

Online “Spaced Education Progress-Testing” of Students to Confront Two Upcoming Challenges to Medical Schools

B. Price Kerfoot, MD, EdM, Kitt Shaffer, MD, PhD, Graham T. McMahon, MD, MMSc, Harley Baker, EdD, Jamil Kirdar, MbChB, MRCP, Steven Kanter, MD, Eugene C. Corbett, Jr., MD, Roger Berkow, MD, Edward Krupat, PhD, and Elizabeth G. Armstrong, PhD

Abstract

Purpose

U.S. medical students will soon complete only one licensure examination sequence, given near the end of medical school. Thus, schools are challenged to identify poorly performing students before this high-stakes test and help them retain knowledge across the duration of medical school. The authors investigated whether online spaced education progress-testing (SEPT) could achieve both aims.

Method

Participants were 2,648 students from four U.S. medical schools; 120 multiple-choice questions and explanations in preclinical and clinical domains were

developed and validated. For 34 weeks, students randomized to longitudinal progress-testing alone (LPTA) received four new questions (with answers/explanations) each week. Students randomized to SEPT received the identical four questions each week, plus two-week and six-week cycled reviews of the questions/explanations. During weeks 31–34, the initial 40 questions were re-sent to students to assess longer-term retention.

Results

Of the 1,067 students enrolled, the 120-question progress-test was completed by 446 (84%) and 392 (74%) of the LPTA and SEPT students, respectively. Cronbach alpha reliability was 0.87.

Scores were 39.9%, 51.9%, 58.7%, and 58.8% for students in years 1–4, respectively. Performance correlated with Step 1 and Step 2 Clinical Knowledge scores ($r = 0.52$ and 0.57 , respectively; $P < .001$) and prospectively identified students scoring below the mean on Step 1 with 75% sensitivity, 77% specificity, and 41% positive predictive value. Cycled reviews generated a 170% increase in learning retention relative to baseline ($P < .001$, effect size 0.95).

Conclusions

SEPT can identify poorly performing students and improve their longer-term knowledge retention.

The National Board of Medical Examiners, the Educational Commission for Foreign Medical Graduates, and the Federation of State Medical Boards have recently endorsed a plan to replace the current three-step licensure examination system with two gateway examination sequences.¹ These will be administered near the end of medical school and at the end of internship. The first gateway exam sequence will assess both basic and clinical science knowledge and the ability to integrate and apply these knowledge sets. Although this new system of gateway exams has many benefits, it poses two substantial challenges to U.S. medical schools. First, it is critical that students in

need of remediation be identified before taking this high-stakes summative exam sequence, the scores from which can be used by residency programs for matching decisions. Second, retention of basic science knowledge learned by current methods of medical education is quite poor.^{2,3} Given this, it is not clear how knowledge learned by students in the first two years of medical school can be effectively retained one to two years later for the first gateway exam sequence.

Spaced education progress-testing (SEPT) has the potential to help medical schools overcome both challenges. SEPT is a fusion of online spaced education with traditional progress-testing of medical students.^{4–7} SEPT consists of two elements: (1) longitudinal progress-testing via questions and answers/explanations e-mailed at regular intervals to students, and (2) cycled reviews of the material over spaced intervals of time to improve long-term retention of learning via the spacing effect. The spacing effect refers to the psychology research finding that

educational encounters that are repeated over spaced intervals of time (spaced distribution) result in more efficient learning and improved retention compared with encounters clustered at a single time point (bolus, or massed, distribution).^{8–11} Prior research has demonstrated that online spaced education improves knowledge acquisition, boosts learning retention for up to two years, and can change clinical behavior.^{12–15} Additionally, spaced education is well accepted by learners.^{13,14}

We hypothesized that SEPT’s longitudinal progress-testing could act as an effective diagnostic tool to prospectively identify those students at risk of performing below standard on their licensure examinations. We also hypothesized that SEPT’s cycled reviews of content could improve long-term retention of core knowledge. To investigate these hypotheses, we conducted a 34-week randomized trial of SEPT at four U.S. medical schools.

Please see the end of this article for information about the authors.

Correspondence should be addressed to Dr. Kerfoot, VA Boston Healthcare System, 150 South Huntington Avenue, 151DIA, Jamaica Plain, MA 02130; telephone: (774) 286-9230; fax: (857) 364-6561; e-mail: price.kerfoot@gmail.com.

Acad Med. 2011;86:300–306.
First published online January 18, 2011
doi: 10.1097/ACM.0b013e3182087bef

Method

Study participants

A total of 2,648 medical students from all four years at four U.S. medical schools were eligible to participate. The University of Alabama School of Medicine and the University of Virginia School of Medicine are public, Harvard Medical School is private, and the University of Pittsburgh School of Medicine functions as a private medical school with a small amount of state funding. All are four-year medical-doctorate-granting schools whose curricula are structured with 1.5 to 2 years of preclinical studies followed by clinical clerkships. Participants were recruited via e-mail. There were no exclusion criteria. Institutional review board approval was obtained to perform this study.

Development and validation of progress-test content

Each progress-test item consisted of an evaluative component (a multiple-choice question usually based on a clinical scenario) and an educational component (the answer and explanation). Four content areas were selected to cover preclinical (anatomy and histology) and clinical (cardiology and endocrinology) domains. The content of the items was targeted to core information that every medical student should know on graduation. A physician content-expert constructed 40 to 42 questions for each topic area; these were independently content-validated by two domain experts/educators. The questions were then pilot-tested by 30 fourth-year medical students. Psychometric analysis of the questions was performed using the Integrity test analysis software (Castle Rock Research, Edmonton, Alberta, Canada). For each topic area, 30 questions were selected for inclusion based on item difficulty, point-biserial correlation, and Kuder-Richardson 20 score. The educational components of the 120 progress-test items were then constructed by physician content-experts and independently content-validated by two domain experts/educators.

The progress-test items and cycled reviews were delivered to students at designated time intervals via an automated e-mail delivery system. The e-mail presented the clinical scenario and question (evaluative component). On

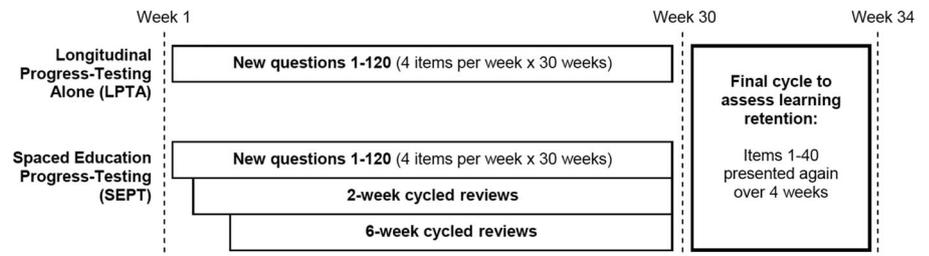


Figure 1 Structure of the multiinstitutional randomized controlled trial described in this report. The trial was conducted across 34 weeks (October 2007 to June 2008) at four U.S. medical schools. Students were stratified by school and year of training and then were randomized into two cohorts. Students in the *longitudinal progress-testing alone* (LPTA) cohort received four new progress-test items each week (one question on each topic on Monday, Tuesday, Thursday, and Friday) for 30 weeks. Students in the *spaced education progress-testing* (SEPT) cohort received the identical four new progress-test items each week for 30 weeks, but in addition they received two-week and six-week cycled reviews of the questions and explanations. During weeks 31 through 34, the initial 40 progress-test questions were re-sent to students in both cohorts to assess their long-term retention of the educational material. Over these four weeks, students received daily e-mails (Monday through Friday) that contained two progress-test questions. Timeline is not drawn to scale.

clicking on a hyperlink, a Web page opened that allowed the student to submit an answer to the question. The answer was downloaded to a central server, and students were immediately presented with a Web page displaying the educational component: the correct answer, a summary of the curricular learning points, explanations why the possible answers were correct/incorrect, and hyperlinks to additional educational material. Because of the question-answer format of the progress-test items, evaluation and education are inextricably linked.

Study design and organization

This multiinstitutional randomized controlled trial was conducted across 34 weeks (October 2007 to June 2008). At enrollment, students reported their Medical College Admission Test (MCAT) and United States Medical Licensing Examination (USMLE) scores. Students were stratified by school and year of training and then were block-randomized (block size = 8) into two cohorts. Students in the *longitudinal progress-testing alone* (LPTA) cohort received four new progress-test items each week (one question on each topic on Monday, Tuesday, Thursday, and Friday) for 30 weeks (see Figure 1). Students in the *SEPT cohort* received the identical four new progress-test items each week for 30 weeks, but in addition they received two-week and six-week cycled reviews of the questions and explanations. For example, a question presented in week 1 was re-sent in week 3 (as a two-week cycled

review) and in week 7 (as a six-week cycled review). During weeks 7 through 30 of the trial, each e-mail to LPTA students contained one question, whereas each e-mail to SEPT students contained three questions (a new question, a two-week cycled review, and a six-week cycled review). The time intervals between cycled reviews were based on psychology research findings to optimize long-term retention of learning.^{9,16}

During weeks 31 through 34, the initial 40 progress-test questions were re-sent to students in both cohorts to assess their long-term retention of the educational material (see Figure 1). Over these four weeks, students received daily e-mails (Monday through Friday) that contained two progress-test questions.

At week 34, students completed a short survey focused on their use of the progress-test items, preferences for future programs, and most recent USMLE Step 1 and Step 2 Clinical Knowledge (CK) scores. Participants received a \$30 bookstore gift certificate on completing the survey and $\geq 85\%$ of the questions.

Outcome measures

The primary outcome measure was student performance on the 120-item progress-test administered over weeks 1 to 30. The secondary outcome measure was student performance on the 40-item test of long-term retention administered over weeks 31 to 34.

Statistical analysis

Power was estimated based on the cross-cohort comparison of students' scores on the 40-item test of learning retention administered over weeks 31 to 34. We calculated that if 172 students entered this parallel-design, randomized controlled trial, the probability would be 90% that the study would detect a score difference at .05 significance (α), given a 5% true difference between scores and a 10% standard deviation (effect size = 0.5). Adjusting for a potential attrition rate of 30%, we would need to recruit 224 students total. To determine the generalizability of SEPT across medical schools, we aimed to recruit at least this number of students from each school (896 total).

Scores for the 120-item progress-test were calculated as the number of SEPT questions answered correctly normalized to a percentage scale. Students were defined as completing the progress-test if they submitted answers to $\geq 80\%$ of the questions. No data for students who submitted answers to $< 80\%$ of the questions were included in the analyses. Unanswered questions were considered to be answered incorrectly. To assess whether missing data influenced the results, reliability and validity analyses were repeated using the expectation maximization algorithm to impute estimated partial credit for missing answers.^{17,18} The results were virtually identical, so we report only the findings based on unadjusted data. Reliability was estimated with Cronbach alpha.^{19,20}

When different MCAT or USMLE scores were reported pre- and posttrial, the score reported at enrollment was used for analysis. To reduce potential errors, self-reported test-score data from 37 students (3%) with extreme/highly improbable MCAT or USMLE scores were eliminated. An audit was performed with the actual MCAT and USMLE scores for participants from one school. For this one school, the students' actual MCAT and USMLE scores were used for analyses in place of their reported scores. Criterion-based validity of the progress-test was assessed via Pearson correlation, partial correlation controlling for MCAT scores, and structural equation modeling. Linear discriminant function analysis, logistic regression, and receiver operator characteristic analyses assessed the predictive validity of SEPT. Because few

Table 1

Baseline Demographic Characteristics of Randomized Cohorts of Medical Students, Four U.S. Medical Schools, 2007–2008*

Characteristic	No. (%) of students in LPTA cohort (n = 534)	No. (%) of students in SEPT cohort (n = 533)
Medical school		
1	166 (31)	165 (31)
2	104 (19)	106 (20)
3	110 (21)	111 (21)
4	154 (29)	151 (28)
Year of training		
1	126 (24)	124 (23)
2	153 (29)	158 (30)
3	141 (26)	138 (26)
4	114 (21)	113 (21)
Degree		
MD	471 (88)	477 (89)
MD-PhD	44 (8)	32 (6)
MD-other	19 (4)	24 (5)
Gender		
Male	263 (49)	259 (49)
Female	271 (51)	274 (51)
Characteristic	Mean (SD)	Mean (SD)
Age	25.1 (2.9)	25.1 (3.0)
MCAT score	33.0 (3.9)	33.5 (3.8)

* Students were stratified by school and year of training and then were randomized into two cohorts. Demographic characteristics of the randomized cohorts were similar. Students in the *longitudinal progress-testing alone* (LPTA) cohort received four new progress-test items each week (one question on each topic on Monday, Tuesday, Thursday and Friday) for 30 weeks. Students in the *spaced evaluation progress-testing* (SEPT) cohort received the identical four new progress-test items each week for 30 weeks, but in addition they received two-week and six-week cycled reviews of the questions and explanations. Percentages may not add to 100% because of rounding.

students had completed the Step 2 exam by the trial's end, the predictive analyses were limited to year 2 students and their Step 1 exam scores. We defined a Step 1 score below the national median as below standard because very few students failed Step 1 or scored below the 25th percentile. Evidence for construct validity was obtained by assessing progress-test performance by year of training.

The analysis of learning retention included data from students who had submitted answers to $\geq 80\%$ of questions 1 through 40 both on initial presentation (weeks 1–10) and in the final cycle (weeks 31–34). Two-tailed *t* tests were used to test the statistical significance of differences in scores across cohorts. Cohen *d* provided the intervention effect sizes.^{21,22} Statistical analyses were performed with SPSS for Windows 18.0 (Chicago, Illinois).

Results

Of the 2,648 eligible students, 1,067 (40%) enrolled in the trial. Five hundred thirty-four and 533 students were randomized to the LPTA and SEPT cohorts, respectively. Demographic characteristics of the randomized cohorts were similar (see Table 1). The 120-question progress-test was completed by 446 (84%) and 392 (74%) LPTA and SEPT students, respectively ($P < .001$; see Figure 2). Completion rate did not vary significantly by students' year of training ($P = .31$) but did vary significantly by medical school (range 70%–89%; $P < .001$).

Overall Cronbach alpha reliability of the 120-question progress-test was 0.87. Mean alpha was 0.85 (SD 0.05) across training years, 0.77 (SD 0.6) across medical schools, and 0.73 (SD 0.10) within year within medical school. There

was no significant difference in reliability by cohort.

The percentage of progress-test items answered correctly was 39.9% (SD 7.8), 51.9% (SD 7.9), 58.7 (SD 9.9), and 58.8% (SD 9.5) for students in years 1 through 4 of medical school, respectively (ANOVA, $P < .001$, Figure 3A). Although this increase in scores over years 1 to 3 provides evidence for construct validity, there was no significant difference in scores between students in years 3 and 4 of medical school ($P = 1.00$, Bonferroni correction). Similar results were found at all four schools (see Figure 3B). In three of the four topic areas (anatomy, histology, and endocrinology), there was no significant difference in scores between years 3 and 4 of medical school (see Figure 3C). In cardiology alone, a small but statistically significant improvement in performance was detected from year 3 to year 4 (54.9% [SD 12.6] to 58.2% [SD 12.2], respectively; $P = .02$), corresponding to an effect size of 0.27.

The score audit found that 96% of students' reported MCAT scores were within 2 points of the actual score (reported mean 35.5 [SD 3.5], actual mean 35.3 [SD 3.4], $r = 0.96$). In addition, the audit demonstrated that 99% of students' reported Step 1 scores were within 10 points of the actual score (reported mean 240 [SD 17], actual mean 240 [SD 17], $r = 0.99$), as were 98% of students' reported Step 2 CK scores (reported mean 242 [SD 20], actual mean 241 [SD 20], $r = 0.97$).

Progress-test performance correlated significantly with MCAT, Step 1, and Step 2 CK scores ($r = 0.23, 0.54$, and 0.60 , respectively; $P < .001$). When controlling for MCAT scores, progress-test performance continued to correlate significantly with Step 1 and Step 2 CK scores ($r = 0.52$ and 0.57 , respectively; $P < .001$). Using a SEPT score of 50% correct as a cutoff, the progress-test correctly identified those second-year students who scored below the mean on Step 1 with 75% sensitivity, 77% specificity, and 41% positive predictive value.

Four hundred five (76%) and 357 (67%) LPTA and SEPT students, respectively, completed $\geq 80\%$ of questions 1 through 40 at initial presentation and in the final cycle (weeks 31–34) and were included in

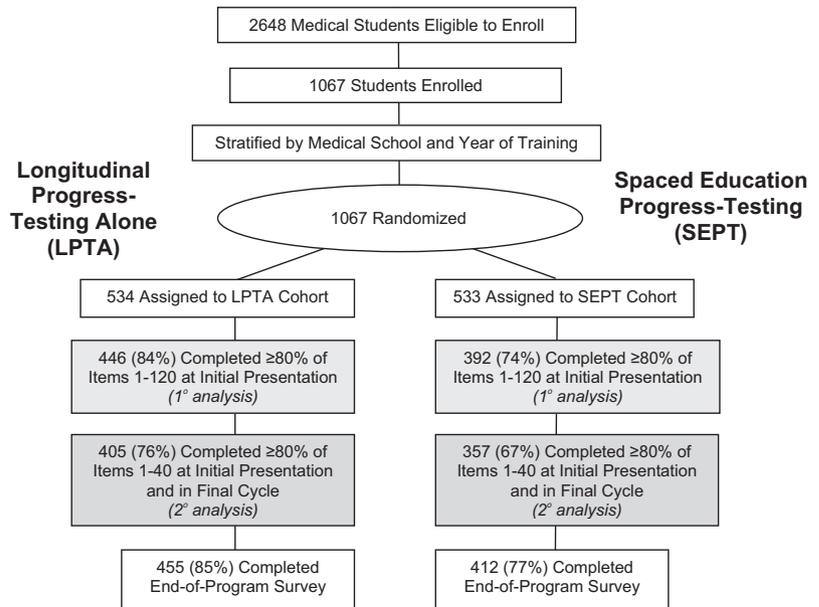


Figure 2 Modified CONSORT (Consolidated Standards for Reporting of Trials) flowchart of the randomized controlled trial described in this report. Students were stratified by school and year of training and then were randomized into two cohorts. Students in the *longitudinal progress-testing alone* (LPTA) cohort received four new progress-test items each week (one question on each topic on Monday, Tuesday, Thursday, and Friday) for 30 weeks. Students in the *spaced education progress-testing* (SEPT) cohort received the identical four new progress-test items each week for 30 weeks, but in addition they received two-week and six-week cycled reviews of the questions and explanations. Students were defined as completing the progress-test if they submitted answers to $\geq 80\%$ of the questions. During weeks 31 through 34, the initial 40 progress-test questions were re-sent to students in both cohorts to assess their long-term retention of the educational material. At week 34, students completed a short survey focused on their use of the progress-test items, preferences for future programs, and most recent USMLE Step 1 and Step 2 Clinical Knowledge scores.

the analysis of learning retention. When initially presented with questions 1 to 40 in weeks 1 to 10, LPTA and SEPT students scored 42.0% (SD 11.3) and 41.0% (SD 11.4), respectively ($P = .19$). When re-presented with the questions in weeks 31 through 34, LPTA students scored 50.0% (SD 12.1), whereas SEPT students scored 62.9% (SD 15.0, $P < .001$), representing a 170% relative increase in learning retention (effect size 0.95; see Figure 4).

The end-of-program survey was completed by 867 (81%) enrolled students. Eighty-seven percent of respondents (754/867) reported never looking up the answers prior to submission online. Students also reported that they used the hyperlinks in a median of 2% of explanations (interquartile range [IQR] 0–10) to access additional topic-related information. Eighty-nine percent (767/867) of survey respondents (72% of all enrollees) requested to participate in future programs.

Discussion and Conclusions

Our results demonstrate that SEPT (i.e., longitudinal progress-testing with cycled reviews) is a reliable, valid, and diagnostically effective method to identify poorly performing students and to significantly improve students' longer-term retention of learning. As a diagnostic tool, SEPT's longitudinal progress-testing was able to prospectively identify year 2 students at risk of performing below the median on their licensure examinations. SEPT's predictive strength ($r = 0.52$ – 0.57) is similar to that of students' course grades in college and medical school,^{23–26} and its sensitivity and specificity are comparable to those of common medical screening tests.^{27,28} As an educational tool, SEPT's cycled reviews of content more than doubled longer-term retention of core knowledge among students. Taken together, these results suggest that SEPT can be of considerable value to U.S. medical schools as they prepare their students for licensure and practice.

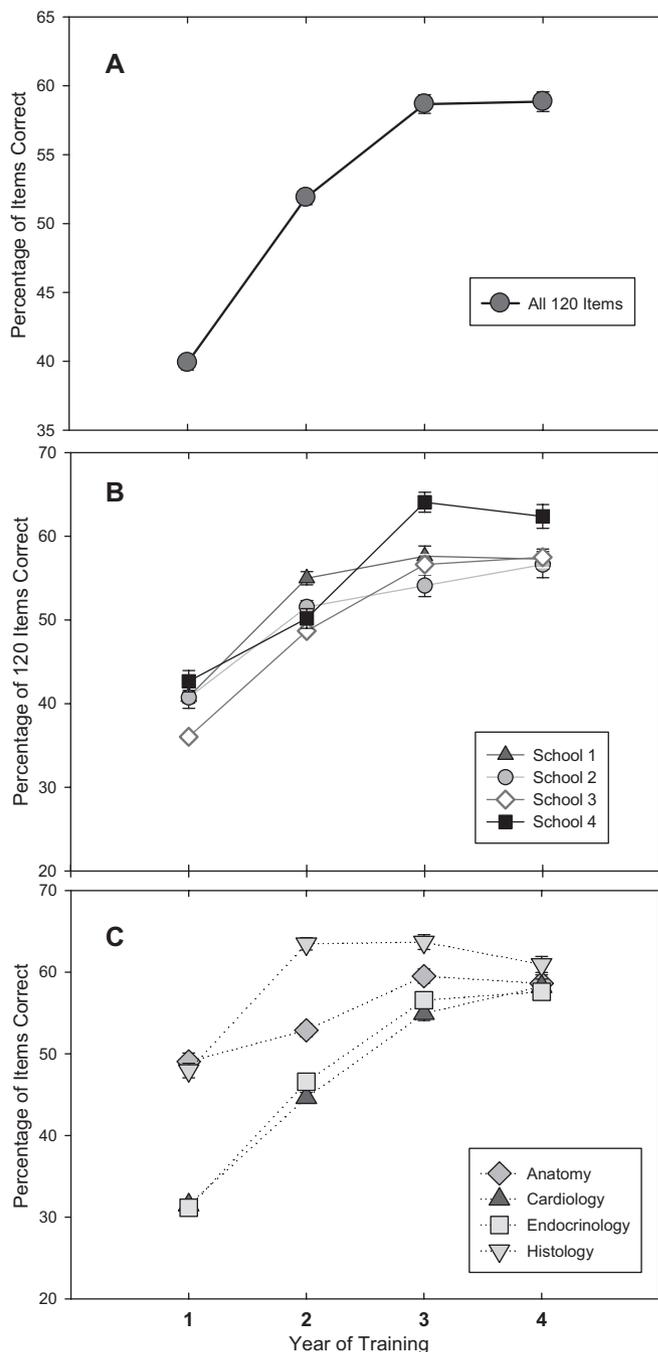


Figure 3 Progress-test performance overall (A), by medical school (B), and by topic (C). During weeks 1 through 30 of the trial, students at four medical schools were sent 30 validated questions in each of the following domains: anatomy, histology, cardiology, and endocrinology (120 questions total). Scores were calculated as the number of questions answered correctly normalized to a percentage scale. No data for students who submitted answers to <80% of the questions were included in the analyses. Error bars represent standard error.

Ideally, SEPT would be used as one part of an overall evaluation program to prospectively identify students who could benefit from remediation. Though performance characteristics of the progress-test may change when used as a summative rather than formative evaluation and as a compulsory rather than voluntary program, our results

indicate that the 120-item progress-test could be used, at a minimum, for moderate-stakes decisions for individual students. We estimate that a level of reliability sufficient for high-stakes assessment could be generated by a longitudinal progress-test of 180 questions or more.²⁹ This increased number of questions is unlikely to

produce “e-mail fatigue,” because current spaced education programs that deliver two questions every day are well accepted by students and physicians. Importantly, SEPT’s diagnostic characteristics (sensitivity, specificity, and predictive value) can be tailored to meet the needs of individual schools. If higher specificity and positive predictive value of SEPT were needed to identify those students most in need of remediation, the cutoff score could be lowered at the cost of reduced test sensitivity.

It is not clear why the progressive improvements in progress-test scores from years 1 through 3 were not sustained into year 4. This finding may reflect SEPT’s inability to capture the types of learning that may be occurring in year 4 in domains such as critical thinking, interpersonal skills, and professionalism. If this finding is replicated across other domains, year 4 may need to be restructured to enhance its educational efficacy.

The substantial improvements in learning retention generated in our trial by the spaced reinforcement of content are consistent with the results of prior studies and may have a distinct neurophysiologic basis.^{12,30,31} A recent study demonstrated that spaced learning by rats improves neuronal longevity in the hippocampus and that the strength of the rats’ memories correlates with the number of new cells in this region of their brains.¹⁰ In addition, the question-based format of the SEPT material may also reduce forgetting via the “testing effect,” the psychology research finding that the process of testing alters the learning process so that new knowledge is retained more effectively.^{32,33}

Our findings are strengthened by the use of a multiinstitutional randomized design and rigorous test-construction methodology. However, the results of the long-term learning assessment in weeks 31 to 34 should be interpreted with caution because the SEPT students had been exposed to the tested material six weeks more recently than had the LPTA students. Ideally, SEPT should be more inclusive than just four content domains, a change that may affect the reliability of the instrument. Given that the methods of question delivery were identical between study arms, the difference in attrition between cohorts

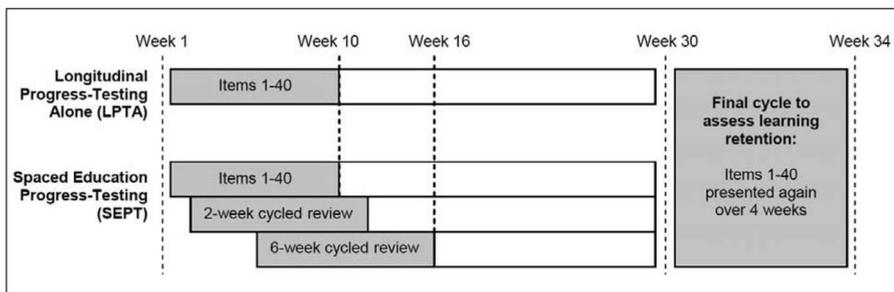
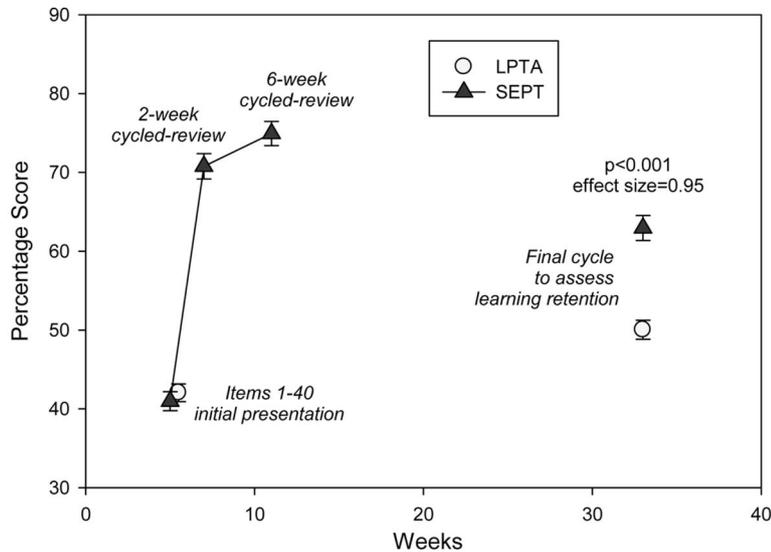


Figure 4 Impact of cycled reviews on long-term retention of learning. During weeks 31 through 34, the initial 40 progress-test questions were re-sent to students in both cohorts as an assessment of their long-term retention of the educational material. *Longitudinal progress-testing alone* (LPTA) students were sent these 40 items over weeks 1 through 10, in e-mails containing one question on each topic on Monday, Tuesday, Thursday, and Friday. *Spaced education progress-testing* (SEPT) students received the identical 40 items over weeks 1 through 10, but in addition they received two-week and six-week cycled reviews of the questions/explanations. For example, a question presented in week 1 was re-sent in week 3 (as a two-week cycled review) and in week 7 (as a six-week cycled review). Timeline is not drawn to scale. Error bars represent standard error.

is likely due to the fact that SEPT students received more than double the number of questions than did the LPTA students. Though our audit of self-reported summative examination scores suggests that most students correctly reported their scores, these self-reports could be subject to a reporting bias.

In summary, our study demonstrates that SEPT can act as an effective diagnostic tool to identify those students who could benefit from remediation and can substantially improve long-term retention of core knowledge. Further work is needed to establish SEPT as an effective tool for continuing medical education and/or for maintenance of certification. Because SEPT is content-

neutral, this methodology may have potential to improve assessment and long-term learning from grade school through higher education.

Dr. Kerfoot is associate professor, Surgical Service, Veterans Affairs Boston Healthcare System, Harvard Medical School, Boston, Massachusetts.

Dr. Shaffer is professor, Department of Radiology, Boston University School of Medicine, Boston, Massachusetts.

Dr. McMahon is assistant professor, Department of Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston, Massachusetts.

Dr. Baker is professor, Department of Psychology, California State University Channel Islands, Camarillo, California.

Dr. Kirdar is assistant professor, Medical Service, Veterans Affairs Boston Healthcare System, Harvard Medical School, Boston, Massachusetts.

Dr. Kanter is vice dean, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania.

Dr. Corbett is professor, Department of Medicine, University of Virginia School of Medicine, Charlottesville, Virginia.

Dr. Berkow is professor, Department of Pediatrics, University of Alabama School of Medicine, Birmingham, Alabama.

Dr. Krupat is director of evaluation and associate professor, Harvard Medical School, Boston, Massachusetts.

Dr. Armstrong is director of educational programs and clinical professor, Harvard Medical School, Boston, Massachusetts.

Acknowledgments: The authors recognize the invaluable work of Ronald Rouse, Jason Alvarez, and David Bozzi of the Harvard Medical School Center for Educational Technology for their development of the spaced education online delivery platform used in this trial. The authors also thank Drs. Daniel Federman, Susan Stearns, and Noelle Granger for their work on content validation.

Funding/Support: This study was supported by Harvard Medical International, the Harvard University Milton Fund, and the Harvard University Presidential Distance Learning Grant Program.

Other disclosures: Dr. Kerfoot is an equity owner and director of Spaced Education, Inc. None of the other authors have conflicts of interest.

Ethical approval: The study protocol received institutional review board approval.

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the position and policy of the United States Federal Government or the Department of Veterans Affairs. No official endorsement should be inferred.

References

- McMahon GT, Tallia AF. Perspective: Anticipating the challenges of reforming the United States medical licensing examination. *Acad Med.* 2010;85:453–456. http://journals.lww.com/academicmedicine/Abstract/2010/03000/Perspective_Anticipating_the_Challenges_of.18.aspx. Accessed November 23, 2010.
- Ling Y, Swanson DB, Holtzman K, Bucak SD. Retention of basic science information by senior medical students. *Acad Med.* 2008;83(10 suppl):S82–S85. http://journals.lww.com/academicmedicine/Fulltext/2008/10001/Retention_of_Basic_Science_Information_by_Senior.20.aspx. Accessed November 23, 2010.
- Custers EJ. Long-term retention of basic science knowledge: A review study. *Adv Health Sci Educ Theory Pract.* 2010;15:109–128.
- van der Vleuten CPM, Verwijnen GM. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Med Teach.* 1996;18:103–109.
- Blake JM, Norman GR, Keane DR, Mueller CB, Cunningham J, Didyk N. Introducing progress testing in McMaster University's problem-based medical curriculum: Psychometric

- properties and effect on learning. *Acad Med.* 1996;71:1002–1007. http://journals.lww.com/academicmedicine/Abstract/1996/09000/Introducing_progress_testing_in_McMaster.16.aspx. Accessed November 23, 2010.
- 6 Arnold L, Willoughby TL. The quarterly profile examination. *Acad Med.* 1990;65:515–516. http://journals.lww.com/academicmedicine/Abstract/1990/08000/The_quarterly_profile_examination.5.aspx. Accessed November 23, 2010.
 - 7 van der Vleuten CP, Schuwirth LW, Muijtjens AM, Thoben AJ, Cohen-Schotanus J, van Boven CP. Cross institutional collaboration in assessment: A case on progress testing. *Med Teach.* 2004;26:719–725.
 - 8 Bjork RA. Retrieval practice and the maintenance of knowledge. In: Gruneberg MM, Morris PE, Sykes RN, eds. *Practical Aspects of Memory: Current Research and Issues*. New York, NY: John Wiley & Sons; 1988:396–401.
 - 9 Pashler H, Rohrer D, Cepeda NJ, Carpenter SK. Enhancing learning and retarding forgetting: Choices and consequences. *Psychon Bull Rev.* 2007;14:187–193.
 - 10 Sisti HM, Glass AL, Shors TJ. Neurogenesis and the spacing effect: Learning over time enhances memory and the survival of new neurons. *Learn Mem.* 2007;14:368–375.
 - 11 Cepeda NJ, Vul E, Rohrer D, Wixted JT, Pashler H. Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychol Sci.* 2008;19:1095–1102.
 - 12 Kerfoot BP. Learning benefits of on-line spaced education persist for 2 years. *J Urol.* 2009;181:2671–2673.
 - 13 Kerfoot BP, Armstrong EG, O'Sullivan PN. Interactive spaced-education to teach the physical examination: A randomized controlled trial. *J Gen Intern Med.* 2008;23:973–978.
 - 14 Kerfoot BP, Kearney MC, Connelly D, Ritchey ML. Interactive spaced education to assess and improve knowledge of clinical practice guidelines: A randomized controlled trial. *Ann Surg.* 2009;249:744–749.
 - 15 Kerfoot BP, Lawler EV, Sokolovskaya G, Gagnon D, Conlin PR. Durable improvements in prostate cancer screening from online spaced education a randomized controlled trial. *Am J Prev Med.* 2010;39:472–478.
 - 16 Landauer TK, Bjork RA. Optimum rehearsal patterns and name learning. In: Gruneberg MM, Morris PE, Sykes RN, ed. *Practical Aspects of Memory*. New York, NY: Academic Press; 1978:625–632.
 - 17 Enders CK. The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educ Psychol Meas.* 2004;64:419–436.
 - 18 Enders CK. Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychol Methods.* 2003;8:322–337.
 - 19 Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika.* 1951;16:297–334.
 - 20 Pedhazur EJ, Schmelkin LP. *Measurement, Design, and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum; 1991.
 - 21 Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Erlbaum; 1988.
 - 22 Maxwell SE, Delaney HD. *Designing Experiments and Analyzing Data: A Model Comparison Approach*. Belmont, Calif: Wadsworth; 1990.
 - 23 Basco WT Jr, Way DP, Gilbert GE, Hudson A. Undergraduate institutional MCAT scores as predictors of USMLE step 1 performance. *Acad Med.* 2002;77(10 suppl):S13–S16. http://journals.lww.com/academicmedicine/Fulltext/2002/10001/Undergraduate_Institutional_MCAT_Scores_as.5.aspx. Accessed November 23, 2010.
 - 24 Mitchell KJ. Traditional predictors of performance in medical school. *Acad Med.* 1990;65:149–158. http://journals.lww.com/academicmedicine/Abstract/1990/03000/Traditional_predictors_of_performance_in_medical.5.aspx. Accessed November 23, 2010.
 - 25 Silver B, Hodgson CS. Evaluating GPAs and MCAT scores as predictors of NBME I and clerkship performances based on students' data from one undergraduate institution. *Acad Med.* 1997;72:394–396. http://journals.lww.com/academicmedicine/Abstract/1997/05000/Evaluating_GPAs_and_MCAT_scores_as_predictors_of.22.aspx. Accessed November 23, 2010.
 - 26 Paolo AM, Bonaminio GA, Durham D, Stites SW. Comparison and cross-validation of simple and multiple logistic regression models to predict USMLE step 1 performance. *Teach Learn Med.* Winter 2004;16:69–73.
 - 27 Lieberman DA. Screening for colorectal cancer. *N Engl J Med.* 2009;361:1179–1187.
 - 28 Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: A summary of the evidence for the U.S. Preventive Services Task Force. *Ann Intern Med.* 2002;137(5 part 1):347–360.
 - 29 Aiken LR. *Psychological Testing and Assessment*. 10th ed. Boston, Mass: Allyn and Bacon; 2000.
 - 30 Kerfoot BP, Brotschi E. Online spaced education to teach urology to medical students: A multi-institutional randomized trial. *Am J Surg.* 2009;197:89–95.
 - 31 Kerfoot BP, DeWolf WC, Masser BA, Church PA, Federman DD. Spaced education improves the retention of clinical knowledge by medical students: A randomised controlled trial. *Med Educ.* 2007;41:23–31.
 - 32 Roediger HL, Karpicke JD. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychol Sci.* 2006;17:249–255.
 - 33 Karpicke JD, Roediger HL 3rd. The critical importance of retrieval for learning. *Science.* 2008;319:966–968.