Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial

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

CONTEXT Laboratory studies in cognitive psychology with relatively brief final recall intervals suggest that repeated retrieval in the form of tests may result in better retention of information compared with repeated study.

OBJECTIVES Our study evaluates if repeated testing of material taught in a real-life educational setting (a didactic conference for paediatric and emergency medicine residents) replicates these findings when measured at a more educationally relevant final recall interval of 6 months.

METHODS Residents participated in an interactive teaching session on two topics: (i) status epilepticus, and (ii) myasthenia gravis. Residents were randomised to two counter–balanced groups which either took tests on status epilepticus and studied a review sheet on myasthenia gravis (SE-T/MG-S group) or took tests on myasthenia gravis and studied a review sheet on status epilepticus (MG-T/SE-S group). Testing and studying occurred immediately after teaching and then at two additional times at intervals of about 2 weeks. Residents received feedback after each test. Tests consisted of short-answer questions and the review sheets consisted of information identical to that on the answer sheets for the tests. At about 6 months residents took a final test on both topics.

RESULTS Nineteen residents in the SE-T/MG-S group and 21 residents in the MG-T/SE-S group completed the study. Collapsing across groups, repeated testing produced final test scores that were an average of 13% higher than those produced by repeated study (39% versus 26%) at > 6 months after the initial teaching session ($t[78] = 3.93$, standard error of the difference = 0.03, $P < 0.001$, d = 0.91).

CONCLUSIONS Repeated testing with feedback appears to result in significantly greater long-term retention of information taught in a didactic conference than repeated, spaced study. Testing should be considered for its potential impact on learning and not only as an assessment device.

*Medical Education 2009: **43**: 1174–1181*
doi:10.1111/j.1365-2923.2009.03518.x

[1]Department of Neurology, Washington University in St Louis, St Louis, Missouri, USA
[2]Department of Psychology, Washington University in St Louis, St Louis, Missouri, USA

Correspondence: Douglas P Larsen, Department of Neurology, Washington University, 660 South Euclid Avenue, Campus Box 8111, St Louis, Missouri 63110, USA. Tel: 00 1 314 454 6120; Fax: 00 1 314 454 2523; E-mail: larsend@neuro.wustl.edu

INTRODUCTION

Factual knowledge plays a critical role in the development of clinical expertise.[1] In order to increase the development of this knowledge base and to standardise the educational experience, residency programmes have developed curricula that typically consist of a series of didactic conferences that cover topics fundamental to resident training. Indeed, in the USA, the Accreditation Council for Graduate Medical Education mandates this type of curriculum.[2] Despite these efforts, multiple studies have shown no difference in the long-term retention of knowledge on various assessment measures between residents who attend didactic conferences and those who do not.[3–7] A remedy that has been suggested in this literature is to make conferences more interactive and clinically relevant.[4]

Increasing interaction and clinical relevance in the conferences, however, may not be sufficient because these interventions focus only on improving initial learning. Research in cognitive psychology suggests that additional processing of information after initial learning is important for long-term retention.[8] Specifically, repeated practice in retrieving information from memory seems to greatly enhance future recall. This retrieval practice often takes the form of tests. Although tests have traditionally been used for assessment tools, this line of research demonstrates that the actual act of taking tests augments retention. We have more thoroughly reviewed this concept of test-enhanced learning elsewhere.[9] Briefly, studies conducted mostly in laboratory settings show that repeated testing produces superior retention relative to repeated study over time periods of 1–6 weeks.[10–13] More recently, a single test in a medical skills training setting has been shown to improve retention after 2 weeks compared with standard instruction and training.[14] However, more research is needed to clarify how the mnemonic benefits of testing translate into real-life educational settings and whether testing improves retention over longer intervals that would be more educationally relevant.

Our current study is a randomised controlled trial to evaluate the effect of repeated testing on the retention of information taught in a resident conference setting with a retention interval of > 6 months. We hypothesised that repeated testing would result in better long-term retention of the information than repeated study.

METHODS

Subjects

Residents from the paediatrics and emergency medicine programmes affiliated with our institution served as subjects. They were recruited from attendees at a didactic conference on paediatric neurological emergencies that was taught as part of the core curriculum of both residency programmes. Residents were from all 3 years of the paediatric residency programme and all 4 years of the emergency medicine programme. The study was approved and granted exempt status as an educational study by the institutional review board at our institution. All residents gave voluntary informed consent before participating in the study.

Materials

Tests and review sheets

The essential clinical content that residents would be expected to know long-term was identified and experimental materials (e.g. tests, review sheets, etc.) based on this content were created. Two neurological situations that might be encountered in an emergency department were covered: the treatment of status epilepticus, and the diagnosis and treatment of myasthenia gravis. Clinical material was based on standard treatments and guidelines at our institution. Tests consisted entirely of questions in which residents were asked to generate answers from memory in a short-answer format. Many questions required the generation of multiple pieces of information. The status epilepticus test and myasthenia gravis test required the recall of 39 and 40 pieces of information, respectively. Review sheets contained information identical to that on the answer sheets for the tests on the same material. Tests and review sheets were identical for all testing and study events.

Questionnaires

In order to obtain additional information about the residents, two questionnaires were created, one for the initial test and the other for the final test. The initial test questionnaire asked residents to rate the quality of the teaching session, the usefulness of the information, the degree of interaction in the conference, and their level of prior knowledge about the topic. These questions were answered using a 9-point Likert scale. Questions covered each topic separately. Residents were also asked their opinion on

participating in regular tests as part of an educational programme.

On the final questionnaire, residents reported how many times they had used the information clinically between the time of the initial teaching and the final test. They also described any adverse circumstances during the initial teaching or subsequent tests that may have affected their learning or performance (e.g. being post-call, having been paged out of the conference, etc.). In order to assess how repeatedly taking tests influenced residents' opinions of testing, residents were again asked about their willingness to take tests on a regular basis.

Teaching sessions

Once the objectives and tests had been created, a 1-hour teaching session was designed to cover both topics. The time was evenly divided between the topics. The teaching session covered exactly the same information as the tests. The session was designed to be highly interactive and involved residents in activities such as creating drawings, practising elements of the examination of myasthenia patients on one another, and using cards with parts of the status epilepticus treatment protocol to complete a chart.

Procedure

The same teaching session was delivered on three separate occasions, once for the paediatric residency programme, once for the emergency medicine residency, and once as a make-up session for paediatric residents who had missed the first conference. One of the authors (DPL) taught all of the conferences and administered the tests. All conferences were taught using the same techniques and teaching plan. Residents were assigned to one of two groups in alternating fashion as they entered the teaching session. Residents were either assigned to be repeatedly tested over status epilepticus and repeatedly study myasthenia gravis (SE-T/MG-S group) or to be repeatedly tested over myasthenia gravis and repeatedly study status epilepticus (MG-T/SE-S group). Residents did not know to which group they had been assigned until they took their initial test at the end of the teaching session. At the end of the teaching session, residents could choose whether or not to participate in the study.

At the end of the teaching session, residents either took a test on the treatment of status epilepticus and studied a review sheet on myasthenia gravis (SE-T/MG-S group) or took a test on myasthenia gravis and studied a review sheet on the treatment of status epilepticus (MG-T/SE-S group). The tests took approximately 10–15 minutes to complete. After finishing the test, residents were given the answer sheet to compare with their responses. Residents were not allowed to keep the answer sheets or review sheets for further study. Residents also completed the initial questionnaire at the end of the teaching session.

At approximately 2-week intervals (16.8 days on average), residents took two additional tests on the same topic and studied the review sheet on the other topic. The same tests and review sheets were used in these follow-up sessions. After each follow-up test, residents again compared their responses with the answer sheet. Approximately 6 months after the initial teaching session (6.7 months on average), residents took a final test (using identical tests to the follow-up tests) on both topics. Residents also completed the final questionnaire at that time.

Two members of the research team (DPL and ACB) independently scored the tests. Discrepancies were resolved by consensus. Scoring of both tests showed very high inter-rater reliability. Cohen's kappa was 0.98 for the myasthenia gravis test and 0.97 for the status epilepticus test.

Statistical analyses

Final test scores were compared using independent *t*-tests between the repeated-test and repeated-study groups by topic and collapsing across topics. *P*-values of < 0.05 were considered significant. Within-subject score differences on repeated tests for each topic were evaluated by repeated-measures analysis of variance followed by post hoc *t*-tests corrected for multiple comparisons. All effect sizes were calculated using Cohen's *d*. Independent *t*-tests were used to compare the differences in prior knowledge between groups. Proportions of clinical exposures after the teaching session and adverse conditions were analysed using a chi-squared test. Because our sample sizes were small, these analyses were performed in order to verify that there were no differences between groups in these factors that might influence the outcome. All statistical analyses were performed using R statistical software (R Foundation, Vienna, Austria) and SPSS Version 16.0 (SPSS, Inc., Chicago, IL, USA).

RESULTS

Subjects

A total of 65 residents (29 in paediatrics and 36 in emergency medicine) attended the three teaching sessions. Of these, 44 residents (68%) agreed to participate in the study, including 23 residents from the MG-T/SE-S group and 21 from the SE-T/MG-S group. Two residents in each group dropped out during the course of the study. Data pertaining to 21 residents from the MG-T/SE-S group and 19 from the SE-T/MG-S group were included in the final analyses.

Effect of testing on retention

Initial learning was moderately high as measured by the test immediately after the teaching session (Test 1). Residents in the repeated testing groups scored an average of 78% correct answers on the status epilepticus test and 62% correct answers on the myasthenia gravis test. A sharp drop in recall occurred approximately 2 weeks later on the same material (Test 2). Residents scored 47% and 44% on the status epilepticus and myasthenia gravis tests, respectively. This was followed by a small and statistically insignificant increase in scores on Test 3, presumably because residents were given feedback after the previous test. Residents scored 55% and 48% on the status epilepticus and myasthenia gravis tests, respectively (Fig. 1).

Approximately 6–7 months after the initial teaching session residents were given a final test on both topics using information identical to that used in the previous tests. On the final status epilepticus test, residents in the SE-T/MG-S group scored on average 11% higher (42% versus 31%) than residents in the MG-T/SE-S group ($t[38] = 2.07$, standard error of the difference [SED] = 0.05, $P < 0.05$). This difference gave a moderate effect size of 0.65. On the final myasthenia gravis test, residents in the MG-T/SE-S group scored on average 17% higher (36% versus 19%) than residents in the SE-T/MG-S group ($t[39] = 4.13$, SED = 0.04, $P < 0.001$). This difference gave a very large effect size of 1.30. Collapsing across groups, repeated testing produced a final test score which was an average of 13% higher than that of repeated study (39% versus 26%) ($t[78] = 3.93$, SED = 0.03, $P < 0.001$). The overall effect size was large at 0.91.

Questionnaires

Initial and final questionnaires were used to better understand factors that may have influenced our results. All questions reported below were assessed in the questionnaires using a 9-point Likert scale. Data for three residents were missing from their responses to questions about the myasthenia gravis topic on the initial questionnaire.

Residents rated the teaching session overall very highly. They felt that the information was very

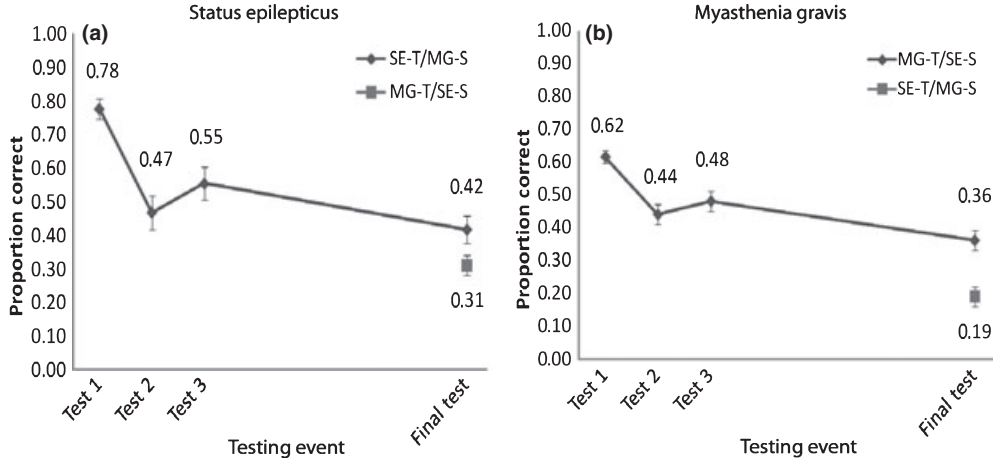

Figure 1 Graphical representation of averaged test results for the (a) status epilepticus and (b) myasthenia gravis tests over approximately 6 months (error bars represent standard errors of the mean). Test 1 was administered immediately following the teaching session. Tests 2 and 3 were administered at approximately 2-week intervals. The final test was given at approximately 6 months after the teaching session. SE-T/MG-S = status epilepticus test/myasthenia gravis study group; MG-T/SE-S = myasthenia gravis test/status epilepticus study group

Table 1 Responses to items on the questionnaire on the initial teaching session

Category	Likert scale	Average resident response (status epilepticus) (n = 40)	Average resident response (myasthenia gravis) (n = 37*)
Overall rating of teaching session	1 = poor 9 = excellent	7.98	8.03
Usefulness of information	1 = not useful at all 9 = very useful	8.35	8.00
Degree of participation	1 = not at all 9 = very active	7.55	7.32
Amount of prior knowledge	1 = no prior knowledge 9 = extremely knowledgeable	5.79	5.92

* Three residents did not provide responses on the myasthenia gravis topic

useful and that they had actively participated in the learning process. They reported a fair level of prior knowledge of the subject matter. There was no significant difference in prior knowledge between groups. Table 1 shows specific results by topic.

In order to assess the role of clinical exposure after the teaching session on our results, residents were asked on the final questionnaire how many times they had used the information on the two topics during the course of the study. Table 2 shows the number of residents reporting post-teaching clinical application by group and topic. None of the differences in clinical exposure were significant except for the difference in exposures to myasthenia gravis and status epilepticus in the SE-T/MG-S group ($\chi^2[1] = 6.69$, $P = 0.01$).

Ten out of 16 residents (63%) in the SE-T/MG-S group (data were missing for three residents) reported adverse events during either the initial teaching session or subsequent tests that may have affected their performance. These included such events as being paged out of the teaching session, attending a session after being on overnight call and feeling pressed for time. In the MG-T/SE-S group, 13 of 18 residents (72%) reported similar adverse events (data were missing for three residents). The difference between groups was not significant ($\chi^2[1] = 0.06$, $P = 0.81$).

On the final questionnaire, when residents were asked if they had studied prior to any of the tests, three residents in the SE-T/MG-S group responded that they had. None of the residents in the MG-T/SE-S group had studied before any of the tests.

Table 2 Number of residents reporting clinical application experiences by topic during the course of the study

	Experience with the status epilepticus material	Experience with the myasthenia gravis material	Significance
SE-T/MG-S (n = 19)	12* (67%)	3† (18%)	P = 0.01
MG-T/SE-S (n = 21)	11 (52%)	7 (33%)	NS
Significance	NS	NS	

* Data missing for one resident
† Data missing for two residents
SE-T/MG-S = status epilepticus test/myasthenia gravis study group; MG-T/SE-S = myasthenia gravis test/status epilepticus study group; NS = not significant

On the initial questionnaire 91% of residents who responded (seven residents did not answer the question) said that they would be willing to take tests on a regular basis as part of an educational programme. On the final questionnaire 87% of residents (data were missing for one resident) expressed willingness to take tests. Any change between the initial and final questionnaires is difficult to interpret because 18% of the participants did not give an initial response.

DISCUSSION

The results of the present study show that repeated testing produced better long-term retention of information learned from a didactic conference relative to repeated, spaced study. In the cognitive psychology literature this result is known as the 'testing effect'. Our experiment shows that this phenomenon, which has been mostly demonstrated in laboratory settings, can also be applied to a real-life educational situation. The testing effect is sufficiently robust that we were able to demonstrate a significant benefit with a relatively small sample size. In addition, most previous studies of the testing effect have shown an increase in retention over 1–6 weeks after initial learning.[10–14] Our study shows improved recall at a more educationally relevant time interval of > 6 months.

In our study, residents showed a moderately high level of initial learning and felt that they had actively engaged in the learning process. Residents felt that the information was highly useful and well taught. Despite optimising initial learning, residents experienced significant forgetting over the course of the 6–7 months of the study. The results of our repetitive tests over time are similar to typical forgetting curves. Forgetting curves usually show a steep decline in retention shortly after initial learning and a more gradual drop over long periods of time[15] (Fig. 1). Our study, however, demonstrates that residents who were repeatedly tested forgot significantly less than those who repeatedly studied the material.

It should be noted that the activities of our repeated study control groups differed from typical resident practice. In our experience, residents do not typically engage in focused study of the information taught in didactic sessions, yet in this case our control group received spaced study experiences that doubtless improved recall over normal levels. Our procedure would tend to underestimate the mnemonic gains to be made from repeated testing in normal circumstances. Nonetheless, our results demonstrate that long-term retention is significantly enhanced by retrieval practice via repeated testing. The overall effect size of repeated testing in our study is large, at 0.91, especially given that the average effect size of most educational interventions is 0.50.[16] Our overall effect size is consistent with those in other studies of the testing effect.[12,13]

The decline in performance between the first and second tests raises questions about whether a different testing schedule might improve long-term retention. In a knowledge-based educational setting, a shorter interval between follow-up tests might substantially reduce forgetting and better maintain gains in knowledge from initial learning (see studies by Landauer and Bjork[17] and Karpicke and Roediger[18] for more on testing intervals). In their study of testing effects in a resuscitation skills course, Kromann *et al.*[14] did not find a drastic decrease in retention after 2 weeks; however, there may be differences in the rate of decay of motor skills versus factual knowledge. The question of whether an increased number of tests with feedback would also help to more fully reverse the effects of initial forgetting remains to be investigated. These issues will need to be explored in future research.

Our study has several limitations. Firstly, in the repeated study groups, we did not have a mechanism to measure how thoroughly residents studied the review sheet. We were unable to use methods that can quantify reading, such as having residents fill in blanks where words have been left out, because the retrieval this requires begins to approximate our test group and would have confounded our results. However, because residents do not typically study materials from didactic conferences, any amount of study is more exposure than they would typically receive. Again, our study design would tend to underestimate the effects of testing in typical resident education conditions.

Secondly, although repeated testing produced significantly better retention for both topics (status epilepticus and myasthenia gravis), there were differences in the testing effect between the two topics. There was greater separation between the test and study groups for myasthenia gravis and the effect size was much larger compared with that for the status epilepticus topic. The reasons for this are not entirely clear. This may be because myasthenia gravis is a rarer clinical condition with which residents may not have as much experience and therefore they may have benefited more from the repeated testing.

Alternatively, intrinsic characteristics of the tests themselves may also have contributed to the differences in testing effect. Although both tests required recall of almost the same number of information items, they had different components dictated by the different subject matter. The status epilepticus test had more questions requiring recall of a single item of information. This would allow for more memory cueing from the questions themselves. The myasthenia gravis test had more questions that required multiple pieces of information to be listed, which would require greater effort to recall.

Thirdly, our questionnaires may have been subject to response biases. Residents may have inflated their estimations of the quality and value of the teaching, as well as their willingness to take repeated tests in order to please the investigators. However, the investigators were not involved in evaluating the residents and therefore intent to please the investigators would be minimised. Additionally, residents' reports of clinical exposures and adverse events would be subject to recall bias. However, the trend of reporting more exposure to status epilepticus than to myasthenia gravis is consistent with the epidemiology of the two conditions. In addition, the frequency of adverse events such as being deprived of sleep and being paged out of conference at some point in the course of the study seems high but reasonable for a population of residents.

Finally, although we did show increased recall with repeated written tests, it is unclear whether this increased retention would transfer to clinical settings. Although we hope that information taught in conferences and retained through written tests will be applied by residents in their clinical work, this issue has not been investigated to our knowledge. If the processes used to learn information are very different from the processes used to retrieve the information, retention may be less likely. Future studies will need to investigate the degree of transfer from didactic settings and written tests to clinical applications.

Although much of educational practice and research is devoted to maximising initial learning and creating valid final assessment measures, our research directs attention to the critical nature of post-learning activities in terms of long-term retention. Our study shows that tests can be used not only for assessment, but that they also have a direct effect on retention. Medical educators sometimes avoid tests because of fears that learners will be unwilling to take repeated tests (in addition to the fact that they have to be graded). This fear appears unfounded. In our study, the vast majority of residents expressed willingness to take regular tests. Further research is needed to identify how testing can be further optimised to produce even greater gains in long-term retention of information. Future studies will also need to demonstrate that the increased knowledge retained through testing can be transferred to patient care situations. These studies should begin to give us an additional tool to help bridge the gap between initial learning and the final outcomes we seek as educators.

Contributors: DPL developed the study concept. All authors contributed to the study design. Teaching and data collection were performed by DPL. DPL and ACB performed the data analysis and interpretation under the supervision of HLR. DPL drafted the paper, and ACB and HLR provided critical revisions. All authors approved the final version of the paper for submission.

Acknowledgements: we would like to thank the residents who participated in the study for their time and effort.
Funding: DPL receives support from the McDonnell Center for Systems Neuroscience at the Washington University School of Medicine. HLR receives grant support from the James S McDonnell Foundation. HLR is also supported by the Institute of Education Sciences, US Department of Education, through grant R305H060080-06 to Washington University in St Louis. The opinions expressed are those of the authors and do not represent the views of the Institute or the US Department of Education.

Conflicts of interest: none.

Ethical approval: the Institutional Review Board of Washington University in St Louis approved exempt status for this research as an educational study.

REFERENCES

1 van der Vleuten CPM, Newble DI. How can we test clinical reasoning? *Lancet* 1995;**345**:1032–4.
2 Accreditation Council for Graduate Medical Education. *Common Program Requirements*, Chicago, IL: ACGME 2007. http://www.acgme.org/acWebsite/dutyHours/dh_dutyhoursCommonPR07012007.pdf. [Accessed 30 October 2009.]
3 Picciano A, Winter R, Ballan D, Bimberg B, Jacks M, Laing E. Resident acquisition of knowledge during a noontime conference series. *Fam Med* 2003;**35** (6):418–22.
4 Winter RO, Picciano A, Bimberg B, Chae M, Chae S, Jacks M, Metz J, Milne C. Resident knowledge acquisition during a block conference series. *Fam Med* 2007;**39** (7):498–503.
5 Fitzgerald JD, Wenger NS. Didactic teaching conferences for IM residents: who attends, and is attendance

related to medical certifying examination scores? *Acad Med* 2003;**78** (1):84–9.
6 Cacamese SM, Eubank KJ, Hebert RS, Wright SM. Conference attendance and performance on the in-training examination in internal medicine. *Med Teach* 2004;**26** (7):640–4.
7 Pollack R, Baker RJ. The acquisition of factual knowledge and the role of the didactic conference in a surgical residency programme. *Am Surg* 1988;**54** (9):531–4.
8 Roediger HL III, Karpicke JD. The power of testing memory: basic research and implications for educational practice. *Perspect Psychol Sci* 2006;**1**:181–210.
9 Larsen DP, Butler AC, Roediger HL III. Test-enhanced learning in medical education. *Med Educ* 2008;**42**:959–66.
10 Karpicke JD, Roediger HL III. The critical importance of retrieval for learning. *Science* 2008;**319** (5865):966–8.
11 Butler AC, Roediger HL III. Testing improves long-term retention in a simulated classroom setting. *Eur J Cogn Psychol* 2007;**19** (4/5):514–27.
12 Roediger HL III, Karpicke JD. Test-enhanced learning: taking memory tests improves long-term retention. *Psychol Sci* 2006;**17**:249–55.
13 McDaniel MA, Anderson JL, Derbish MH, Morrisette N. Testing the testing effect in the classroom. *Eur J Cogn Psychol* 2007;**19** (4/5):494–513.
14 Kromann CB, Jensen ML, Ringsted C. The effect of testing on skills learning. *Med Educ* 2009;**43**:21–7.
15 Nairne JS, Pandeirada JNS. Forgetting. In: Roediger HL III, ed. Byrne J, ed-in-chief. *Cognitive Psychology of Memory. Volume II of Learning and Memory: A Comprehensive Reference*. Oxford: Elsevier 2008;179–94.
16 Albanese M. Problem-based learning: why curricula are likely to show little effect on knowledge and clinical skills. *Med Educ* 2000;**34**:729–38.
17 Landauer TK, Bjork RA. Optimum rehearsal patterns and name learning. In: Gruneberg MM, Morris PE, Sykes RN, eds. *Practical Aspects of Memory*. London: Academic Press 1978;625–32.
18 Karpicke JD, Roediger HL III. Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *J Exp Psychol Learn Mem Cogn* 2007;**33** (4):704–19.

Received 26 February 2009; editorial comments to authors 9 April 2009; accepted for publication 27 July 2009