Revista Latinoamericana
de Psicología

http://revistalatinoamericanadepsicologia.konradlorenz.edu.co/

ORIGINAL

Improving retention by placing retrieval practice at the end of class: a naturalistic study

Roberta Ekuni[a][b]*, Sabine Pompeia[b]

[a] *Universidade Estadual do Norte do Paraná, Brazil*
[b] *Universidade Federal de São Paulo, Brazil*

Received 1 April 2019; accepted 29 November 2019

KEYWORDS
Testing effect,
memory,
evidence-based learning

Abstract Attempting to recall previously encountered information by being tested or quizzed (retrieval practice: RP) enhances memory. In real classrooms, however, it is unclear when testing should take place (placement) in order to elicit better learning. We tested students using authentic undergraduate-course materials with two placements, followed by collective feedback: (a) at the end of the class in which content was taught; or (b) at the beginning of the next class. Re-teaching (RT) the same content through lecturer led-reviews at the same placements was used as a control condition. RP and RT plus feedback took 15 min of 100 min-long classes and were applied during 12 classes after which retention was assessed by exams. Participants were 114 students enrolled in a biweekly taught course. Testing (RP) once at the end of the same class in which content was taught boosted academic scores by around 10% compared to the other manipulations.

Mejorando la retención al realizar la práctica de recuperación al final de la clase: un estudio naturalístico

PALABRAS CLAVE
Efecto de las pruebas,
memoria,
aprendizaje basado
en la evidencia

Resumen El intento de recordar la información presentada previamente mediante pruebas o cuestionarios (práctica de recuperación, RP) mejora la memoria. Sin embargo, en las clases reales no está claro en qué momento debe realizarse las pruebas para obtener un mejor aprendizaje. Probamos a los estudiantes utilizando materiales auténticos de cursos de licenciatura en dos momentos, seguidos de una retroalimentación colectiva: (a) al final de la clase en la que se enseñó el contenido; o (b) al principio de la siguiente clase. La reenseñanza (RT) del mismo contenido a través de revisiones dirigidas por el profesor en el mismo momento se utilizó como condición de control. La RP y la RT más la retroalimentación tomaron 15 minutos de clases de 100 minutos de duración y se aplicaron durante 12 clases después de las cuales la retención fue evaluada por exámenes. Los participantes fueron 114 estudiantes inscritos en un curso impartido cada dos semanas. Las pruebas (RP) realizadas una vez al final de la misma clase en la que el contenido se enseñó, aumentó los resultados académicos en alrededor de un 10% en comparación con las otras manipulaciones.

* Autor para correspondencia.
e-mail: robertaekuni@uenp.edu.br

https://doi.org/10.14349/rlp.2020.v52.3

Evidence-based teaching requires research findings to move from the laboratory into the classroom (Agarwal, Bain, & Chamberlain, 2012). One way of boosting learning in naturalistic settings is to use retrieval practice, also known as the testing effect (Roediger, Finn, & Weinstein, 2012; Roediger & Karpicke, 2006). Retrieval practice involves testing or quizzing students on previously presented content. This not only enables learners to assess the extent of their retention (Roediger, Agarwal, McDaniel, & McDermott, 2011), but also makes it more likely that information will be remembered for longer periods of time (Adesope, Trevisan, & Sundararajan, 2017; Roediger & Karpicke, 2006; Roediger, Putnam, & Smith, 2011; Rowland, 2014). Some authors have suggested that retrieval practice is not always effective or robust in educational settings, in which complex information is taught (van Gog & Sweller, 2015). However, these arguments have not withstood scrutiny and the benefits of retrieval practice seems to clearly offset any disadvantages (Karpicke & Aue, 2015; Rawson, 2015). In this paper, we will refer to “tests” or “testing” as opportunities to practice retrieval of information before final assessment of memory retention, which we will term “exams” (corresponding to finals or similar terms in usual teaching practices).

Despite the large body of data showing retrieval practice’s efficiency for promoting lasting learning in laboratory experiments and in classroom settings (Adesope et al., 2017; Roediger & Karpicke, 2006; Rowland, 2014), teachers generally use testing and exams as a means of evaluating acquired knowledge (Roediger & Karpicke, 2006) rather than a teaching or studying technique. To change this, teachers should realize that testing could also be an effective teaching tool that does not involve changing *what* is taught. They should simply get students to practice remembering information more often. This need not take up too much of teachers limited classroom time (see Leeming, 2002). In fact, retrieval practice may lead to more efficient use of time since improved retention reduces the need to re-explain content (Pyc, Agarwal, & Roediger, 2014). Furthermore, repeated testing decreases test-anxiety (Agarwal, D’Antonio, Roediger, McDermott, & McDaniel, 2014) and can even improve students’ perception of their teachers (Bangert-Drowns, Kulik, & Kulik, 1991).

One testing opportunity pertaining to each taught content before long-term evaluation of retention has been shown to have the strongest effect on learning (Adesope et al., 2017; Bangert-Drowns et al., 1991; Bjork, Little, & Storm, 2014). Repeating testing of the same content more times can also improve retention, but to a much smaller extent (Rowland, 2014). Therefore, to maximize retrieval-practice benefits while minimizing their burden, teachers should test students on previously taught content at least once before retention is assessed in the longer term (exams). Regarding test format, cued- or free-recall (short-answer tests) questions favour retention (Kang, McDermott, & Roediger, 2007; Rowland, 2014) and multiple-choice tests also work well (Adesope et al., 2017; Karpicke, 2017; Rowland, 2014). However, the number and plausibility of alternative answers in multiple-choice tests can influence results and, in some cases, even lead students to acquire incorrect knowledge (see Karpicke, 2017). Hence, feedback should be provided after retrieval practice to enable students to correct any misunderstandings (Butler & Roediger, 2008). If teachers are pressed for time, collective feedback can be used, that is, presentation of test answers to the whole class instead of having to review each students’ answers (Butler, Karpicke, & Roediger, 2007). When short-answer questions are used for testing during class, exams may consist of multiple-choice questions, which are easier to mark. Varying the format of tests and exams is a good way of assessing retention because students are led to diversely retrieve what they have learned (see Kang et al., 2007). As such, exams also become additional learning opportunities (McDaniel, Anderson, Derbish, & Morrisette, 2007).

One question that has not yet been fully answered, which we aim to assess in this paper, is when (placement) to apply tests in real classroom environments. Although Rowland’s (2014) meta-analysis showed that the lag between presenting information and testing it did not moderate testing effects, many laboratory studies have shown this lag to be important. Tests in these laboratory studies were usually applied only seconds or a few minutes after content was presented (Karpicke & Roediger, 2010; Pyc & Rawson, 2009), while retention intervals (from testing to exam) typically spanned minutes to hours, or sometimes up to a couple of days (McDaniel, Roediger, & McDermott, 2007). The findings from these studies are evidently of little use when it comes to deciding what to do in classroom situations. Testing students every few minutes can interrupt the flow of lessons. Also, the aim of education is to make information retrievable in the long term rather than after just a few minutes, hours, or days. Additionally, teachers have access to students only on certain days of the week, so they have little flexibility in deciding when to place testing.

Some naturalistic studies propose more practical solutions to overcome this issue of placement, such as testing at the end of each class (Lyle & Crawford, 2011; McDaniel, Agarwal, Huelser, McDermott, & Roediger (2011) (experiment 1, 2a, and 2b) and Roediger, Agarwal, et al. (2011) (experiments 1 and 2)) or in the next class (Bjork et al., 2014; Leeming, 2002; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013). Both placements could have advantages. Testing some days after teaching content introduces some desirable difficulties (see Bjork & Bjork, 2011), since answering questions becomes harder after time has elapsed and forgetting (Roediger & Karpicke, 2011) and/or interference (Hays, Kornell, & Bjork, 2012) have taken hold. In contrast, testing for retention of concepts soon after they have been presented may be facilitated by factors such as: (1) more recent activation of previously formed semantic networks while learning new content (Carpenter, 2009); (2) temporal and contextual cues from a recent learning episode (Karpicke, Lehman, & Aue, 2014; see also Radvansky & Zacks, 2017); (3) less interference from new content encountered by students after a class (Hays et al., 2012); and/or (4) less time during which forgetting can occur (Roediger & Karpicke, 2011). All these factors may enhance the effects of retrieval practice in the classroom, possibly by allowing associative networks to be strengthened, reconsolidated, and more efficiently re-organised, which is in line with the proposed mechanism through which retrieval practice works (Bjork, 1975; McKenzie & Eichenbaum, 2011; Roediger & Butler, 2011; Rowland, 2014). Studies that tested the effects of retrieval practice at the end of class or the beginning of the next class, however, did not directly compare these manipulations, which was the objective of the present study.

We sought to determine the best testing/quizzing placement to enhance retention in a manner that can feasibly be applied by teachers, using authentic study materials in undergraduate classes. Students were submitted to one test or retrieval practice opportunity (short-answer questions) pertaining to content taught in one class that was part of a sequence of 12 Educational Psychology classes. We measured how this affected retention of information in final, multiple-choice exams when testing was placed either: (a) at the *end* of the *same* classes in which the content was taught; or (b) at the *beginning* of the *next* classes. Thus, twelve retrieval practice opportunities occurred many minutes to days after students were exposed to new course information. However, each content was tested only once. Spreading testing over time in this way has been shown to be advantageous for lasting learning (Adesope et al., 2017).

Our control condition was re-teaching the same content that was tested at both placements. We chose this condition because re-reading content, the usual control condition for the type of experiment conducted in this paper (Adesope et al., 2017; Roediger, Agarwal, et al., 2011), has been shown to be inefficient (Rowland, 2014; Adesope et al., 2017). We considered this manipulation to be unethical because scores on final exams in the present naturalistic experiment were part of students' final grades. Furthermore, although retrieval practice is more beneficial when compared to "shallow" encoding conditions, such as rereading, in naturalistic studies there is less evidence that testing works better than activities that are usually employed in the classroom, such as concept mapping and lecturing (Moreira, Pinto, Starling, & Jaeger, 2019). Because teachers often highlight or direct attention to critical facts (Kornell, Rabelo, & Jacobs, 2012), which can be characterized as a "deeper" form of encoding than rereading, we opted to use lecturer-led reviews of content (Bol & Hacker, 2001) as a control condition. This type of re-exposure to content, which we will call "re-teaching", has been shown to be a powerful means of improving learning, and is on a par with testing (Bol & Hacker, 2001; Kang et al., 2007). However, studies that have compared these manipulations have not varied testing placement. Therefore, we reasoned that if retrieval practice in one of our placements were found to improve learning more than usual teaching practices, this could increase the credibility of this technique as a teaching tool. We hypothesized that testing during the next class would be more advantageous because studies tend to show that longer intervals between being exposed to content and being tested enhances long-term storage (Adesope et al., 2017). This occurs in part because learning seems to be strongly dependent on the amount of effort required to answer tests (see Adesope et al., 2017).

In order to maintain control over the re-teaching manipulation, the same experienced lecturer taught all classes. She had access to two classrooms of students in each year, so the experiment took place over two consecutive years. Each classroom had lessons at a different time of day (afternoon or evening). Undergraduates tend to be chronically sleep deprived when they have to wake up early in the morning to study or work due to their chronotype being shifted two to three hours later in the day during this phase of life (Evans, Kelley, & Kelley, 2017). Because there is a strong body of evidence that students tend to learn best when they are most alert (Levandovski, Sasso, & Hidalgo, 2013; Preckel et al., 2013; Rahafar, Maghsudloo, Farhangnia, Vollmer, & Randler, 2016), our study design considered the time of day at which participants' classes were scheduled (by reversing the manipulations considering time of class in the two consecutive years). Sex was also taken into account because it can influence academic performance. Female students consistently obtain better grades than male students (Voyer & Voyer, 2014), and men and women also differ in terms of academic self-efficacy, that is, in their belief in their ability to achieve intended results (Huang, 2013), which can affect how they study. Additionally, age can have significant effects on how people remember information (e.g. Old & Naveh-Benjamin, 2008), so we controlled for age in our analyses.

Method

Participants

Participants were 131 undergraduate students majoring in Biological Sciences or Information Systems at the Universidade Estadual do Norte do Paraná. Their native language was Portuguese and they were taking an Educational Psychology course. The Ethics Committee approved the experimental procedures (approval number: 670,895), and all students provided written informed consent. They were told at the beginning of the course that all the manipulations they would be exposed to had been found to be good for learning.

Experimental Design and Procedure

We manipulated three variables: type of re-exposure to materials (retrieval practice (RP) versus re-teaching (RT)), placements (at the end of the *same* class (RPsame or RTsame)) or at the beginning of the *next* class (RPnext or RTnext), and time of day when courses were taught (afternoon versus evening), which coincided with the degrees participants were pursuing (see Figure 1).

We had access to four classes taught twice a week by the same experienced lecturer (the first author) for two consecutive years. In each year there were two types of students: one class was undergraduates studying Information Systems (with classes on Monday and Wednesday evenings), and the others were Biological Sciences undergraduates (afternoon course on Mondays and Thursdays). Therefore, with a few exceptions (e.g. bank holidays), the lags between classes and testing or re-teaching were three and four days for the Information System degree (Monday to Thursday and Thursday till the following Monday) and two and five days for the Biological Sciences degree course (Monday to Wednesday and Wednesday till the following Monday).

Each course had 24[1] lessons during each semester, divided into two blocks with different sequential manipula-

[1] The RPnext or RTnext conditions included an initial class in which content was taught but no retrieval practice or re-teaching took place. The manipulations were applied to the following 12 classes, the last of which included only RP or RT relative to content from the prior class, followed by other activities that did not involve teaching new information. The RPsame and RTsame conditions included an initial class in which no content was taught, followed by 12 classes with taught content, and at the end there was RP or RT. Therefore, manipulations (RP or RT) were applied only during 12 classes in each block.

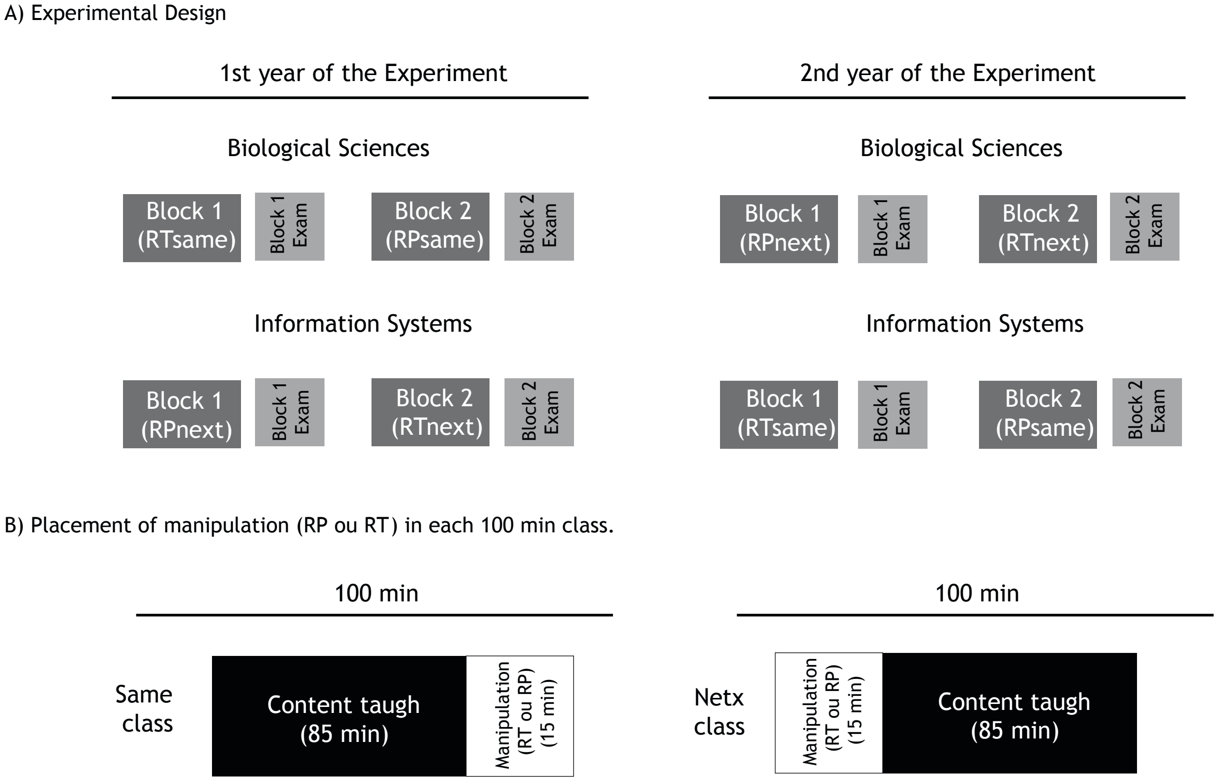

Figure 1. A) Experimental design: four groups of students were enrolled for two consecutive blocks of 12 classes each (two classes per week) during which testing was manipulated. Retention of taught material was assessed at the end of each block by using multiple-choice tests including three questions on content tested (RP) or Re-Taught (RT) in the block (totalling 36 questions). B) Placement of manipulations: In the Retrieval Practice condition (RP), students answered three short-answer questions (followed by blocked feedback) either at the end of the *same* class during which tested material was taught (RPsame) or at the beginning of the *next* class (RP next). The control lecturer-led review condition (Re-Teaching – RT) also occurred either at the end of the class in which content was taught (RTsame) or at the beginning of the next class (RTnext).

tions. In the first year of the experiment, each of the two classrooms of students was submitted to either the *same* (RTsame followed by RPsame) or the *next* (RPnext followed by RTnext) condition. This was done so as not to confuse students who, in Brazil, often turn up late or leave before the end of class. They were informed that during the course they should not be late (*next* conditions) as manipulations occurred at the beginning of class, or that they should stay until the end of class (*same* conditions) as the manipulations took place at the end of class. Hence, in the *next* condition, participants always experienced retrieval practice in the first block and re-teaching in the second, while in the *same* conditions, participants always experienced a block of re-teaching followed by a block of retrieval practice. After the twelfth class in every block, participants underwent a block exam (respectively, block exam one and two) which included questions related to the content taught over the prior 12 classes. This exam was used to assess retention of information under each manipulation.

In the second year of the study, we kept the same protocol (with new students pursuing the same undergraduate degrees); the only exception was that we alternated manipulations across the two degree options taught at different times of day (indicated by arrows on Figure 1). The purpose was to control for time of day in which classes were held, a factor that is mostly ignored in the retrieval practice literature. We believed that time of class could be a more determinant factor for learning than the order of manipulations within blocks in each of the *same* and *next* conditions. Hence, although we controlled for time of class/course in our model and, consequently, for the number of days (lags) between the classes and the *next* manipulations, our manipulations were confounded with order: each classroom of students was subjected to two sequential manipulations and the total counterbalancing of testing/re-teaching, *same* and *next* conditions was not obtained (see Figure 1). This was taken into account in the statistical analysis, which controlled for the between/within-participant and sequential nature of the variables (see below).

Each class lasted 100 minutes; 85 min of which were for teaching content, and the remaining 15 min were for retrieval practice (testing) or re-teaching. This happened at either the beginning of class (*next* conditions, pertaining to information presented in the last class) or at the end of each class (*same* conditions, pertaining to content presented in that same class). In the retrieval practice conditions, students used the first ten minutes of the 15 minute period to answer three short-answer questions (self-paced), while the final five minutes were used for collective feedback (delayed blocked-practice feedback; Kornell & Vaughn, 2016; Rowland, 2014). To ensure students were paying attention during feedback, they were asked to mark their answers as correct, partly correct, or wrong. The lecturer reviewed all answers and scored them 1 (correct), 0.5 (partially correct), or 0 (wrong) to make sure students had marked them correctly. Consequently, scores for short-answer questions for each class ranged from 0 to 3. Students were not graded on these results (no-stakes testing). We compared mean retrieval accuracy for the daily retrieval practice tests in the RPsame and RPnext conditions because getting test answers right can influence retrieval practice effects, although exactly how is still unclear (see Karpicke, 2017; Rowland, 2014).

In order to equate overall exposure to materials in the retrieval practice and re-teaching conditions (see Rowland, 2014), during re-teaching the lecturer reviewed/summarized the same information that had been assessed in retrieval practice tests by briefly reviewing content. During this time, students were allowed to interact with the lecturer and ask questions, but the lecturer did not ask any questions because this would characterize retrieval practice (Smith, Roediger, & Karpicke, 2013).

The participants, under all manipulations, could keep notes taken in class and during re-teaching. However, under the retrieval practice conditions they were not allowed to keep copies of the quizzes because the questions focused on the content of the target material that would be used in the block exams. Hence, if they used these tests to study, they would have an advantage over those exposed to re-teaching. Note taking during RT was different, as it was based on what aspects of the re-taught information was regarded *by the student* as important and not only what the teacher selected as being the core of the class (the target material).

Retention of material (percentage of correct answers) was measured based on the final exams following the end of each block, which corresponded to two thirds of students' grades in the course. These exams were multiple-choice, self-paced tests covering all content that had been tested/re-taught in that particular block (three questions from each of the 12 classes, totalling 36 multiple-choice questions). As the classes during which the final block exam was applied lasted 100 minutes, students had approximately three minutes to choose answers for each multiple-choice question.

Materials

The Educational Psychology course content was based on Consenza and Guerra (2011) and Santrock (2009) and divided into 24 classes (Appendix, Table 1A). Based on the content in each class, three relevant facts were used to compose retrieval-practice short-answer questions and/or to be retaught. For the final exams, we composed multiple-choice tests from all selected relevant facts that were tested/re-taught in all classes in that block; there were 36 questions, each with five alternative answers. The order of the alternatives was randomly positioned among participants. Thus, we used a cross-format assessment (short answers in the tests, and multiple-choice for exams measuring retention).

To reduce the possibility of the experimental design and material influencing results, with a few exceptions, the tested/re-taught content and short-answer and multiple-choice exam questions were different for the two years during which the experiment was carried out. An example is shown in Table 1. However, the same questions were used for each year when two parallel manipulations were conducted at the same time in different classes/blocks.

Care was taken to create test questions based on content presented in the same/prior class, depending on the manipulation. However, as this was a naturalistic study, content from various classes was inevitably repeated throughout the course, which included unstructured, semantically themed information, and also semantically related materials (see Rowland, 2014). Therefore, some questions may have been easier to answer if students had understood content from previous classes.

Statistical Analyses

The level of significance was set at 5%. The proportion of male and female students in classrooms was compared using a Chi square test, and the other data were analysed using univariate General Linear Models (GLM): factors and levels are detailed below). *Post hoc* contrasts were carried

Table 1 Examples (translated from Portuguese) of short-answer questions used during a class and the multiple-choice final test question on the same content (cross-format assessment, similar to McDaniel et al., 2007). The correct multiple-choice alternative is marked in bold.

Short-Answer	Multiple-choice
What is indirect instruction? Name one positive and one negative characteristic of this method.	Considering the concept of indirect instruction, choose the correct alternative: a) Indirect instruction is centred on the teacher; a positive point of this method is that learning is faster. b) Indirect instruction is centred on the teacher; a negative point of this method is that it does not use playful activities. **c) Indirect instruction is student** centred**; a positive point of this method is the interpersonal relationship between the teacher and the student.** d) Indirect instruction is student centred; a negative point of this method is that it is not focused on the student's reflective abilities. e) Indirect instruction is student centred; a positive point of this method is that learning is faster.
What are the three dimensions of executive functions that are worth highlighting?	What are the three dimensions of executive functions that are worth highlighting? a) Executive memory, shifting, and attentional control b) Working memory, attentional control, emotions c) Updating, planning, language **d) Working memory, inhibitory control, mental flexibility** e) Mental flexibility, shifting, and updating

out with Tukey's Honest Significance Difference (HSD) tests for samples of different sizes, a test that corrects for multiple comparisons. Effect sizes were measured with Cohen *d* values (Cohen, 1998), which were either corrected (or not) between/within participant comparisons, depending on the contrasting values.

To compare students' ages in the four classrooms, we used a between-participant univariate GLM with the classroom as a categorical factor (four levels). As there was no difference among classes, age was not entered as a covariate into the following models. To compare the mean number of correct answers to short-answer tests in the RP conditions, we used a univariate GLM including *same* vs. *next* as a two-level factor. We then included sex and time of class as categorical predictors to determine whether they influenced results.

Differently, to determine the effects of retrieval practice and re-teaching at different placements (same and next class), the dependent measure used was retention (percentage of correct answers) in the final block exams, which was analysed with a mixed (between- and within-participant) GLM. We obtained data on retention under four conditions: RPsame, RPnext, RTsame, and RTnext. However, some participants took part in two consecutive manipulations that we aimed to compare (within-participant), while other comparisons involved data obtained from different people (between-participant). Thus, treating these four manipulations as independent would be statistically incorrect because part of the data was obtained from the same person and part was not. To account for this, our mixed model included the following factors: (1) order of manipulation, a between-subject factor (two levels: RTsame followed by RPsame vs. RPnext followed by RTnext); and (2) block, which was a within-participants factor (two levels): first block (either RPsame or RPnext, depending on the two orders of testing in each class) vs. second block of manipulation (either RTsame or RTnext). This statistical model, therefore, takes into consideration the between- (different classrooms) and within-participant (different blocks) consecutive design.

If only order of manipulation were found to be significant, this would mean that possible carryover effects had occurred. In other words, one sequence was different from the other. If we only obtained a block effect, this would mean that students faired differently in learning the content presented in the first and second blocks, or that there were carry-over effects. Conversely, an interaction of these factors would reveal a difference between the four manipulations that was true, even when considering that part of the data was from the same or from different individuals and that the effect was not solely due to order of manipulations or block, but to the nature of the manipulation in terms of retention of information. Next, we added to the model the categorical factor time of day at which classes were scheduled, which corresponded to the different courses (two levels: biological sciences (afternoon) vs. information systems (evening)) and sex, because the proportion of men and women differed between classrooms.

Results

All enrolled students agreed to participate, but 17 dropped out of the class for reasons unrelated to the experiment, so our final sample consisted of 114 students (aged 20.0±3.0: mean±SD) who were majoring in: Biological Sciences (afternoon) (first year: *n* = 23; 8 men; second year: *n* = 30; 8 men); Information Systems (evenings) first year: *n* = 28; 27 men; second year: *n* = 33; 25 men). There were no age differences between students in the four classes (*p* =.38), so age was not used as a covariate in the statistical models mentioned below. However, there were more males majoring in Information Systems and more females in Biological Sciences (Chi Square (*df* = 3) = 38.76, *p* < .0001), so sex was always included in the models. In the first year of the experiment, there was a teachers' strike half-way through the course, so data were not collected for the last six classes of block one. In these cases, retention was measured as a proportion of correct answers based on information given in the classes. Block two took place normally. In the second year of the experiment, six participants did not sit one of the block exams. Since there were at most two missing data points for each of the four experimental conditions, we substituted their missing scores for their classmates' mean scores. This was done because GLM exclude all data from a participant if there are missing data, so our sample would have been reduced by six participants even though they had valid data in the other blocks.

Scores on short-answer questions throughout the course

In the class by class testing, mean accuracy was higher for the RPsame (mean±SD: 2.29±0.33) than it was for the RPnext condition (1.56±0.46, or 52%), *F*(1,110) = 91.98, *p* < .001, a difference that reached a large effect size (*d* = 1.82). Sex and time of day when classes were scheduled were not significant factors.

We reran this analysis considering mean correct answers only in the first half of the blocks (first six classes) because we had less data on the second half of the two blocks due to the teacher strike during the first year of the experiment. The same findings were observed, *F*(1,109) = 81.22,*p* < .001; *d*=1.38. Importantly, there were very few cases in which students marked their answers in the collective correction differently from the teacher, so collective feedback worked well. Because of the great similarity between both sets of corrections, there was no variance in the model, so a statistical comparison between the teacher's and students' ratings was not possible.

Retention Scores on the Multiple-Choice Block Exams

Regarding retention (proportion of correct answers) in the final block exams, the analysis only showed a two-way interaction between order of testing and block: *F*(1,112) = 6.50, *p* < .02. *Post hoc* contrasts with corrections for multiple comparison indicated that the interaction resulted from higher retention in the RPsame condition compared to all the other manipulations (Tukey test *p* values < .03). Effect sizes were medium to large: RPsame vs. RTsame (*d* = .78, corrected for within-subject analysis), RPsame vs. RPnext (*d* = .85), and RPsame vs. RTnext (*d* = .57) (Figure 2). When we included the time of class/course and sex as controls in the model, these factors had no effect (*p* values > .21), and the above-mentioned interaction was unchanged. We reran this analysis considering the percentage of correct answers in the block's final exams that only pertained to the first six classes in each block. These classes were further in time from the final block exams, so scores could indicate longer-term retention. The interaction was once again observed $F(1,106) = 4.91$, $p < .03$, with the same pattern of results: RPsame led to better recall than all other manipulations.

Discussion

In this experiment, testing content at the *end* of the classes in which the content was taught [RPsame, following McDaniel et al., 2011 (in experiments 2a and 2b); Roediger, Agarwal et al., 2011 (in experiments 1 and 2); Lyle & Crawford, 2011] was more effective in promoting lasting learning than testing in the next class (as in Bjork et al., 2014; Leeming, 2002; McDaniel et al., 2013) and re-teaching the same content at either placements. The increase in retention reached medium to high effect sizes and was around 10% higher than the other conditions, mirroring gains in retrieval practice studies compared to manipulations such as rereading (Adesope et al., 2017; McDaniel et al., 2011 (experiment 2b); Roediger, Agarwal, et al. 2011).

In contrast to the advantage of RPsame over the other manipulations, retrieval practice did not have the same advantage when it took place in the class following the one in which the tested content was taught (RPnext). This is not surprising for the following reason: learning can be strengthened by drawing attention to relevant material in a wide range of ways (see Bol & Hacker, 2001; Butler & Roediger, 2007; Kang et al., 2007), which, in the present case, was undertaken through re-teaching content. This indicates that students were engaged by both lecturer-led reviews and retrieval practice but that, nevertheless, the RPsame manipulation enhanced retention more than RPnext, which is a novel finding and will be further detailed below. However, lower retention in the RPnext condition contrasts with the advantage of retrieval practice at this placement reported by Leeming (2002) and Bjork et al. (2014). However, these authors compared retention after retrieval practice with doing nothing, so simple re-exposure to content rather than retrieval practice as such could account for their results, as discussed by Moreira et al. (2019) and Adesope et al. (2017). Likewise, studies that showed beneficial retrieval practice effects by placing testing at the end of each class compared this manipulation with either testing before the class was taught or testing all content during the last class before the exam (McDaniel et al., 2011; Roediger, Agarwal, et al. 2011). These studies did not manipulate different testing placements *after* the content was taught, as we did, so their results do not contradict ours.

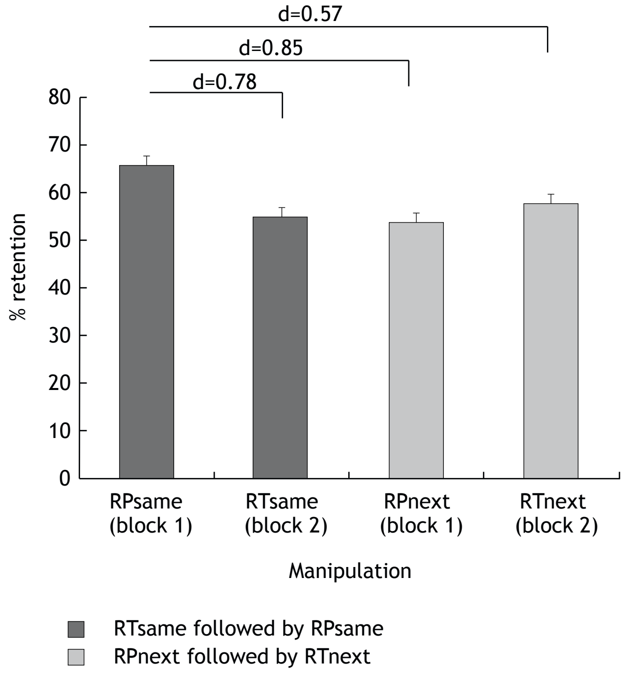

Figure 2. Mean (±SE) percentage retention in the four experimental conditions: Retrieval Practice at the end of the same class (RPsame); Retrieval Practice at the beginning of the next class (RPnext); Re-Teaching at the end of the same class (RTsame); Re-Teaching at the beginning of the next class (RTnext).

N.B. There was an interaction of block and order of manipulation (see text). RPsame> all other manipulations (*post hoc p* values <0.03). Effect sizes of the differences are indicated by Cohen *d* values.

The finding that testing soon after presenting information (end of class: RPsame) enhanced long-term retention compared to testing days later (RPnext) did not concur with our initial hypothesis, which was based on results from some laboratory studies that tend to show that longer intervals between presentation and testing (initial test lags) makes retrieval practice more advantageous (e.g. Adesope et al., 2017; Roediger & Karpicke, 2011). In laboratory studies, the intervals between presentation of to-be-remembered content and testing were much shorter and their material was much simpler than ours (e.g. a small number of word pairs or vocabulary tested after a few seconds or minutes (Karpicke & Roediger, 2010; Pyc & Rawson, 2009)). It follows that requiring answers at the end of class can also constitute a lag that is long enough to maximize long-term recall, especially as the content was more complex than usual information studied in most investigations conducted in the laboratory. This lag was found to be better than the one separating the class in which content was presented from the next class (a few days). We are unaware of studies that have discussed the best initial test lags in relation to the type of content or memory (semantic or episodic) that is involved in different kinds of retrieval practice conditions. Therefore, there are no published results to contrast our findings with.

It makes sense that learning associations between words presented in an experiment, for example, improves with contextual reinstatements of a specific episode, which can be strengthened by testing at longer lags when contextual and temporal cues have shifted (see Karpicke et al., 2014). Conversely, material learned in class must be bound to students' semantic knowledge and not to the episode or class in which the concept was taught. In this sense, context reinstatement may not be a major driver in promoting better retention. Learning facts regardless of episodes may, in turn, benefit more from semantic elaboration (Carpenter, 2009) during testing. Our findings show that this may be enhanced if retrieval practice takes place shortly after or during the same learning episode, possibly because recently activated semantic networks would have suffered less from forgetting (Hays et al., 2012) and/or retroactive interference (Szpunar, McDermott, & Roediger, 2008) than when testing occurs days later. This may explain why, in Rowland's (2014) meta-analysis, initial test lags, or placement of the first testing opportunity did not moderate testing effects. Naturalistic and laboratory investigations may differ in this respect, leading to high variability in results that preclude findings from becoming statistically significant.

Another factor that could have influenced better long-term recall after the RPsame manipulation is that students answered more tests correctly than in the RPnext manipulation, which was probably due to there being less time for interference and forgetting to occur. Higher accuracy at testing can positively influence long-term retention, strengthening memory traces (e.g., Karpicke, 2017; Karpicke et al., 2014) and facilitating reconsolidation (McKenzie & Eichenbaum, 2011), corroborating our findings. But the opposite has also been shown (see Rowland, 2014), perhaps because of the differing importance of context for recall in the laboratory and naturalist studies discussed above, which may involve more episodic and semantic memory, respectively. We thus propose that, in the classroom, processes that facilitate retrieval (RPsame) during testing may override the benefits of context reinstatement and/or larger retrieval effort (Radvansky & Zacks, 2017) needed to counteract forgetting or interference. Nonetheless, our proposal and results warrant confirmation, as we found no studies that contrasted comparable test placement manipulations in the classroom and in the laboratory or that discussed in detail the different types of material learned in these kinds of study.

A final potential explanation for the better recall at test in the RPsame condition, compared to the other manipulations, is that we used no-stakes tests. Under these conditions, participants may not have been encouraged enough to exert the extra-effort needed to recall information in the next class; this contrasts to the easier recall conditions for those who had just learned the lessons in the same class. It remains to be established whether using low-stakes testing, during which cognitive demands are met with more incentive (e.g. Roediger, Agarwal, et al., 2011), could alter the results found in this paper. We believe this is unlikely because the final block exams were high stakes (corresponded to two thirds of participants' grades), so using the proposed study techniques in class should have been motivation enough to do well. After all, all students were told that the techniques worked before entering the experiment. Also, there are countless studies mentioned throughout this paper that involved no incentives to remember, and they still found retrieval practice effects.

Despite strong associations between sleep and academic achievement (Evans et al., 2017; Levandovski et al., 2013; Preckel et al., 2013; Rahafar et al., 2016), having classes/practicing retrieval in the afternoon or evening did not alter the pattern of results, suggesting that, irrespective of alertness, undergraduate students benefit from being tested at the end of class, soon after learning content. Results might well have differed had we used classes that took place in the morning when undergraduates are usually not at all alert (Evans et al., 2017). Sex also failed to affect findings, which shows that differences in academic achievements (Voyer & Voyer, 2014) and self-efficacy (Huang, 2013) between men and women cannot explain our results and are unlikely to influence the benefits of end-of-class retrieval practice.

There are limitations to our study, which are mainly due to its naturalistic nature. Ideally, total counterbalancing of our manipulations should have been obtained, but it would have been unrealistic to get students to comply with not being late for class in part of the course and then staying until the end of classes in the other part. For this reason, we maintained the *same* and *next* conditions fixed in each class and controlled for this in our statistical model. It must be stressed that our participants were not usual experimental participants. They were volunteers who accepted to take part in the experiment during their undergraduate studies. Moreover, testing effects seem to be robust irrespective of a blocked or mixed design. Despite our fixed ordering of manipulations, there were no spillover effects (see McDaniel, Anderson et al., 2007; McDermott et al., 2014) from the use of retrieval practice from block one to block two because we found no benefit for retention in the RTnext (block two) in comparison to RPnext (block one). There was also no way of reliably assessing out of class studying (see McDaniel, Anderson et al., 2007), or controlling for the possibility that understanding prior content enhanced retention of subsequent material, but neither of these factors can account for our results. Another factor that was beyond our control was a teacher strike that interrupted part of the experiment, but our statistical analyses considered this and found the same effects. A possible concern could be that test scoring was not blind (the first author was the teacher), but the results did not confirm our predictions so it is unlikely that this affected results: we expected that a longer interval between content exposure and testing (lag) would be most beneficial, as this has been repeatedly shown by laboratory studies (Karpicke & Roediger, 2010; Pyc & Rawson, 2009), but, in fact, testing in the same class led to better retention. Additionally, in the RT condition, students could take notes and keep them, while those in the RP condition could not take home the tests done in class. This was the case because studying by using tests would give students an advantage over those who were involved in re-teaching, as tests comprised the target information that was used in the block exams. However, this was true of both RPsame and RPnext conditions, so this probably did not lead to our results. Finally, future studies should seek to replicate our findings by manipulating all conditions within the same participants.

In conclusion, our ecological study found that retrieval practice at the end of class with tests that involved content presented in that same class was the best manipulation in terms of boosting academic performance. This should be confirmed in more counterbalanced experimental designs. We believe that this finding is not restricted to learning content in Educational Psychology by undergraduates. Dunlosky, Rawson, Marsh, Nathan, and Willingham, (2013) and Adesope et al. (2017) have shown that retrieval practice works irrespective of type of to-be-remembered information and age. Importantly, this advantage was not affected by participants' sex or the time of day of the class (afternoon or evening). Our findings also speak to the controversies regarding the effectiveness of retrieval practice in educational and laboratory settings (see van Gog & Sweller, 2015), which may stem from differences between learning mainly based on semantic or episodic memory, respectively.

The present study raises important points in terms of practical utility for guiding teaching practices. Testing students only once on each taught content with short-answer questions at the end of the same class in which the content was taught was more effective that re-teaching this content and revising it in the next class. Classroom time used for this kind of testing may be offset by enhanced learning, so less time will be spent re-explaining content. Furthermore, frequent testing reduces test anxiety and has even been found to increase students' evaluation of their teachers (Agarwal et al., 2014). Students should also have access to feedback, to correct any possible misunderstandings, but this can be done collectively to reduce teacher burden. Multiple-choice tests in final exams may be used for ease of marking, and have the added benefit of assessing the ability to recall facts in a different way to how they were learned and tested. This constitutes a better means of evaluating what was learned, provides students with an extra opportunity for practicing retrieval and, therefore, benefits learning.

Acknowledgements

We are grateful to Henry L. Roediger III for helpful comments on the draft version. Thanks to Universidade Estadual do Norte do Paraná for allowing us to conduct research on campus and the undergraduates who took part in the study on a voluntary basis.

Funding details

This work was supported by the Fundação Araucária - Governo do Estado do Paraná, CNPq; CAPES through Grant: Finance Code 001 and Process 88881.133607/2016-01; AFIP and Universidade Estadual do Norte do Paraná – UENP/PROPG/EDITORA UENP.

Disclosure statement

None

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research, 87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (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. https://doi.org/10.1007/s10648-012-9210-2

Agarwal, P. K., D'Antonio, L., Roediger, H. L., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, *3*, 131–139. https://doi.org/10.1016/j.jarmac.2014.07.002

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research*, *85*(2), 89–99. https://doi.org/10.1080/00220671.1991.10702818

Bjork, E. L., & Bjork, R. A. (2011). Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, J. R. Pomerantz (Eds.) & FABBS Foundation, *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers. https://doi.org/10.1017/CBO9781107415324.004

Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition*, *3*, 165–170. https://doi.org/10.1016/j.jarmac.2014.03.002

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. *Information Processing and Cognition: The Loyola Symposium*, 123–144.

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *Journal of Experimental Education*, *69*(2), 133–151. https://doi.org/10.1080/00220970109600653

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273–281. https://doi.org/10.1037/1076-898X.13.4.273

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Cohen, J. (1998). *Statistical Power Analysis for the Behavioral Sciences: Second Edition*. NJ, U.S.A.: Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9780203771587

Cosenza, R., & Guerra, L. (2011). *Neurociência e educação*. Artmed Editora.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, Supplement*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Evans, M. D. R., Kelley, P., & Kelley, J. (2017). Identifying the best times for cognitive functioning using new methods: Matching university times to undergraduate chronotypes. *Frontiers in Human Neuroscience*, *11*(April), 1-11. https://doi.org/10.3389/fnhum.2017.00188

Hays, M. J., Kornell, N., & Bjork, R. A. (2012). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 290–296. https://doi.org/10.1037/a0028468

Huang, C. (2013). Gender differences in academic self-efficacy: A meta-analysis. *European Journal of Psychology of Education*, *28*(1), 1–35. https://doi.org/10.1007/s10212-011-0097-y

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D. (2017). Retrieval-Based Learning: A Decade of Progress. In *Learning and Memory: A Comprehensive Reference* (Third Ed.) (pp. 1–28). United Kingdom: Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., & Aue, W. R. (2015). The Testing Effect Is Alive and Well with Complex Materials. *Educational Psychology Review*, *27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). *Retrieval-Based Learning. An Episodic Context Account. Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 61). https://doi.org/10.1016/B978-0-12-800283-4.00007-1

Karpicke, J. D., & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition*, *38*(1), 116–124. https://doi.org/10.3758/MC.38.1.116

Kornell, N., Rabelo, V. C., & Jacobs, P. (2012). Tests enhance learning — Compared to what? *Journal of Applied Research in Memory and Cognition*, *1*, 257–259. https://doi.org/10.1016/j.jarmac.2012.10.002

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *The Pschology of Learning and Motivation*, 1–33. https://doi.org/10.1016/bs.plm.2016.03.003

Leeming, F. C. (2002). The Exam-A-Day Procedure Improves Performance in Psychology Classes. *Teaching of Psychology*, *29*(3), 210–212. https://doi.org/10.1207/S15328023TOP2903_06

Levandovski, R., Sasso, E., & Hidalgo, M. P. (2013). Chronotype: a review of the advances, limits and applicability of the main instruments used in the literature to assess human phenotype. *Trends in Psychiatry and Psychotherapy*, *35*(1), 3–11. https://doi.org/10.1590/S2237-60892013000100002

Lyle, K. B., & Crawford, N. A. (2011). Retrieving Essential Material at the End of Lectures Improves Performance on Statistics Exams. *Teaching of Psychology*, *38*(2), 94–97. https://doi.org/10.1177/0098628311401587

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399-414. https://doi.org/10.1037/a0021782

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*(2), 200–206. https://doi.org/10.3758/BF03194052

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in Middle-School Science: Successful Transfer Performance on Classroom Exams. *Applied Cognitive Psychology*, *27*(3), 360–372. https://doi.org/10.1002/acp.2914

McKenzie, S., & Eichenbaum, H. (2011). Consolidation and Reconsolidation: Two Lives of Memories? *Neuron*, *71*(2), 224–233. https://doi.org/10.1016/j.neuron.2011.06.037

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval Practice in Classroom Settings: A Review of Applied Research. *Frontiers in Education*, *4*(February). https://doi.org/10.3389/feduc.2019.00005

Old, S. R., & Naveh-Benjamin, M. (2008). Differential Effects of Age on Item and Associative Measures of Memory: A Meta-Analysis. *Psychology and Aging*, *23*(1), 104–118. https://doi.org/10.1037/0882-7974.23.1.104

Preckel, F., Lipnevich, A. A., Boehme, K., Brandner, L., Georgi, K., Könen, T., … Roberts, R. D. (2013). Morningness-eveningness and educational outcomes: The lark has an advantage over the owl at high school. *British Journal of Educational Psychology*, *83*(1), 114–134. https://doi.org/10.1111/j.2044-8279.2011.02059.x

Pyc, M. A., Agarwal, P. K., & Roediger, H. L. (2014). Test-enhanced Learning. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying Science of Learning in Education: Infusing Psychological Science into the Curriculum* (pp. 78–90). American Psychological Association Society for the Teaching of Psychology.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*, *17*, 133–140. https://doi.org/10.1016/j.cobeha.2017.08.006

Rahafar, A., Maghsudloo, M., Farhangnia, S., Vollmer, C., & Randler, C. (2016). The role of chronotype, gender, test anxiety, and conscientiousness in academic achievement of high school students. *Chronobiology International*, *33*(1), 1–9. https://doi.org/10.3109/07420528.2015.1107084

Rawson, K. A. (2015). The Status of the Testing Effect for Complex Materials: Still a Winner. *Educational Psychology Review*, *27*(2), 327-331. https://doi.org/10.1007/s10648-015-9308-4

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. https://doi.org/10.1037/a0026252

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., Finn, B., & Weinstein, Y. (2012). Applications of cognitive science to education. In S. Della Sala & M. Anderson (Eds.), *Neuroscience in Education: the good, the bad and the ugly* (pp. 128–151). Oxford: Oxford University Press.

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

Roediger, H. L., & Karpicke, J. D. (2011). Intricacies of spaced retrieval: A resolution. In A. S. Benjamin (Ed.), *Successful Remembering and Successful Forgetting: Essays in Honor of Robert A. Bjork* (pp. 23–48). New York: Routledge.

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten Benefits of Testing and Their Applications to Educational Practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (Vol. 55, pp. 1–36). Oxford: Elsevier. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Rowland, C. A. (2014). The Effect of Testing Versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect. *Psychological Bulletin*, *140*(6), 1–32. https://doi.org/10.1037/a0037559

Santrock, J. W. (2009). *Educational psychology*. McGraw-Hill Education.

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(6), 1712–1725. https://doi.org/10.1037/a0033569

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1392–1399. https://doi.org/10.1037/a0013082

van Gog, T., & Sweller, J. (2015). Not New, but Nearly Forgotten: the Testing Effect Decreases or even Disappears as the Complexity of Learning Materials Increases. *Educational Psychology Review*, *27*(2), 247–264. https://doi.org/10.1007/s10648-015-9310-x

Voyer, D., & Voyer, S. D. (2014). Psychological Bulletin Gender Differences in Scholastic Achievement: A Gender Differences in Scholastic Achievement: A Meta-Analysis. *Psychological Bulletin*, *140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Appendix

Table 1A Content taught in each class assessed by Retrieval Practice or Re-Teaching (lecturer-led review), per block.

Class	Theme
Block 1	
1	Introduction to Educational Psychology
2	Planning, Teaching and Technology
3	Classroom management
4	Classroom assessment
5	Human development I (children)
6	Human development II (adolescents)
7	Piaget and cognitive development
8	Vygotsky and cognitive development
9	Constructivism and relationship between affect and cognition in learning
10	Behaviourism and Learning I
11	Behaviourism and Learning II
12	Contributions of Behavioural Analysis to Education
Block 2	
1	Motivation, teaching, and learning
2	Educational Psychology and Special Education: Dealing with diversity in the educational context
3	Brain, behaviour, and cognition
4	Plasticity, emotions, and learning
5	Perception
6	Attention
7	Intelligence and creativity
8	Memory
9	Executive Functions
10	Thought and language
11	Neuromyths and education
12	Neuroscience and education