RESEARCH DIRECTIONS

Spaced Repetition: towards more effective learning in STEM

Alison Voice[1]*and Arran Stirton[2]

[1]School of Physics & Astronomy, University of Leeds, Leeds, LS2 9JT
[2]School of Physics & Astronomy, University of Leicester, Leicester, LE1 7RH

Corresponding author: a.m.voice@leeds.ac.uk

Keywords: *Spaced repetition; Long term memory; Cramming; Web app*

Abstract

The use of spaced repetition within a physics higher education thermodynamics module has been analysed for: its pattern of use by students; its effect on memory and performance in the end of module exam; and performance in a delayed test after the summer vacation. A custom-built web app with the facility to generate a personalised repetition timetable was used to deliver practice questions on the material throughout the module. Just over a quarter of students, spanning the whole ability range of the class, made use of the app in some way, about half using it in a spaced manner and half using it for massed practice just before the exam. Students who engaged in a spaced manner had an adjusted mean exam score of 70%, compared to 64% for massed usage and 61% for non-usage. The spaced usage represents a positive effect size of 0.47 over non-usage, which is statistically significant (p = 0.000056). For the delayed test the mean adjusted scores for spacers and non-users were 45% and 34% respectively. Whilst less material had been retained over the summer, this revealed a statistically significant (p = 0.021) positive effect size of 0.54. This work provides evidence and mechanisms to involve students in repetitive practice during the learning phase of a course to advantage their long term retention of material.

Introduction

The benefit of spaced repetition on learning has been recognised for over a century, with Ebbinghaus (1885) making the first systematic investigation of memory, developing a ‘forgetting curve’. This shows exponential decay of information from the memory when no effort is made to revisit that information. However, if the information is revisited the rate of decay reduces, thus allowing for ever increasing time intervals between repetitions to retain long term memory of it. This psychology research was first applied to education by Mace (1968) suggesting that regular revision of curriculum material would be much more effective than ‘massed’ study (all at once). He proposed that revision should be spaced in gradually increasing intervals.

In higher education long term memory is needed for students to have a sound platform of understanding on which to build new advanced knowledge, and to apply when solving problems. Indeed, the Benchmark Statements for Physics (Quality Assurance Agency, 2017) and Chemistry (Quality Assurance Agency, 2014) specifically mention this skill. Students, however, are often prone to studying material in a massed way, in a ‘cram for the exam’ mentality despite the potential for diminishing the long term benefit. A survey of 189 physics undergraduates (Curtis *et al.*, 2018) revealed that students in all years study predominantly by reading and writing notes during term time, with 69% leaving any form of

https://doi.org/10.29311/ndtps.v0i15.3376

self-testing until the month before the exam. This learning strategy is hardly surprising since both students and teachers tend to focus on short term performance, and rarely consider the effect on long term retention (Kornell & Bjork, 2007; Roediger & Karpicke, 2018).

Further investigation into the nature of this spaced practice reveals that recall (testing) is more beneficial than re-reading (Roediger & Karpicke, 2006a) even in the absence of marking or feedback. This is explained (Rawson *et al.*, 2015) as the process of recall improving and increasing the retrieval pathways in the brain.

In STEM classroom situations recent studies have demonstrated the benefit of spaced repetition on long term memory with Medical students (Kerfoot *et al.*, 2007), Maths students (Rohrer & Taylor, 2007; Gallo & Odu, 2009; Rohrer, 2009), Engineering students (Hopkins *et al.*, 2015) and Natural Science students (Kapler *et al.*, 2015).

In the 1970s Leitner (1972) developed a system of flash cards, sorted into five boxes, to employ the spacing effect to develop long term memory. The first box contained information least known, the second box was known a little better, and so on. Cards in the first box were then reviewed daily, the second box every two days, etc. When confident that information on a particular flash card could be recalled it was promoted to the next box, but failing to recall the information on any card in any box, demoted it back to the first box. With the advent of personal computers and mobile technology, apps have been created to replicate this process, the most well-known probably being *Anki* and *Duolingo*.

Based on this wealth of evidence that regular repetition and recall of material promotes better long term memory, the research reported here investigates how first year undergraduate (UG) Physics students engage with practice questions supplied via a custom-built spaced-repetition app. The effect of app usage on performance in the end of module exam, and an impromptu delayed test after the summer vacation is examined.

Methods

To examine the effects of spaced repetition in a university STEM module setting, trials were conducted on three successive cohorts of first year physics students at the University of Leeds over the academic years 2016-17, 2017-18, and 2018-19. For each trial, the students were given access to the intervention (a bespoke spaced-repetition web app) to assist them with their semester two thermodynamics module.

Outcomes were assessed according to a controlled before-after design, comparing semester one and module-specific semester two exam performance between the control and intervention groups. Additionally for the 2018-19 cohort, a delayed test on thermodynamics content was administered at the beginning of their second year to assess how well material was retained over the summer.

Student participation

In each trial all first year students enrolled on the thermodynamics module were invited to use the web app to assist their learning. Each cohort was informed about the app through an initial in-lecture presentation and reminded of its availability at various points throughout the module. The exact timing of when the app was introduced to students varied from year to year.

Participation in the intervention group was voluntary (and thus not randomized) and could commence at any point from the initial introduction of the software, right up to the exam. Faculty were blind to individual participation, but aware of overall usage and self-reported student usage. The work was conducted in line with ethical requirements and approval from the University of Leeds.

Provision of the spaced repetition web-app

A purpose-built web application was used for the trials, consisting of a bank of questions with the facility to deliver a personalized review schedule. The questions were designed to give students practice of the core laws, equations and concepts of the module, while the scheduling system informed the students individually of the best time to review each question, in line with the spacing effect.

https://doi.org/10.29311/ndtps.v0i15.3376

As the questions were being developed alongside the trials, the number and content of the questions varied slightly from year to year; starting with 50 questions in 2016-17, rising to 67 questions in 2017-18 and dropping to 62 in 2018-19. Questions were both conceptual and mathematical in nature, containing a mix of self-assessed free-recall questions alongside computer-assessed mathematical ones. To mitigate the issue of students purely memorising the answers to particular questions, the parameters in mathematical questions were randomised during each review, so that the information being recalled was the process to solve and not the numerical answer. For the self-assessed questions students had to decide for each question whether they were sufficiently familiar with that part of the syllabus, or whether they needed more practice.

The scheduling system was designed to present for repeat the next day, any wrongly answered mathematical question or a question self-assessed as not sufficiently familiar. Successful/familiar questions were set for review at ever increasing time intervals according to the following pattern: 1 day after initially completing the question, then 6 days after the first review, and then subsequently at intervals of 2.5 times the previous interval (i.e. 15 days, 37 days, etc.). Students were permitted to review questions within 10% of the scheduled date.

Outcomes to be tested

The primary outcome measure for each trial was improvement in student performance in the thermodynamics section of the second semester exam, as controlled by their overall first semester exam performance in Physics. Usage data from the web app, consisting of the time of review and pass/fail for each question, was collected to determine how students made use of the intervention.

For the 2018-19 cohort there was an additional outcome measure: student retention of thermodynamics material over the summer, measured by student performance in a delayed test at the beginning of their second year and controlled by their first year, semester one performance.

Statistical methods

Students that did not complete either of their first or second semester exams were dropped from the analysis. Scores for both the first semester physics exam and the thermodynamics exam were normalized to a percentage score to facilitate a before-and-after comparison.

As the control and intervention groups were identified post-hoc, two analyses were carried out: an *intention-to-treat* analysis where all students that used the app at least once were assigned to the intervention group, regardless of how they used it; and a *per-protocol* analysis, where students that massed their usage of the app (i.e. all their usage was on a single day) were excluded from the analysis.

Comparisons of the mean semester one physics exam performance between the control and intervention groups were undertaken using a two-tailed Welch's t-test (a modification of Student's t-test for unequal sample variances) and differences in performance were calculated. Analysis of Covariance (ANCOVA) was used to test for significant difference between intervention and control groups in their thermodynamics performance. To measure how large an impact the intervention had, the thermodynamics results were controlled for the first semester physics exam performance yielding the ANCOVA-adjusted means for each group. The effect size was then calculated as the standardized mean difference between the groups (Maxwell & Delaney, 1990).

Results and Discussion

Which students used the app?

Data on 429 students was collected across all three trials, of which 94% (405) completed both the semester one and thermodynamics exams and were carried forward in the analysis. Of these 108 (26%) were considered *app users*, having completed at least one question on the app. See Table 1 for cohort details.

Academic year	Cohort size	App users	% Users
2016-17	103	34	33
2017-18	142	24	17
2018-19	160	50	31

Table 1 Number of students using the spaced repetition app in each cohort.

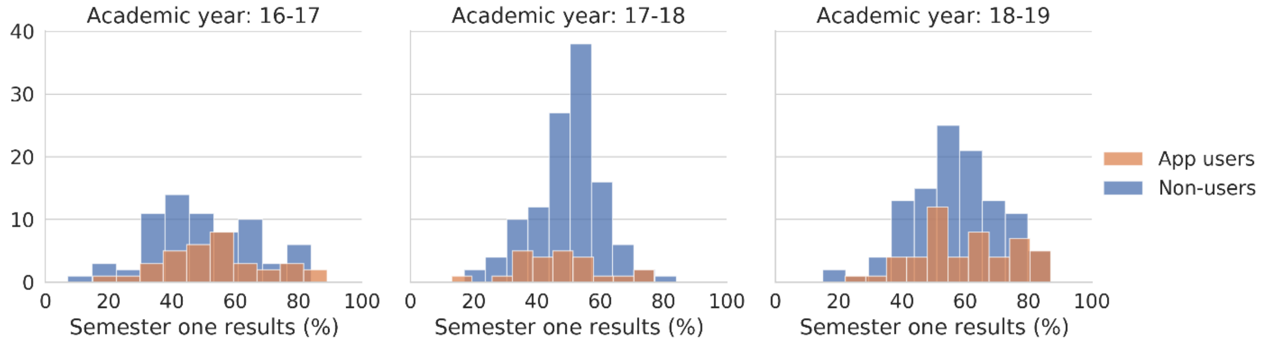

Figure 1 Histograms comparing the distributions of semester 1 physics exam marks for students that subsequently used the app vs. those that did not.

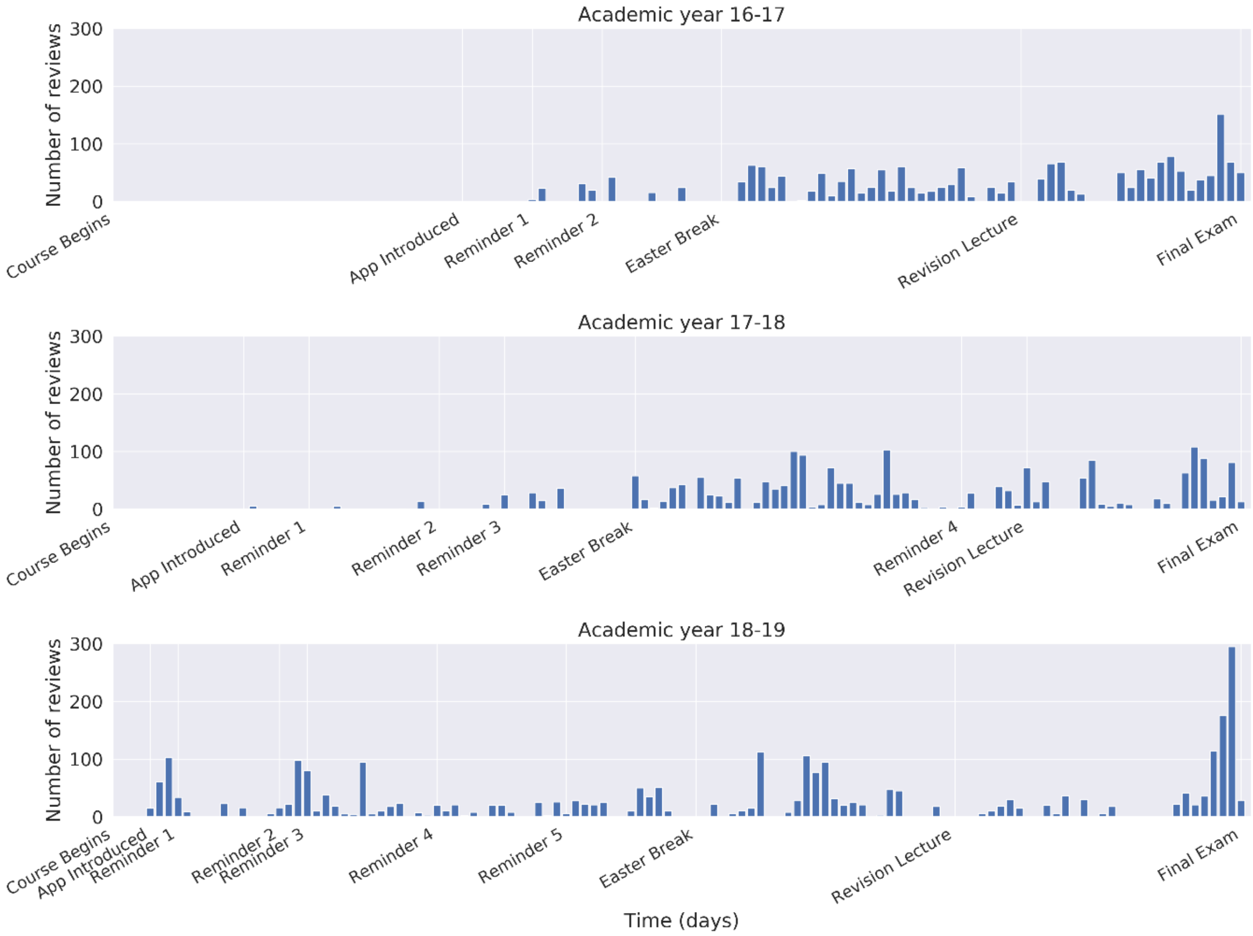

Figure 2 Timelines of the collective usage of the web app for each cohort.

https://doi.org/10.29311/ndtps.v0i15.3376

Academic year	Crammers	Spacers	% Crammed	% Spacers
2016-17	16	18	47	53
2017-18	9	15	37	63
2018-19	24	26	48	52

Table 2 Number and percentage of spacers and crammers in each cohort of app users

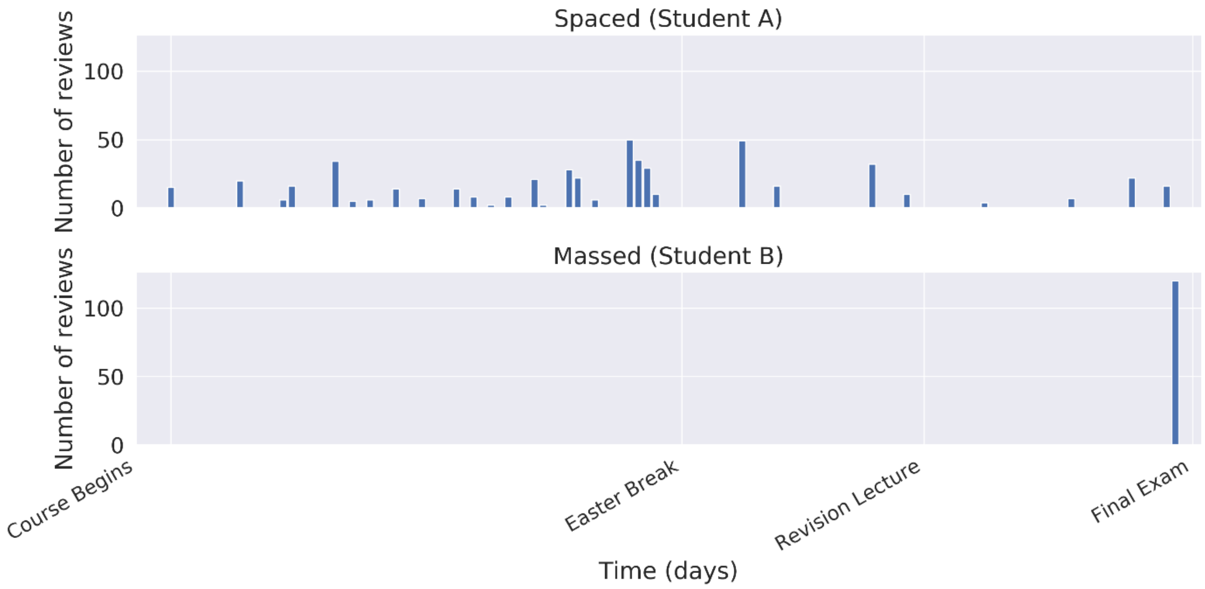

Figure 3 Timeline of app usage for two students typifying spaced practice (Student A) and massed practice (Student B).

Figure 1 shows the distribution of the semester one physics exam marks for students who subsequently became *app-users* and *non-users* within the semester two thermodynamics module. This demonstrates that students who chose to engage with the spaced repetition app in semester two spanned the full ability range of the class, as determined by their semester one performance. No statistically significant differences in the average semester one exam results for subsequent *app-users* and *non-users* were found in any of the trials (with p = 0.26, p = 0.26 and p = 0.30 respectively for 2016-17, 2017-18, and 2018-19) corresponding to mean differences of +4.1%, -2.9% and +2.6% respectively).

When did students engage?

Figure 2 shows a timeline of app usage for each cohort, revealing reasonably steady usage throughout the term with several spikes. In particular, for each trial spikes in usage were observed in the days immediately before the final exam, indicating that some students used the app to ‘cram for the exam’, despite being encouraged to space their practice throughout the term.

In fact almost half (45%) of the *app users* only used the app for massed, rather than spaced, practice (Table 2).The difference between students engaging in spaced practice (*spacers*) and those engaging in massed practice (*crammers*) is highlighted in Figure 3, which shows the usage timelines for two students most typifying these different behaviours.

Anecdotally, when surveyed at the end of the module both these students agreed that using the app enhanced their understanding (Student A ‘strongly agreed’, Student B ‘mostly agreed’), but when asked if the app kept topics fresh in their mind, Student A ‘strongly agreed’ and Student B ‘mostly disagreed’.

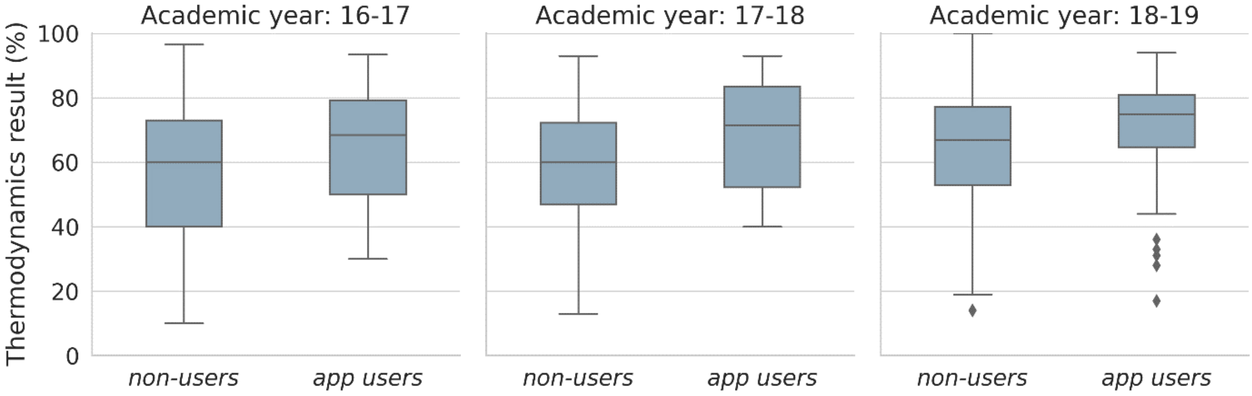

Figure 4 Box plots of unadjusted thermodynamics results, plotted for app users and non-users

Academic year	Adj. mean *non-users* (%)	Adj. mean *app users* (%)	Mean difference (%)	Effect size	p-value
2016-17	57	64	+7	0.33	0.039
2017-18	58	70	+12	0.71	0.00057
2018-19	66	68	+2	0.13	0.36

Table 3 ANCOVA-adjusted mean thermodynamics scores for app users and non-users, effect size and statistical significance of each trial under an intention-to-treat analysis.

Intention-to-treat analysis

To look in more detail at the impact of the app, the effect on the end of module thermodynamics exam was analysed. Firstly, students who made any use of the app (*app users*) were compared with students who did not use it at all (*non-users*). Figure 4 shows box plots of the thermodynamics exam scores for users and non-users in each cohort, revealing an increased mean for app users in all cases.

To control for any innate ability in physics or superior exam technique, the thermodynamics results were scaled by the semester one physics exam scores and labelled as 'adjusted' results. For the 2016-17 and 2017-18 cohorts *app users* had statistically significant higher ANCOVA-adjusted mean thermodynamics scores compared to *non-users* (p = 0.039 and p = 0.00057 respectively) corresponding to effect sizes of 0.33 (+7%) and 0.71 (+12%). In 2018-19 the difference in performance was positive but not found to be statistically significant. See Table 3 for complete results.

Combining student results across all three trials, *app users* and *non-users* were found to have adjusted mean scores of 67% and 61% respectively; corresponding to a statistically significant (p = 0.00045) positive effect size of 0.33. With app users spanning the ability range of the whole class this enhanced performance of over half a 'classification grade' is evidence of the benefit to all students.

Per-protocol analysis

Since only just over half of the *app users* engaged with the app in a spaced manner (Table 2), this second analysis focusses more closely on the effect of spaced repetition. It removes those students who undertook all their app usage on one day (massed practice), and thus compares *spacers* with *non-users*. Despite the spacers being a subset of app users, they still spanned the full range of performance in the semester one physics exam, from fail to 1st class. Statistically, no significant differences were found between the average semester one physics results for *spacers* and *non-users* in any of the trials (with p = 0.49, p = 0.88 and p = 0.88 respectively for 2016-17, 2017-18, and 2018-19; corresponding to mean differences of +3.1%, +0.47% and +0.47% respectively). Figure 5

https://doi.org/10.29311/ndtps.v0i15.3376

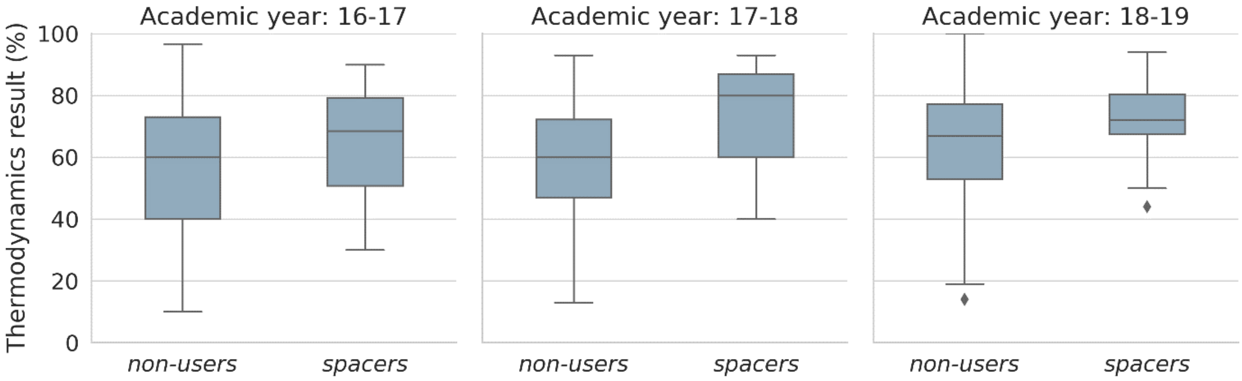

Figure 5 Box plots of unadjusted thermodynamics results, plotted for spacers and non-users.

Academic year	Adj. mean *non-users* (%)	Adj. mean *spacers* (%)	Mean difference (%)	Effect size	p-value
2016-17	57	64	+7	0.34	0.070
2017-18	58	73	+15	0.80	0.0014
2018-19	66	72	+6	0.38	0.037

Table 4: ANCOVA adjusted mean thermodynamics scores, effect size and p values for spacers vs non-users in each trial under a per-protocol analysis.

shows box plots of raw thermodynamics scores for *spacers* and *non-users*, revealing an increased effect for spaced usage, over all usage, implying spaced practice is superior to massed practice in terms of preparation for the exam.

Table 4 gives data adjusted for semester one performance in physics. In 16-17 the difference in adjusted performance between *spacers* and *non-users* was not found to be statistically significant p = 0.07. However, for both 2017-18 and 2018-19 the *spacers* performed significantly better (p = 0.001 and p = 0.037 respectively) with effect sizes of 0.80 and 0.38.

When combining student results across all three trials, *spacers* and *non-users* were found to have adjusted mean scores of 70% and 61% respectively, corresponding to a statistically significant (p = 0.000056) positive effect size of 0.47. This difference spans a whole degree class and is strong evidence to encourage students and staff alike to engage in spaced repetition during the learning of STEM material. In contrast, *crammers* did not show a statistically significant effect when compared to *non-users* (p = 0.26 adj, mean = 64%).

Delayed test

Since this intervention and analysis was undertaken in a real classroom situation, where students were free to study in other ways in addition to using the app, perhaps the ultimate test of the long term memory effect of this spaced repetition app, comes from the delayed test, delivered without warning at the start of the second year, after the summer vacation.

This delayed test, delivered to the 18-19 cohort, showed a positive difference in adjusted means between *app-users* (adjusted mean 40%) and *non-users* (adjusted mean 34%) with and effect size of 0.33 although this was not statistically significant (p = 0.084). However, when comparing *spacers* with *non-users* the adjusted means were 45% and 34% respectively, indicating a statistically significant effect size of 0.54 (p = 0.021). This compares favorably with the effect size of 0.41 determined by Kerfoot (2009) for delayed testing of medical students. Students that crammed did not perform significantly higher than *non-users* (p = 0.77, adj. mean = 36%).

This demonstrates the long-term memory benefit of spaced repetition within a STEM module under real teaching conditions. The lower adjusted means (delayed test vs thermodynamics exam) reveal that students have forgotten some of the material over the summer, in line with Ebbinghaus' forgetting curve (Ebbinghaus, 1885). But the effect of using the app in a spaced manner has significantly enhanced students' relative performance over non-users. Such evidence is important to encourage more students (and staff) to engage in spaced repetition practice, to help students develop a deeper knowledge and understanding of their discipline to advantage them throughout their degree. And in line with Roediger and Karpicke (2006b) this shows the benefit of testing as part of the learning process rather than just to assess what has been learned.

Conclusions

This research was undertaken to investigate student engagement with spaced repetition within a 1st year physics module in an attempt to foster better long-term memory retention of material. Such retention is desirable as a platform for more advanced study and to improve students' problem solving skills, a key graduate outcome and employability attribute. A custom-built web app was used to deliver questions on core laws, equations and concepts in thermodynamics, which had the facility to generate a personalised repetition timetable, and to monitor usage.

Trials were conducted on three successive cohorts of physics students (2016-17, 2017-18, 2018-19) with analysis encompassing 405 students in total. Of these typically 26% made use of the app in some way. For analysis students were post-identified as *app users*, *non-users*, and *spacers*, with spacers being a subset of users. Students in all three categories spanned the full ability range of the class, as determined by reference to their score in the physics exam immediately preceding the thermodynamics module.

To evaluate the effect of the app on memory and performance, the end of module exam scores in thermodynamics were 'adjusted' according to performance in the previous semester physics exam, to control for inherent physics ability and/or exam technique. Over the three cohorts, students who had engaged with the app in a *spaced* manner had an adjusted mean thermodynamics exam score of 70%, compared to 61% for *non-users*. This is essentially a full classification grade difference and represents a statistically significant (p = 0.000056) positive effect size of 0.47, demonstrating the power of spaced repetition in a STEM classroom setting.

More extended long-term memory was assessed in an impromptu delayed test after the summer vacation, yielding mean adjusted scores for spacers and non-users of 45% and 34% respectively. Whilst less material had been retained over the summer, this still revealed a statistically significant (p = 0.021) positive effect size of 0.54 for engagement with spaced repetition.

Students have a strong tendency to 'cram for exams' as this technique has served them well through high school and brings short term reward. But in higher education, and for employment, long term memory is far more beneficial and finding ways for students to engage and trust such a method is important. This research thus contributes key findings that should prove useful and persuasive to students, and staff, to develop and expand their study techniques. And in the absence of a spaced repetition app, students should still be encouraged and facilitated to engage in regular testing during the learning phase, rather than cramming immediately prior to an exam.

Acknowledgements

The authors would like to thank Derek Raine for his insight, interest and inspiration in this project, from its inception and throughout its execution. Thanks also go to Jasmin Curtis and Adrienne Lowe for their undergraduate project work in surveying and analysing student study behaviour, highlighting the tendency for students to leave self-testing until close to the exams.

References

Curtis, J., Lowe, A. and Voice, A. M. (2018) *Analysis of student study habits*. Unpublished manuscript, University of Leeds.

Ebbinghaus, H. (1885) *Memory: a contribution to experimental psychology (H.A.Ruger, C.E.Bussinius & E.R.Hilgard, Trans in 1964.)*. New York: Dover Publications

Gallo, M. & Odu, M. (2009) *Examining the Relationship Between Class Scheduling and Student Achievement in College Algebra,* Community College Review, 36(4), pp. 299-325.

Hopkins, R.F., Lyle, K.B., Hieb, J.L. & Ralston, P.A.S. (2015) *Spaced Retrieval Practice Increases College Students' Short- and Long-Term Retention of Mathematics Knowledge*, Educational Psychology Review, 28(4), pp. 853-873. DOI: 10.1007/s10648-015-9349-8.

Kapler, I.V., Weston, T. & Wiseheart, M. (2015) *Spacing in a simulated undergraduate classroom: Long-term benefits for factual and higher-level learning,* Learning and Instruction, 36, pp. 38-45. DOI: 10.1016/j.learninstruc.2014.11.001.

Kerfoot, B.P. & Brotschi, E. (2009) *Online spaced education to teach urology to medical students: a multi-institutional randomized trial*, Am J Surg, 197(1), pp. 89-95. DOI: 10.1016/j.amjsurg.2007.10.026.

Kerfoot, B.P., DeWolf, W.C., Masser, B.A., Church, P.A. and Federman, D.D. (2007) *Spaced education improves the retention of clinical knowledge by medical students: a randomised controlled trial*, Medical Education, 41(1), pp. 23-31 DOI: 10.1111/j.1365-2929.2006.02644.x.

Kornell, N. & Bjork, R.A. (2007) *The promise and perils of self-regulated study*, Psychonomic Bulletin & Review, 14(2), pp. 219-224. DOI: 10.3758/bf03194055.

Leitner, S. (1972) *So lernt man lernen*. Edited by Erfolg, A.L.f.z. Angewandte Lernpsychologie fuhrt zum Erfolg. Herder.

Mace, C.A. (1968) *The psychology of study*. Penguin.

Maxwell, S.E. and Delaney, H.D. (1990) *Designing experiments and analyzing data: A model comparison perspective*. US: Lawrence Erlbaum Associates Publishers *p434*.

Quality Assurance Agency (2014) *Subject Benchmark Statement: Chemistry*. [Online]. Available at: https://www.qaa.ac.uk/docs/qaa/subject-benchmark-statements/sbs-chemistry-14.pdf?sfvrsn=99e1f781_14.

Quality Assurance Agency (2017) *Subject Benchmark Statement: Physics, Astronomy and Astrophysics.* [Online]. Available at: https://www.qaa.ac.uk/docs/qaa/subject-benchmark-statements/sbs-physics-astronomy-and-astrophysics-17.pdf?sfvrsn=2f94f781_12.

Rawson, K.A., Vaughn, K.E. and Carpenter, S.K. (2015) *Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis*, Memory & Cognition, 43(4), pp. 619-633. DOI: 10.3758/s13421-014-0477-z.

Roediger, H.L. & Karpicke, J.D. (2006a) *The Power of Testing Memory Basic Research and Implications for Educational Practice*, Perspectives on Psychological Science, 1(3), pp. 181-210. DOI: 10.1111/j.1745-6916.2006.00012.x.

Roediger, H.L. & Karpicke, J.D. (2006b) *Test-enhanced learning - Taking memory tests improves long-term retention*, Psychological Science, 17(3), pp. 249-255. DOI: 10.1111/j.1467-9280.2006.01693.x.

Roediger, H.L. & Karpicke, J.D. (2018) *Reflections on the Resurgence of Interest in the Testing Effect*, Perspectives on Psychological Science, 13(2), pp. 236-241. DOI: 10.1177/1745691617718873.

Rohrer, D. (2009) *The Effects of Spacing and Mixing Practice Problems*, Journal for Research in Mathematics Education, 40(1), pp. 4-17.

Rohrer, D. & Taylor, K. (2007) *The shuffling of mathematics problems improves learning,* Instructional Science, 35(6), pp. 481-498. DOI: 10.1007/s11251-007-9015-8.