AC 2011-1428: PRELIMINARY ANALYSIS OF REPEATED TESTING AS A METHOD TO ENHANCE LONG-TERM RETENTION OF KNOWLEDGE

Paul M. Santi, Colorado School of Mines

Paul Santi is a professor of Geology and Geological Engineering at the Colorado School of Mines. He has 16 experience teaching at the university level and 6 years experience in the geotechnical and environmental consulting industry. He obtained a B.S. in Geology and Physics from Duke University, an M.S. in Geology from Texas A&M University, and a Ph.D. in Geological Engineering from the Colorado School of Mines. His research areas include ways to enhance learning in the field of geological engineering, as well as understanding and mitigation of natural hazards.

Preliminary Analysis of Repeated Testing as a Method to Enhance Long-Term Retention of Knowledge

Introduction

One of the primary goals of education, along with developing a deep understanding of material, is enhancing long-term retention of principles, processes, vocabulary, and factual information. Unfortunately, we are not always effective at meeting this goal: studies have shown that typically 90% of material is lost within 30 days of learning it[1].

Repetition has been shown to be a successful method of improving retention[2, 3, 4]. Physiologically, repetition has the effect of establishing permanent neuron pathways that allow access to (and thereby, recall of) the data, repetition recruits more neurons so the pathway is wider, and the environmental factors surrounding the learning experience result in development of multiple pathways to the data, enriching the ability to access the material[5]. Furthermore, studies have shown that repetition should occur within deliberately spaced intervals to best establish neuron pathways[6, 7].

It is suspected that the lack of repetition, especially within the critical time frame, is a weakness in the typical university curriculum, and as a result, long-term retention suffers. Compartmentalization of material between courses and reluctance to spend limited class time reviewing previously-taught material mean that students are only exposed to some important topics a single time.

The goal of this study is to demonstrate that repeated learning of material during a single semester improves retention when tested 16-18 months later. Repeated learning will be achieved by the simple process of repeated testing. Long-term retention will be measured by re-testing after three semesters have passed. This paper presents preliminary baseline data, as results from testing the long-term retention of groups who were repeatedly tested are not yet available (these results will be presented at the meeting).

Relation to present state of knowledge in the field

Learning involves a multi-stage process. First, material is encoded, in which information is fragmented and distributed to various areas in the brain and associated with the learning experience, setting, and previous information. Then it is stored, where neuron links are temporarily established. Later, it may be retrieved, as links are accessed[8]. The problem with long-term retention is that links will fade with time unless the learning experience is emotionally strong[9] or elaborately and deeply encoded, perhaps through repetition[8]. This concept is relied on for licensing of medical doctors, who must pass a series of exams over a course of approximately three years, each of which requires more sophisticated recall and application of the body of knowledge learned through medical school and early residency.

The benefits of repeated testing for retention were demonstrated by Roediger and Karpicke[10, 11], who showed that students who were tested repeatedly on material scored higher on later retention tests than students who studied repeatedly and were only tested once. This improvement has been termed the "testing effect" and has been shown by other research studies as well[12, 13, 14]. The type of recall required by testing is different and more intense than the type of recall experienced by repeat-reading or by studying in a relaxing environment. The intensity induced by the testing environment entrenches neuron links, thereby enhancing retention of information. A recent paper suggests that students who repeatedly read their notes ("cramming") experience "illusions of competence" but do not engage in the critical skills of retrieval of information that testing develops[11]. In one test, the study-only group read a sample passage an average of 14 times, recalling 40% a week later, while the repeat-testing group read the passage an average of 3.4 times, yet recalled 61% after a week. Interestingly, the study-only group performed better than the repeat-testing group on tests given five minutes after the last study session[10].

These published studies have two significant differences from the preliminary research reported here. First, the Roediger and Karpicke studies tested retention at intervals of five minutes, two days, and one week, which really is only short-term review and recall. Other retrieval and memory research also relies on these short time frames[6, 12, 15, 16]. These studies consider recall periods on the order of two days to be "long term"[17]. Two exceptions are Butler and Roediger[18], who tested material from three lectures one month after the instruction, and Bahrick[19], who reports on recall of high school algebra and Spanish vocabulary over a period of one to 50 years. Repeat testing was not part of either study. We conducted repeated testing over an entire semester, with a final retention test 16-18 months later. A second distinction is that these previous studies involved clinical rather than classroom teaching settings, with a very limited breadth of study material; for example, the Roediger and Karpicke studies use vocabulary word pairs or testing on single prose passages. We tested actual classroom teaching with a full range of course material covered in a semester. Consequently, this is a set of field trials of their narrowly-defined clinical studies.

Other measurements of knowledge-base across an entire field, such as licensure exams or even comprehensive degree examinations do not measure long-term retention of knowledge. Students study for these exams, so the testing measures their ability to retain a larger amount of information, but only over the short time frame between studying and taking the exam.

Tested groups

The course used in this study, GEGN 202 – Geologic Principles and Processes, has had a recent enrollment of 60-70, including all sophomores in both Geological Engineering and Geophysical Engineering (Table 1). All students are tested in the class, but only the geological engineers are tracked for long-term retention testing. Traditionally, this fall semester course has included two non-cumulative exams (in addition to two late-semester exams on topics not tracked for this study). More recently, the course was modified to include six exams in 2009 and five exams in 2010, all of which were cumulative, covering all material already introduced in the course. Consequently, some topics were tested as many as six times, later material tested five times, and so on, with the material at the end of the semester tested only once.

Table 1. Summary of tested groups.

Group	Date of enrollment in GEGN 202	In-class testing on material tracked for retention	Date of retention exam
2007 control group (N=25)	Fall 2007	2 exams, not cumulative	April 2009
2008 control group (N=38)	Fall 2008	2 exams, not cumulative	March 2010
2009 test group (N=50)	Fall 2009	6 cumulative exams	April 2011 (expected)
2010 test group (N=50)	Fall 2010	6 cumulative exams	April 2012 (expected)

A long-term retention exam was prepared, consisting of 30 multiple-choice questions, divided evenly among the general course topics. The retention exam was administered to two cohorts of students (Table 1): those who had taken the course in Fall 2007 and were tested in April 2009 (25 students in the "2007 control group"), and those who had taken the course in Fall 2008 and were tested in March 2010 (38 students in the "2008 control group"). Scores from these two control groups are the baseline levels representing typical retention of information over a period of approximately 18 months. The two control groups are the baseline against which we expect to demonstrate improved scores as a result of repeated testing during the original class.

In addition to the two control groups, two test groups will represent students who were tested more frequently during the semester. The retention test will be given to Fall 2009 students in April 2011 (40 students in the "2009 test group") and the Fall 2010 students in April 2012 (50 students in the "2010 test group") and their scores will be compared to the control groups to demonstrate the expected improvement in retention.

Baseline 18-month retention exam

The 30 questions in the long-term retention exam were divided among the 10 topical categories covered in the class. The number of questions in each category reflects the amount of lecture time spent on that category. The topical categories, in chronological order of presentation, include:

1. Glacial (4 questions)
2. Periglacial (1 question)
3. Volcanic and igneous (2 questions)
4. Climate, weathering and soils (3 questions)
5. Slopes (3 questions)
6. Folding and faulting (2 questions)
7. Drainage basins and fluvial (5 questions)
8. Arid (3 questions)
9. Karst (3 questions)
10. Coastal (4 questions)

The students had repeated exposure or testing of only a small subset of topics before the retention exam. That is, other courses in their curriculum delivered between GEGN 202 and the retention exam did not provide in-depth review of many of the topics covered and tested in GEGN 202, except for topic 6, and to a lesser degree for topic 3. The students were not aware they would be taking the retention exam and did not study for it.

The results of the baseline retention tests are summarized on Figures 1 and 2, which display the percentage of correct answers for each question and for each topical category, respectively.

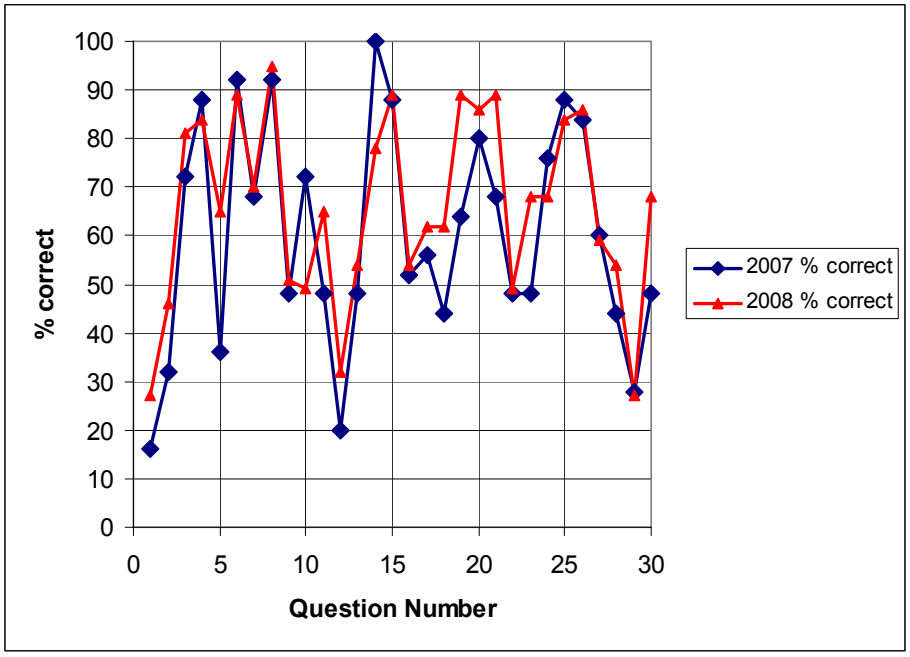

Figure 1. Summary of retention exam scores for individual questions.

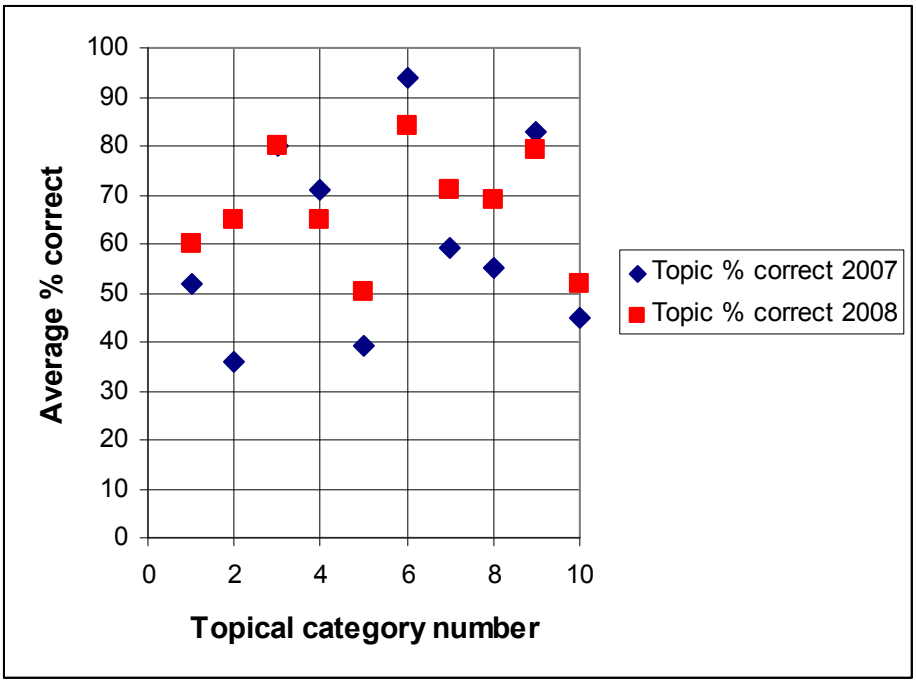

Figure 2. Summary of retention exam scores for each topical category.

A few observations are intriguing and worthy of comment:

1. The mean score on the test was 60% for the 2007 control group and 66% for the 2008 control group. For comparison, the average of test scores for the class three semesters

earlier was 79% and 82%, respectively. From these differences it may be concluded that there is a measurable performance difference, even between two groups learning under close to identical conditions. There are no clear indicators why the 2008 grouped seemed to perform better than the 2009 group.

2. While the scores of the two control groups showed similar higher and lower values from question to question on Figure 1, there are obvious differences in the percent correct for each group of students. On average, there is a 10% range between scores on each question. The 2008 control group scored higher on the majority of the questions, resulting in their higher average score, while the 2007 control group scored higher on only six of the questions.
3. Scores on the retention test did not correlate with the order of presentation of topics during class, as shown by the randomly varying scores on Figure 2. Topics presented early or late in the semester were not retained better or worse than other topics.
4. Scores on the retention test showed only a weak correlation ($R^2 = 0.12$) with scores on the in-class exams taken 18 months earlier (Figure 3). When the control groups are separated, the trend for the 2007 group is even weaker ($R^2 = 0.033$). This is an important observation: success in class does not necessarily indicate success in retaining information after class.

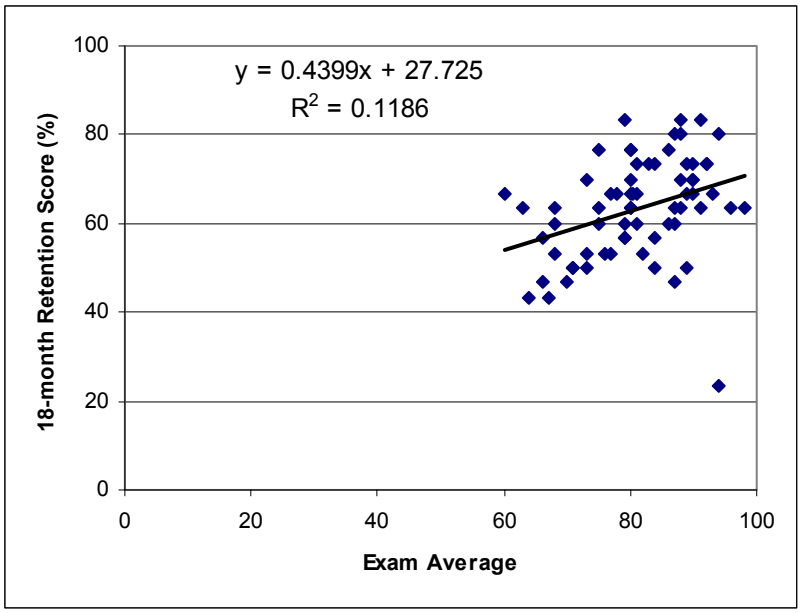

Figure 3. Correlation of the average exam grades students received in GEGN 202 with the retention exam score 18 months later. The trend is weakly statistically significant.

Cumulative testing during 2009 and 2010

A short evaluation was given to the cohort of students who were given the multiple cumulative exams in 2009 and 2010 (the "2009 test group" and the "2010 test group"). The evaluation consisted of a series of questions to be answered on a Likert scale (strongly agree, agree, neutral, disagree, strongly disagree). Notable results are in Table 2.

Table 2. Results of survey of 2009 and 2010 test groups who had taken 6 and 5 cumulative exams, respectively (n = 103).

Question	Strongly disagree or disagree	Neutral	Strongly agree or agree	Summary
Do you feel you learned this subject better by having many, comprehensive exams?	4%	9%	87%	Overwhelmingly "yes"
Did you review old material from previous exams as much as you should have?	39%	19%	42%	As many did as did not
Did you have to give up time for learning the new material in order to have time to review the old material?	35%	23%	43%	Nearly as many did as did not
Should we keep giving comprehensive exams?	8%	8%	83%	Overwhelmingly "yes"
How many exams do you think is ideal?	61% responded with the number of exams they took 24% suggested one fewer than the number they took 8% suggested two fewer			5 or 6 seems ideal

The students overwhelmingly agreed that they learned the subject better by being tested frequently (87% responded "agree" or "strongly agree"). Likewise, 83% responded that multiple comprehensive exams should continue to be given in the class. On the negative side, only 42% felt they reviewed old material from previous exams as much as they should have (although there was a strong difference in the two groups: 26% from 2009 and 57% from 2010). Moreover, when asked if they had to give up time for learning new material in order to review old material, 35% agreed that they did. For the most part, students were accepting of taking frequent exams, with 61% agreeing that the number they took was ideal (6 exams for the 2009 group and 5 exams for the 2010 group). A quarter of them suggested taking one fewer exams than the number they took.

2011 and 2012 retention exams

The retention exams for the 2009 and 2010 test groups will be given in April 2011 and April 2012, respectively, so results are not available at the time of this writing. Results for the 2009 test group will be presented at the conference.

Potential limitations and confounding factors

Although this experimental design has distinct advantages in simplicity and ease of implementation by others, it also has some inherent weaknesses. One of the foremost is that students have no compelling reason to do their best on the 18-month retention exam. It has no bearing on their grades and they are told that it is part of an educational study. Furthermore, it takes them 10-20 minutes, so they may lose whatever motivation of concentration they had before they are finished. On the other hand, as a multiple-choice exam it is straightforward for them to complete, and the novelty of seeing what they can recall and of being part of a scientific study may be sufficient motivators to inspire full efforts. These students all take summer field course two months after the retention exam, and they have been quite interested in seeing their (and their colleagues!) scores.

Another potential confounding factor is that the multiple in-class exams may simply be improving students' abilities as test-takers, and taking five or six exams written by the same instructor may prepare them better for the 18-month exam than taking only two in-class exams. Their support for multiple exams may stem from their desire to have more opportunities to improve their grades or to reduce the impact of a single poor performance. However, the average grades for each of the five exams for the Fall 2010 class showed almost no change through the semester, except for a dip in the score on exam 2 (the average grades, in chronological order, are 84.5, 79.7, 82.9, 84.3, and 84.4).

A third possible confounding factor is the reinforcement of topics in other classes during the intervening 18 months. This certainly was the case for topic 6 in Figure 2, and to a lesser degree for topic 3, and both topics show higher retention scores than average.

The degree of sophistication of the material being tested may be a limitation to applying the results of this study in broader scenarios. Most of the items in the current retention exam test students at Bloom's[20] taxonomy levels 1 and 2 ("knowledge" and "comprehension"), which focus on factual recall, definitions, vocabulary, and basic conceptual understanding. The demonstration of impacts of repeated testing on retention of higher level skills (application, analysis, synthesis, and evaluation) will be more tenuous.

The intensity of instruction is another factor that may influence retention, but is not gauged for this study. Previous research has shown the memory benefits when "challenges," such as more diverse or intense learning experiences, are included in the teaching process[4, 21].

Conclusions

While the improvement of long-term retention resulting from multiple, cumulative exams remains to be shown, the baseline studies have revealed several points related to retention.

- Retention exam scores were 16-19% lower than scores on in-class exams three semesters earlier. This may indicate a measurable loss in information retention over the elapsed time.
- In-class exam scores are poor predictors of the level of retention over a long time period (16-18 months in this case).
- Retention is not better for topics presented later in the class.
- Students were amenable to taking a larger number of cumulative exams, although they may face time constraints that limit their review of old material or their learning of new material.

Bibliography

1) Plucker, J., 2007, Hermann Ebbinghaus, http://www.indiana.edu/~intell/ebbinghaus.shtml, accessed 5/6/10.

2) Hilgard, E.R., 1980, The trilogy of mind: cognition, affection and conation, J. Hist. Behav. Sci 16, 107 – 117.

3) Lechner, H.A., Squire, L.R., and Byrne, J.H., 1999, 100 years of consolidation – remembering Muller and Pilzecker, Learn. Mem 6, 77 – 87

4) Halpen, D.F. and Hakel, M.D., 2003, Appling the Science of Learning to the University and Beyond: Teaching for Long-Term Retention and Transfer, Change, July/August 2003, pp. 37-41.

5) Squire, L., and Kandel, E., 1999, Memory: From Mind to Molecules, Freeman, New York.

6) Landauer, T. K., & Bjork, R. A., 1978, Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes, eds., Practical aspects of memory (pp. 625–632). London: Academic Press.

7) Wagner, A.D., Schacter, D.L., Rotte, M., Koutstaal, W., Maril, A., Dale, A.M., Rosen, B.R., and Buckner, R.L., 1998, Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity, Science, 21, 1188-1191.

8) Medina, J., 2008, Brain Rules, Pear Press, Seattle, 301 p.

9) Wagner, A.D., Mariland, A., and Schacter, D., 2000, Interactions between forms of memory: when priming hinders new learning, J. of Cognitive Neuroscience 12, 52 – 60.

10) Roediger, H.L. and Karpicke, J.D., 2006, Test-enhanced learning: Taking memory tests improves long-term retention, Psychological Science, 17, 249-255.

11) Karpicke, J.D., Butler, A. C., and Roediger, H. L., 2009, Metacognitive strategies in student learning: Do students practice retrieval when they study on their own?, Memory 17, 471-479.

12) Bjork, R. A., 1975, Retrieval as a memory modifier. In R. Solso, ed., Information processing and cognition: The Loyola Symposium (pp. 123-144). Hillsdale, NJ: Lawrence Erlbaum Associates.

13) McDaniel, M. A., & Masson, M. E. J., 1985, Altering memory representations through retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 371–385.

14) 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.

15) Cull, W. L., 2000, Untangling the benefits of multiple study opportunities and repeated testing for cued recall. Applied Cognitive Psychology, 14, 215–235.

16) Logan, J., and Balota, D.,2008, Expanded vs. equal interval spaced retrieval practice: exploring different schedules of spacing and retention interval in younger and older adults. Aging, Neuropsychology and Cognition, 15(3), 257–280

17) Karpicke, J.D. and Roediger, H.L., 2007, Expanding Retrieval Practice Promotes Short-Term Retention, but Equally Spaced Retrieval Enhances Long-Term Retention Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(4), 704–719

18) Butler, A.C. and Roediger, H.L., 2007, Testing improves long-term retention in a simulated classroom setting, European Journal of Cognitive Psychology, 19, pp. 514-527.

19) Bahrick, H.P., 2000, Long term maintenance of knowledge, in Tulving, E. and Craik, F.I.M., eds., The Oxford handbook of memory, Oxford University Press, pp. 347-362.

20) Bloom, B.S. and Krathwohl, D.R., 1984, Taxonomy of educational objectives: Handbook I: Cognitive domain. New York, Addison Wesley.

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