Simona CHERA, PhD
Department of Clinical Science, UiB

Analyzing the impact of recurrent Kahoot tests on student performance

Abstract

Here we performed a small pilot study assessing the impact of recurrent kahoot quizzes on medical students' performance in cutting-edge subjects (not in the typical curricula). Albeit several inherent design limitations, our research suggests that recurrent testing significantly improves student's results, especially when combined with active learning and practice.

1. Introduction

Teaching regeneration and tissue repair to medical and science students requires the presentation of diverse regenerative model systems. We observed that students have difficulties to follow through and reach the course objectives (such as extract a common denominator for the cellular and molecular basis of regeneration), which translates into low scores during end-class kahoot! (Wang 2015) testing. Moreover, even following the lecture, common misconceptions regarding regeneration in adult mammals were still present in some of the students, despite being actively debunked during the classical lecture.

Testing students (quizzes) was proven to be highly useful for improving students' performances when used along the semester (Roediger, Agarwal et al. 2011, McDaniel, Thomas et al. 2013). Moreover, using frequent quizzes or repeated testing prompted better final test results when compared to additional reading (McDaniel, Anderson et al. 2007) assignments or repeated study (Karpicke and Roediger 2008, Karpicke and Blunt 2011, Pastotter and Bauml 2014). Interestingly, the same study also showed that short answer questions have higher impact than multiple-choice questions. Of note, the use of tests improved also the performance in relation to connected untested material, indicating an overall beneficial impact of the method.

The improvement of retention triggered by active retrieval as compared to passive reading, otherwise called the "testing effect", was observed and studied for more than a century. At the beginning of the 20th century Abbott followed by Gates and

Jones published a series of studies highlighting the utility of testing in learning processes (Abbott 1909, Gates 1917, Jones 1923). Since then, an ever-increasing body of research analyzed the topic, further proving better results, especially in delayed tests, in tested cohorts as compared to the untested ones, regardless if the students restudy the course material or not (Hogan and Kintsch 1971, Wheeler, Ewers et al. 2003, Roediger and Karpicke 2006, Agarwal, Karpicke et al. 2008, Agarwal, Bain et al. 2012). Relevant for our work, there is a direct relationship between the number of tests and the increase in performance, especially if the tests are followed by feedback (Roediger and Karpicke 2006).

In this study, we aimed at investigating the impact of recurrent testing on the performance improvement on the students enlisted in the "Regeneration Strategies" class, by using a series of kahoot quizzes. Briefly, during the class (4x45 min) one group of students were exposed to three quizzes containing identical questions, with the difficulty ranging from basic to subtle knowledge of regeneration processes. The tests were separated by (i) classical lecture (presentation) and (ii) article and group debates, with the first quiz (pre-test) being taken at the beginning of the class and the last (end-test) just before the end. In contrast, the control group was only tested at the end of the class (end-test), while being exposed to all the other activities (from now on ET group). In order to study if periodic testing following distinct class activities affects individual performance, a longitudinal analysis of the recurrently tested cohort (from now on RT group) was performed. Subsequently, we completed a comparative analysis of the end-test results between the two groups (i.e. RT vs ET).

In certain aspects this design is similar to the one described in (Roediger, Agarwal et al. 2011).

2. Method

2.1 The course format

We recently implemented a new optional yearly course at the Faculty of Medicine, University of Bergen (ELMED303, Future Medicine). It is a two weeks course, covering 10

hot topics in Biomedical sciences, each class being given by an expert in the field. The session format mimicks the structure of the Harvard Nanocourses(Bentley, Artavanis-Tsakonas et al. 2008), with two hours (2x45 min) of lecture, one hour (1x45 min) of hot topic scientific article discussion and one hour (1x45 min) free discussion on topic, including a test. Nevertheless, in our case, the structure is rather fluid, allowing teachers to swap the above succession of events. Similarly, the type of test used at the end of the class is optional, with teachers adhering to what they think is fit, according to the topic.

Besides organizing this course, I teach the "Regeneration Strategies" class. The standard class retains the above format, with two hours of classical (presentation-based) teaching, followed by commenting one recent scientific review and free discussions. During the free discussions, the students are randomly assigned to two teams and they are asked to debate different fictive medical cases requiring cell replacement. The debates are centered on the choice of the regenerative strategy employed, each group being demanded to defend a certain approach using scientific and ethical reasoning.

2.2 Participants

The course is aimed at year five medical students as well as master students from the biomedical field at University of Bergen. The number of participants varies between 10 and 20, depending on the year. Importantly, as this is an elective course, the students opt to enlist based on a short description offered in advance by the Faculty of Medicine. Thus, it is expected that they present a fair initial interest in the topic. Nevertheless, it should be also stated that the course might be also attractive to students due to its simple fail/pass system dependent on students' attendance. We noticed a balanced gender ratio.

2.3 Design

The study considered students over 2 successive years. The students were verbally informed and verbally accepted their participation in the experiment. One group underwent recurrent testing (RT group), while the other received just an identical end-test (ET group).

Quiz design: The testing was performed using Kahoot!, an on-line game-based learning platform. The participants received 10 multiple-choice questions (some questions with multiple correct answers, 20 sec / question) all related to the class topic. The difficulty of

the questions was incremental, addressing a range from simple to more subtle concepts (please refer to two such examples in Figure 1).

Procedure: For the RT group, the first quiz (pre-test) was applied at the beginning of the class, following teacher's presentation. The students were informed that this test is meant

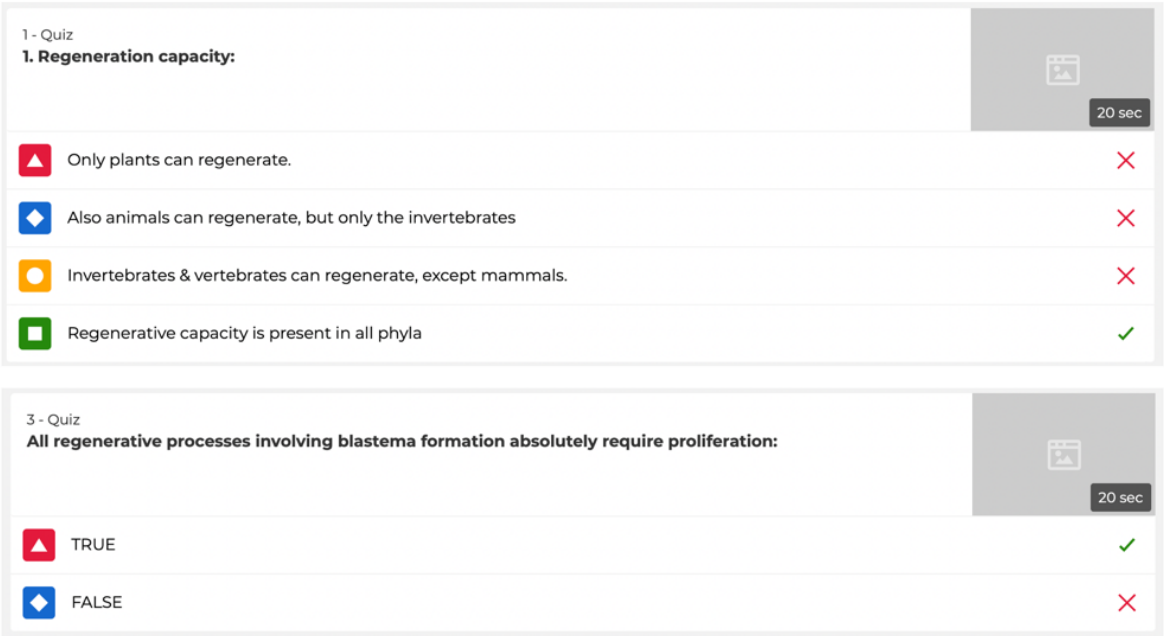

Figure 1. Examples of quiz questions.

to assess their initial knowledge about regeneration in mammals/humans as well as to familiarize them with the class content. It should be stated that the students have no previous systematic expertise on the subject, as there is no course properly covering this topic in the basic curricula at the Faculty of Medicine. As such, it is expected that the basic knowledge students' have on regeneration is acquired from informal and not necessarily specialized alternative sources (such as action movies, news, general knowledge documentaries, web) and hence can be impacted by generally accepted misconceptions (such as "humans cannot regenerate"). These pre-test results were not discussed with the students immediately following the test. The pre-test was succeeded by an interactive, however classical lecture (slide show based) and a second test (mid-test) following the same format as the pre-test (Figure 2, left). Similarly, the results of this test were not immediately discussed with the students. Last, following the discussion of a research article and the team debate of several medical cases and their best regenerative-based treatment approaches, the students were tested one final time (end-test). The end-test followed the same format

as the previous two. The dynamic evolution of the students' answers following the integration of the three tests results was discussed (overall trends for each question).

For the ET group, only the final test was applied, following a similar succession of class events (Figure 2, right). The feedback on each question was provided and discussed immediately after.

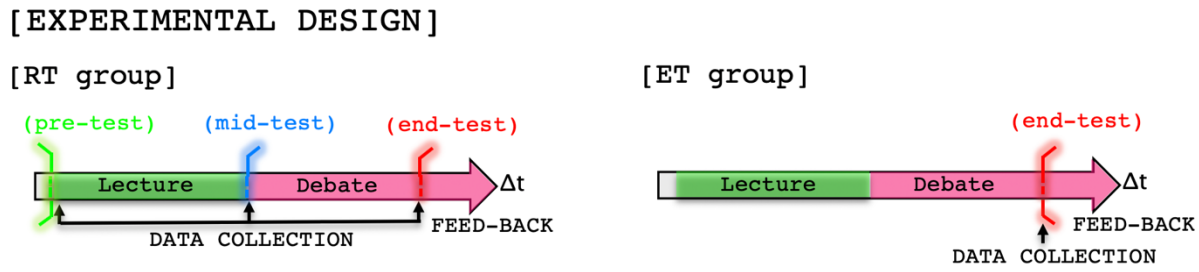

Figure 2. The experimental design employed for the RT and ET group.

Analysis design: Both intra- and inter-group analyses were performed. For the intra-group analysis, the overall and individual dynamic of the answers' pattern was studied. Moreover, a direct comparison between the end-test results was performed between the RT and ET groups (Figure 3). The answers were assessed both globally and specific, according to question difficulty classes. Graphs were generated in GraphPad Prism 8, using data exported from Kahoot. Column graphs display mean and standard deviation. For statistical tests we performed Anova, non-parametric tests and paired t-test, according to the biological question addressed.

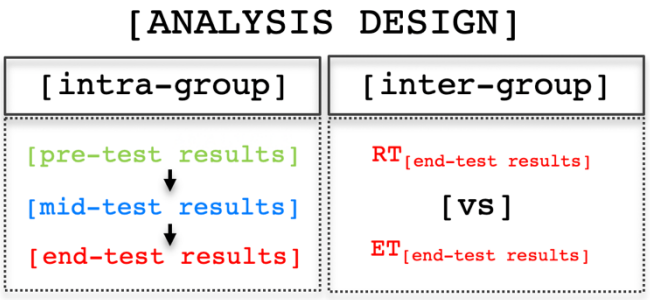

Figure 3. Analysis workflow

3. Results

*3.1. End-class test in the **ET cohort***

Eleven students were enrolled in the ET group. The students were given the test in the last 15 minutes of the class, following the lecture, article discussion and debates (Figure 2, right). The overall group performance was assessed at 61.82% (Figure 4A) with an average score of 6105.3 points. No question (0/10 questions) received only wrong answers, while all students responded one question correctly (Figure 4B). In addition, no student failed all

questions, neither responded all correctly, with the minimum of 40% (4/10 correct answers) and a maximum of 80% (8/10 correct answers). The distribution exhibited a slightly bimodal profile (Figure 4C), with the first mode at score 50% and the second at 70%, suggesting a non-stochastic heterogeneity in the performance inside the assessed cohort.

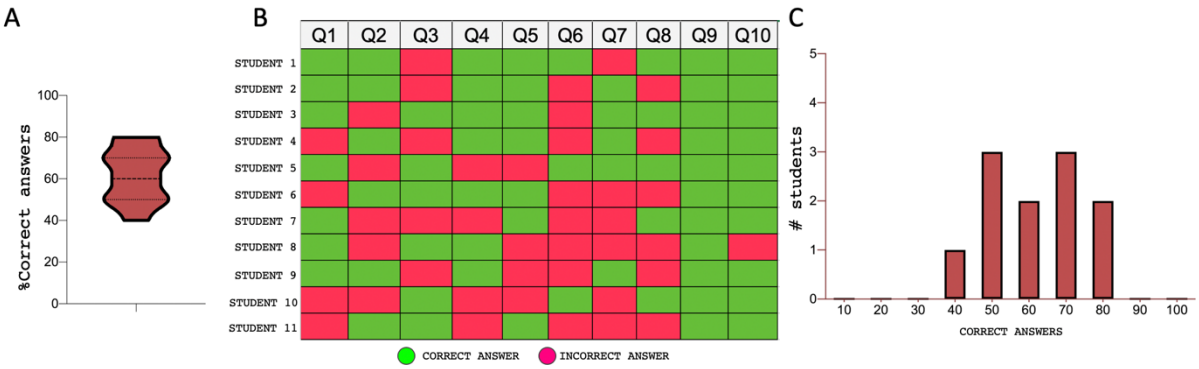

Figure 4. End-test results in the ET group.

Despite the limited number of participants, which impedes on the quality of the statistics, these results were below our expectations at the end of the class.

*3.2. Pre-test in the **RT cohort***

77.77% of the students enrolled in the RT group (7/9 students) took this test, the rest (2/7 students) arriving late to class and, consequently, missing the first test. Overall, the initial group performance was 60%, with an average score of 5331.4 points (Figure 5A). No question (0/10 questions) was answered wrongly by all students, while two questions (2/10 questions) were correctly answered by everybody (Figure 5B). Moreover, no student failed consistently to reply to the questions, with the minimum score being 50% (5/10 correct answers) and the maximum of 70% (7/10 correct answers). The cohort followed a rather Gaussian behavior with 28.6% of the students (2/7 students) reached a 70% score, while 42.8% were at 60% (3/7 students) and 28.6% at 50% (2/7 students) (Figure 5C).

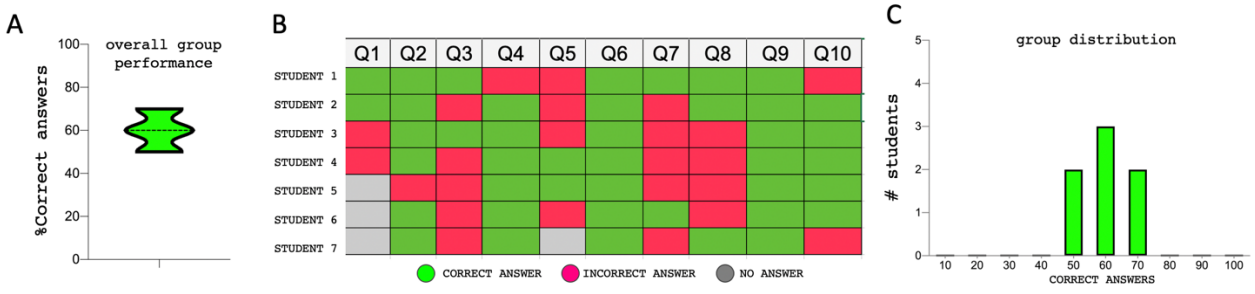

Figure 5. Pre-test results in the RT group.

Despite the low number of participants that is not allowing proper meaningful statistics, these results suggest that the pre-test was well balanced, with all participants being able to reply to at least 50% of the questions. Moreover, the fact that all quiz items received at least one correct answer, indicated that no question was an outlier (i.e. too difficult). As the questions' difficulty was heterogeneous, this test also showed that the level of knowledge of this student cohort was as desired for the beginning of the lecture.

*3.3. Mid-test in the **RT cohort***

The mid-test was performed after the standard lecture (mid-class). All students enrolled in the RT group participated. The overall group performance was 65.5% (Figure 6A). As before, no question received only wrong answer, nevertheless only one question was correctly answered by all students (Figure 6B). No student failed all questions, with the lowest score being 50% (5/10 questions) and highest 90% (9/10 questions), higher than the ones recorded at pre-test. The cohort distribution was skewed to the right with the mode at 60% (Figure 6C).

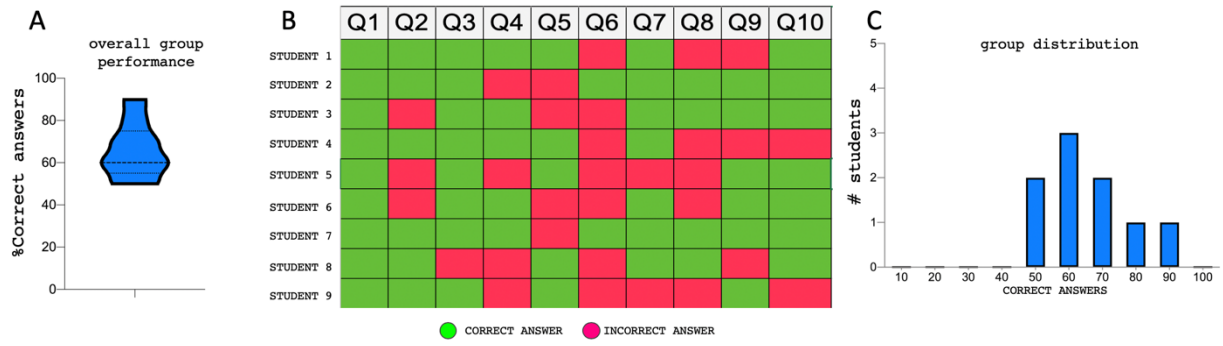

Figure 6. Mid-test results in the RT group.

*3.4. End-test in the **RT cohort***

The end-test was performed following the debates around two successive cases, where two groups of students had to justify a certain scheme of treatment to each other. All students participated at this test with an overall group performance of 80% (Figure 7A). Two questions were answered by all students (Figure 7B). The lowest score was 60%, while the highest was 100% (2/9 students), higher than the ones recorded during the mid-test. Interestingly, the cohort followed a trimodal distribution with the median, mean and mode at 80% (4/9 students) (Figure 7C).

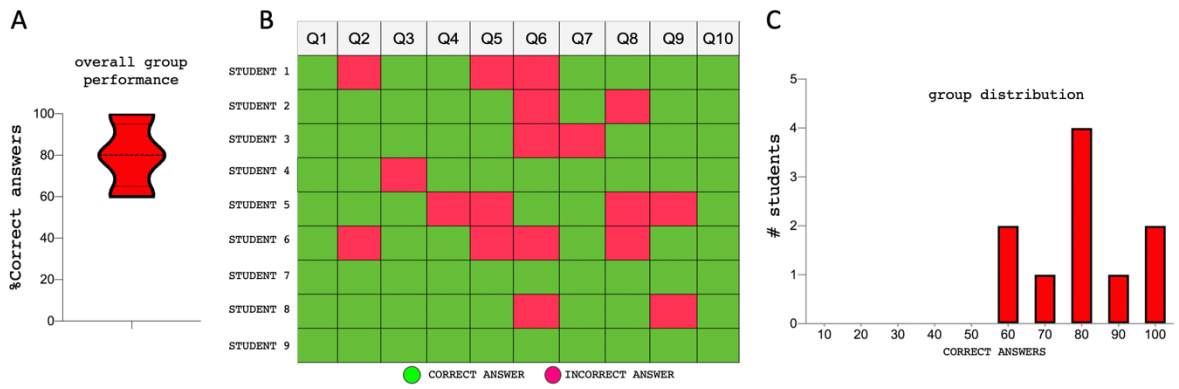

Figure 7. End-test results in the RT group.

*3.5. RT-group Tests Comparison shows improved student performances on the end-test in the **RT cohort***

Comparing the dynamic overall group performance indicated a significant improvement of the students' answers at the end-test as compared with both pre- and mid-test (both by ANOVA and non-parametric test direct comparison, Figure 8A). Indeed, the group performance climbed from 60% at the beginning to 80% at the end of the class. Moreover, the group distribution shifted slightly to the right along the tests, transforming from an apparent Gaussian distribution to a trimodal one (Figure 8B and compare 5C, 6C, 7C). This suggests a non-stochastic difference in students' involvement/performance, probably caused by several variables that will be addressed in the discussion of this study. We further compared the individual student performance along the class (Figure 8C). This confirmed that most students (67%) performed better at the last test (end-test), some exhibiting significant improvements. Moreover, all but one performed better than in the pre-test.

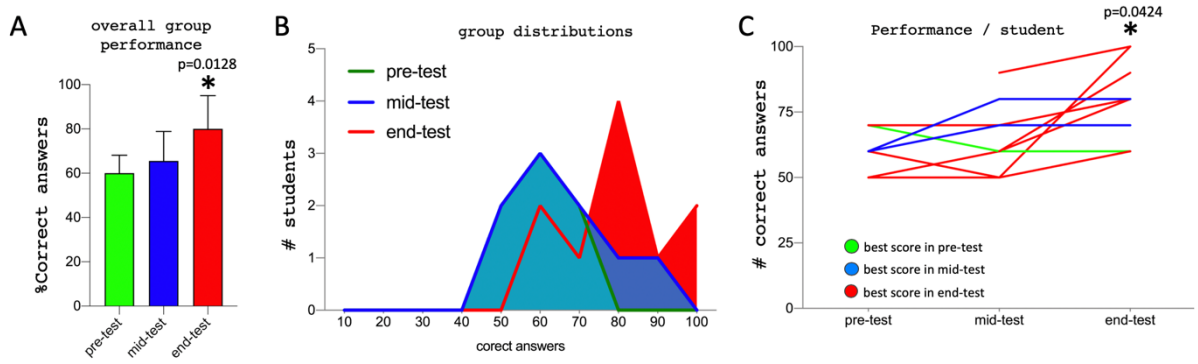

Figure 8. Dynamic comparison of the tests results in the RT group.

3.6. ET- vs RT-group Comparison

Last, to assess the performances of the ET and RT groups we compared the end-test results between the two cohorts. The recurrently tested group (RT group) exhibited significantly improved their performance during the end-test (Figure 9A). This result was also confirmed by the comparison between the two cohort's distributions, showing a shift to the right in the RT group, indicating that, overall, the students performed better when recurrently tested. Both distributions were non-Gaussian, being either bimodal (ET group) or trimodal (RT group), indicating a heterogeneity in the students' cohorts, suggestive of different responses to the class materials. Interestingly, there was no significant difference between the end-test of the ET cohort and the pre-test of the RT-cohort.

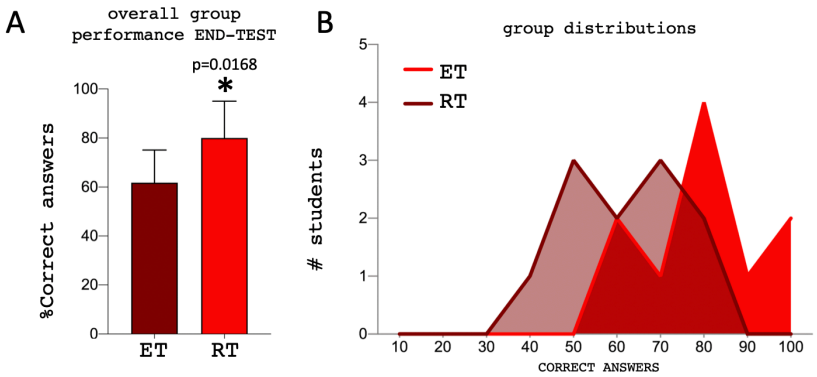

Figure 9. Direct comparison of the end-test results between ET and RT group.

Discussion

Here, we observed a significant difference in students' performance in the groups receiving recurrent Kahoot quiz tests as compared to a control group, which received a single end-class quiz. In our opinion, the main drawback of this pilot study is the low number of participants, which does not allow robust statistics. To properly demonstrate the utility of recurrent testing in this setup, one would require larger number of students over several years over a wider range of subjects.

Nevertheless, based on the current results, the students performed much better when tested along the class day. This is in accordance to previous studies, such as (Roediger, Agarwal et al. 2011), where the power of re-testing was demonstrated in larger cohorts of

students. In contrast with this study, due to the compact course format (2 weeks, no final exam) we were not able to properly assess the "test effect" in long term. In our opinion, testing the effects of recurrent quizzes on long-term concept retention will be a very interesting parameter to investigate, especially for medicine students.

A second difference with the above-mentioned study is the structure of the class. In the previous study, the class was rather homogenous, while in our case we had a standard lecture, followed by and equal amount of time of discussions and debates. This points towards a second weakness of tour study, as this mixed structure might cause a confounding effect on student performance. This problem is also indicate by the fact that the longitudinal comparison of the RT group tests results, indicated a better concept understanding as the class progresses. Of course, this can be simply attributed to the actual "test effect", especially that both the ET and RT group followed an identical class structure, the only difference being the quizzes frequency. However, the situations might be more complex, if one considers that the initial quizzes will help students (i) understand better the course objectives, (ii) focus on most relevant class content and (iii) get familiar with the type of concepts the professor values. In contrast, the ET-group focus on the course objective is merely based on their display on a slide and professors' rhetoric.

Interestingly, supporting this scenario, both the qualitative observations of both the professor and an independent support person (not included in this study as they could not be properly measured) pointed at the debate part of the course being much more focused in the RT-group than ET-group. Indeed, the recurrently test students were using better the specific terms, addressed very relevant questions, were overall more involved in the debates and highly enthusiastic.

Testing students at the beginning and end of the class (no mid-test) and comparing the end-test results with the ones in the RT-groups might help to better understanding the relationship between the test effect and dynamic class structure. This points towards another drawback of this pilot study, represented by the lack of initial assessment of the student cohort in the ET group. It is highly improbable, but not impossible, that by serendipity, the ET students are overall less informed about the class subject than their RT counterparts. In the lack of a pre-test this possibility cannot be properly and formally excluded. We decided to not perform the pre-test in the ET cohort as we wanted to be

initially as close as possible to the typical format of the course (i.e. one test at the end of each class).

The pre-test results in the RT cohort were encouraging as the participants scored well for students that were never formally exposed to the subject. We consider this test extremely useful for assessing the overall level of the cohort and especially for identifying potential misconceptions that tend to usually characterize the latest discoveries in medicine and biology.

Another interesting result of this pilot study is the distribution of the students in the end-test, which failed to follow a Gaussian shape. A potential explanation is that students maintained different levels of interest along the class. The dynamic of the RT group shows that most improved following the debate, where they actively needed to search and use information to convince their peers of their "truth". Classical lecture, although improving slightly the students' scores, did not reach the same levels of impact. Nevertheless, despite the improvements observed in the vast majority, the levels of involvement were quite different between students.

To finalize, based on our pilot study we advance the hypothesis that students focus better in class if they receive recurrent tests helping them familiarize with the course objectives and concepts. Moreover, the improvement seems to be highest when they need to actively use the learned concepts, such as during debates. These might form a positive feedback loop in which the students are more focused on the debate structure, which in turn improves their performances in tests.

References

Abbott E. E. (1909). On the analysis of the factors of recall in the learning process. Psychol. Monogr. **11**, 159–177 10.1037/h0093018.

Agarwal, P. K., P. M. Bain and R. W. Chamberlain (2012). "The Value of Applied Research: Retrieval Practice Improves Classroom Learning and Recommendations from a Teacher, a Principal, and a Scientist." Educational Psychology Review **24**(3): 437-448.

Agarwal, P. K., J. D. Karpicke, S. H. K. Kang, H. L. Roediger III and K. B. McDermott (2008). "Examining the testing effect with open- and closed-book tests." Applied Cognitive Psychology **22**(7): 861-876.

Bentley, A. M., S. Artavanis-Tsakonas and J. S. Stanford (2008). "Nanocourses: a short course format as an educational tool in a biological sciences graduate curriculum." CBE Life Sci Educ **7**(2): 175-183.

Gates, A. I. (1917). Recitation as a factor in memorizing. Archives of Psychology, **6,** No. 40.

Jones, H. E. (1923). The effects of examination on the performance of learning. Archives of Psychology, **10**, 170

Hogan, R. M. and W. Kintsch (1971). "Differential effects of study and test trials on long-term recognition and recall." Journal of Verbal Learning and Verbal Behavior **10**(5): 562-567.

Karpicke, J. D. and J. R. Blunt (2011). "Retrieval practice produces more learning than elaborative studying with concept mapping." Science **331**(6018): 772-775.

Karpicke, J. D. and H. L. Roediger, 3rd (2008). "The critical importance of retrieval for learning." Science **319**(5865): 966-968.

McDaniel, M. A., J. L. Anderson, M. H. Derbish and N. Morrisette (2007). "Testing the testing effect in the classroom." European Journal of Cognitive Psychology **19**(4-5): 494-513.

McDaniel, M. A., R. C. Thomas, P. K. Agarwal, K. B. McDermott and H. L. Roediger (2013). "Quizzing in Middle-School Science: Successful Transfer Performance on Classroom Exams." Applied Cognitive Psychology **27**(3): 360-372.

Pastotter, B. and K. H. Bauml (2014). "Retrieval practice enhances new learning: the forward effect of testing." Front Psychol **5**: 286.

Roediger, H. L., 3rd and J. D. Karpicke (2006). "The Power of Testing Memory: Basic Research and Implications for Educational Practice." Perspect Psychol Sci **1**(3): 181-210.

Roediger, H. L., P. K. Agarwal, M. A. McDaniel and K. B. McDermott (2011). "Test-enhanced learning in the classroom: long-term improvements from quizzing." J Exp Psychol Appl **17**(4): 382-395.

Roediger, H. L. and J. D. Karpicke (2006). "Test-enhanced learning: taking memory tests improves long-term retention." Psychol Sci **17**(3): 249-255.

Wang, A. I. (2015). "The wear out effect of a game-based student response system." Computers & Education **82**: 217-227.

Wheeler, M. A., M. Ewers and J. F. Buonanno (2003). "Different rates of forgetting following study versus test trials." Memory **11**(6): 571-580.