of repeated testing because students cannot successfully retrieve the information (cf. Larsen et al. 2009). It should be noted, though, that providing some spacing between the tests is helpful and likely contributed to the superior effect of testing that we found after 6 months. Studies in medical education that have used repeated tests without spacing have not found a long-term benefit to retention (Schmidmaier et al. 2011). Indeed, cognitive psychology studies have shown that for retention intervals of months to years, retrieval practice should be spaced on the order of weeks to months (Cepeda et al. 2008). Future research should directly investigate different feedback regimens and testing intervals to explore how these factors can be optimized to improve the efficacy of retrieval practice.

Finally, when considering the practical application of our findings, it is important to recognize the positive response of the students in our study to repeated testing both with written tests and SP encounters. Eighty percent of students stated that they would be willing to engage in repeated testing as part of a course. Ninety-three percent of students were willing to engage in repeated SP tests. The majority of students listed SPs and written tests as their preferred means of learning compared to studying review materials. Thus,

student willingness should not be seen as a barrier to implementing test-enhanced learning in medical education.

Limitations

Our study has some limitations. First, our analyses showed a significant interaction between learning activities and topics, indicating that the relative effect of the different learning activities varied by topic. Given that we used real-life clinical topics that varied in many dimensions, such an interaction is somewhat expected. Nevertheless, as noted earlier, both of the initial testing conditions produced better final test performance than the review sheet condition with only one exception—the myasthenia gravis topic on the SP final test. This disparate result seemed to be driven by the fact that the written test group performed worse than the review sheet group on the items pertaining to the physical exam. Although it is not entirely clear why this result occurred, we speculate that when students learned the physical exam items through written testing, they may have become dependent on the structure of the written test. Without the benefit of these cues on the final SP test, students who learned through written testing performed even worse than students who had learned the physical exam items through studying. Indeed, studies in cognitive psychology have demonstrated that memory performance depends upon a match between the cues that were present during initial learning and those that are provided at retrieval (e.g., Tulving and Thomson 1973). Overall, the variation in results by topic suggests that when implementing test-enhanced learning in medical education, educators should be aware that certain topics may not be ideal for learning through written testing.

All of our results must also be considered in the context of our extremely conservative control condition in which students were simply instructed to study the review sheet as they normally would for their courses. However, most students would not typically engage in repeated, spaced study as occurred in our study design. Often in medical education, students do not engage in any further learning activities after a topic has been covered in a didactic conference or a course. Furthermore, seventy-one percent of students reported that they studied the sheet by quizzing themselves, often multiple times. Therefore, the control condition in our investigation does not represent a pure re-reading activity but rather a mix of reading and self-testing. Given this consideration, the fact that both testing conditions produced significantly better retention and application by comparison is remarkable. However, it is important to note that the self-testing that occurred in the study control condition may have reduced the magnitude of the effects. This finding also raises important questions about the adequacy of relying on students to use self-testing as a learning strategy compared with tests that are given by an instructor or another external source. Additionally, the self-testing in the study condition may have contributed to some of the variability between topics in that it may have been easier to self-test for some of the topics than others.

Conclusions

Our results show that repeated testing, both through simulation with SPs and written tests, improves long-term retention and application of knowledge relative to repeated study. Test-enhanced learning holds great promise for medical education. We believe that our findings provide educators with an evidence-based tool that can help to bridge the gap between initial learning and final application of knowledge.

Acknowledgments This study was funded by an Educational Research Grant from the American Academy of Neurology and by the McDonnell Center for Systems Neuroscience at Washington University in St. Louis School of Medicine.

References

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

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

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *The European Journal of Cognitive Psychology, 19*(4/5), 514–527.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*, 604–616.

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

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effect in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*, 1095–1102.

DeMaria, S., Jr, Bryson, E. O., Mooney, T. J., Silverstein, J. H., Reich, D. L., Bodian, C., et al. (2010). Adding emotional stressors to training in simulated cardiopulmonary arrest enhances participant performance. *Medical Education, 44*, 1006–1015.

Fallucco, E. M., Hanson, M. D., & Glowinski, A. L. (2010). Teaching pediatric residents to assess adolescent suicide risk with a standardized patient module. *Pediatrics, 125*, 953–959.

Feddock, C. A., Hoellein, A. R., Griffith, C. H., Wilson, J. F., Lineberry, M. J., & Haist, S. A. (2009). Enhancing knowledge and clinical skills through an adolescent medicine workshop. *Archives of Adolescent and Pediatric Medicine, 163*(3), 256–260.

Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Lee, G. D., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher, 27*(1), 10–28.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*, 772–775.

Karpicke, J. D., & Roediger, H. L., I. I. I. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151–162.

Karpicke, J. D., & Roediger, H. L., I. I. I. (2008). The critical importance of retrieval for learning. *Science, 15*, 966–968.

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2010). The testing effect on skills might last 6 months. *Advances in Health Sciences Education, 15*(3), 395–401.

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2008). Test-enhanced learning in medical education. *Medical Education, 42*(10), 959–966.

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2009). Repeated testing versus repeated study: A randomized, controlled trial. *Medical Education, 43*(12), 1174–1181.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., I. I. I. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*, 399–414.

McGaghie, W. C. (2008). Research opportunities in simulation-based medical education using deliberate practice. *Academic Emergency Medicine, 15*, 995–1001.

McGaghie, W. C., Issenberg, S. B., Cohen, E. R., Barsuk, J. H., & Wayne, D. B. (2011). Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Academic Medicine, 86*, 706–711.

O'Sullivan, P., Chao, S., Russell, M., Levine, S., & Fabiny, A. (2008). Development and implementation of an objective structured clinical examination to provide formative feedback on communication and interpersonal skills in geriatric training. *Journal of the American Geriatric Society, 56*(9), 1730–1735.

Price, J. W., Price, J. R., Pratt, D. D., Collins, J. B., & McDonald, J. (2010). High-fidelity simulation in anesthesiology training: A survey of Canadian anesthesiology residents' simulator experience. *Canadian Journal of Anesthesiology, 57*, 134–142.

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.

Roediger, H. L., I. I. I., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20–27.

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

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

Safdieh, J. E., Lin, A. L., Aizer, J., Marzuk, P. M., Grafstein, B., Storey-Johnson, C., & Kang, Y. (2011). Standaridized patient outcomes trial (SPOT) in neurology. *Medical Education Online, 16*. doi:10.3402/meo.v16i0.5634.

Schmidmaier, R., Ebersbach, R., Schiller, M., Hege, I., Holzer, M., & Fischer, M. R. (2011). Using electronic flashcards to promote learning in medical students: Retesting versus restudying. *Medical Education, 45*, 1101–1110.

Stevens, D. L., King, D., Laponis, R., Hanley, K., Zabar, S., Kalet, A. L., et al. (2009). Medical students retain pain assessment and management skills long after an experiential curriculum: A controlled study. *Pain, 145*(3), 19–24.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 352–373.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3*, 240–245.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organization processes. *Memory & Cognition, 38*, 995–1008.