



Kent Academic Repository

Ramgopal, Varun, Fastrich, Greta M., Higham, Philip A., Modirrousta-Galian, Ariana, Pattni, Mansi, Potts, Rosalind and Seabrooke, Tina (2026) *Examining the benefits of spaced retrieval practice on transfer and the student experience in the classroom*. *Memory* . pp. 1-21. ISSN 0965-8211.

Downloaded from

<https://kar.kent.ac.uk/116768/> The University of Kent's Academic Repository KAR

The version of record is available from

<https://doi.org/10.1080/09658211.2026.2733327>

This document version

Publisher pdf

DOI for this version

Licence for this version

CC BY (Attribution)

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in **Title of Journal** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).



Examining the benefits of spaced retrieval practice on transfer and the student experience in the classroom

Varun Ramgopal, Greta M. Fastrich, Philip A. Higham, Ariana Modirrousta-Galian, Mansi Pattni, Rosalind Potts & Tina Seabrooke

To cite this article: Varun Ramgopal, Greta M. Fastrich, Philip A. Higham, Ariana Modirrousta-Galian, Mansi Pattni, Rosalind Potts & Tina Seabrooke (28 Sep 2026): Examining the benefits of spaced retrieval practice on transfer and the student experience in the classroom, *Memory*, DOI: [10.1080/09658211.2026.2733327](https://doi.org/10.1080/09658211.2026.2733327)

To link to this article: <https://doi.org/10.1080/09658211.2026.2733327>



© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 28 Sep 2026.



Submit your article to this journal [↗](#)



Article views: 186



View related articles [↗](#)



View Crossmark data [↗](#)

Examining the benefits of spaced retrieval practice on transfer and the student experience in the classroom

Varun Ramgopal ^a, Greta M. Fastrich ^a, Philip A. Higham ^a, Ariana Modirrousta-Galian ^{a,b},
Mansi Pattni ^a, Rosalind Potts ^b and Tina Seabrooke ^a

^aCentre for Perception and Cognition, School of Psychology, University of Southampton, Southampton, UK; ^bDivision of Psychology and Language Sciences, University College London, London, UK

ABSTRACT

Spaced retrieval practice is a learning strategy whereby information is retrieved multiple times with each retrieval attempt separated by a time gap. We conducted a semester-long classroom intervention to investigate the benefits of spaced retrieval practice on memory and transfer. Each week, students engaged in three, spaced learning practice sessions (LPSs) during which they either answered multiple-choice (MC) questions about that week's course content (retrieval practice) or were shown the MC questions with the correct answer selected (restudy). Three weeks after the last LPS, participants completed a final MC test comprising (1) repeated questions from the LPSs, (2) transfer questions that had the same answer options as corresponding LPS questions and queried related topics, but had different correct answers, or (3) new questions about lecture content. A typical retrieval practice effect was observed for repeated questions (retrieval > restudy). Although we did not observe a transfer effect on overall final test accuracy (related = new), further exploratory analyses that used purer measures of transfer revealed evidence of both positive and negative transfer. Furthermore, retrieval practice (vs. restudying) contributed to an enhanced student experience (greater sense of mastery, motivation, and attentional control, and reduced anxiety).

ARTICLE HISTORY

Received 5 February 2026
Accepted 30 August 2026

KEYWORDS

Spaced learning; retrieval practice; transfer; testing effect; spaced retrieval practice

A substantial body of research has investigated the effectiveness of learning strategies in promoting long-term memory retention (Biwer et al., 2020; Bjork, 1994; see Dunlosky et al., 2013 for a review). This research has shown that spaced learning and retrieval practice are two highly potent learning techniques. The *spacing effect* refers to the finding that repeated information is typically remembered better when it is studied over multiple spaced intervals than when it is studied in a single session, even though total study time is equated (e.g., Ebbinghaus, 1885, 1964; Sobel et al., 2011; see Kim & Webb, 2022 for a meta-analysis). For example, in Experiment 3 of his study, Dempster (1987) demonstrated the spacing effect in a vocabulary learning paradigm. Participants studied 38 unfamiliar English words with their definitions (e.g., “loggia – balcony”) three times separated by either 0 (massed condition) or 37 (spaced condition) intervening items. Following a short distractor task, participants completed a final test in which they were presented with all the vocabulary words they encountered at study and were instructed to recall their definitions. Items studied in the spaced condition were correctly recalled

more often ($M = 22.17$) than those studied in the massed condition ($M = 13.42$). The spacing effect is robust and has been experimentally demonstrated with many final test formats, including cued recall (Jacoby, 1978), free recall (Toppino & Bloom, 2002), and recognition (Russo & Mammarella, 2002).

Retrieval practice, on the other hand, refers to the act of attempting to bring previously studied information to mind from memory. Information that is learned with retrieval practice is typically recalled better than information that has been restudied – the *retrieval practice effect* (e.g., Carrier & Pashler, 1992; Roediger & Karpicke, 2006; see Adesope et al., 2017 for a meta-analysis).¹ For example, in a classic study, Roediger and Karpicke showed that completing free recall tests after studying prose passages led to better performance on a delayed memory test than restudying the passages an equal number of times. These memorial benefits were observed in this study even though participants received no corrective feedback on their recall performance after each test. The benefits of retrieval practice have also been seen in real classroom settings. For example, Agarwal and

CONTACT Varun Ramgopal  v.ramgopal@soton.ac.uk  Centre for Perception and Cognition, School of Psychology, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

colleagues have demonstrated the beneficial effects of retrieval practice using quizzes on memory performance in multiple classroom-based studies involving middle school students (see Agarwal et al., 2012, for a useful review of their studies; see Yang et al., 2021, for a meta-analytic review on the retrieval practice effect in classroom settings). Similar to the spacing effect, the retrieval practice effect, both with and without feedback, has also been widely demonstrated in cued recall (Allen et al., 1969; Pyc & Rawson, 2010), free recall (Glover, 1989, Experiments 4a, 4b and 4c), and recognition memory tests (Jacoby et al., 2010).

Spaced retrieval practice

While hundreds of experiments have demonstrated the individual benefits of spaced learning and retrieval practice for long-term memory (see Carpenter & Pan, 2024 and McDermott, 2021 for reviews on spaced learning and retrieval practice, respectively), a smaller body of evidence has shown that combining them is a particularly effective learning strategy (Carpenter & DeLosh, 2005; Glover, 1989, Experiment 3; Higham et al., 2023; Thios & D'Agostino, 1976; see Latimier et al., 2021 for a meta-analysis). For example, Carpenter and DeLosh demonstrated the effectiveness of spaced retrieval practice in a face-name learning paradigm. Participants first studied each face-name pair and were then provided with three repetitions of restudy and retrieval practice items at massed or spaced intervals. Retrieval practice at spaced intervals led to the best performance on a final cued-recall test – better than either restudying at spaced intervals or retrieval practice at massed intervals. This study, therefore, illustrates the benefits of integrating spaced learning and retrieval practice.

In addition to demonstrating the utility of spaced retrieval practice in the laboratory, a small number of studies have tested the efficacy of the technique in classroom settings (e.g., Hopkins et al., 2016; Lyle et al., 2020, 2022). For example, in Experiment 2b of their study, McDaniel et al. (2011) found that end-of-semester recall performance was significantly higher on quizzed items compared to non-quizzed items among K-12 students (eighth-grade science students), regardless of whether the quizzing was administered immediately after a lecture or as a delayed review quiz. However, the retrieval practice effect was more pronounced in the delayed quiz condition compared to the immediate quiz condition. Therefore, the findings from this study demonstrate the beneficial effects that a single, spaced review quiz can have on a semester exam taking place three months later.

The benefits of spaced retrieval practice have also been demonstrated among undergraduate students. For example, Lyle et al. (2022) administered problems pertaining to a specific learning objective in a calculus course. The problems were either all administered at once within the same quiz (massed condition) or were separated and

administered across three separate quizzes spread weeks apart (spaced condition). Performance in the practice quizzes was superior in the massed condition relative to the spaced condition. However, spaced practice led to better performance than massed practice on a final criterial test that was conducted in the last week of the course. These findings highlight the beneficial effects of spacing retrieval practice for long-term memory in both K-12 and college students, even when such spaced retrieval does not benefit retention during initial practice.

Apart from studies demonstrating the benefits of spaced retrieval practice, several other studies that have found that a specific form of spaced retrieval practice called *successive relearning* is a particularly effective learning schedule (e.g., Bahrck, 1979; Bahrck et al., 1993; Rawson et al., 2018; see Rawson & Dunlosky, 2022 for a review). Successive relearning requires participants to achieve *mastery* during spaced learning sessions, which involves correctly recalling all answers to a specified *criterion level* during learning. The criterion level can be one correct recall or more. Usually, to assist participants in achieving mastery, a *dropout methodology* is used. That is, items that reach criterion are eliminated from further testing during a given session, whereas incorrectly answered questions are presented again (later in the session) until they are recalled to criterion. These additional features encourage learners to focus their efforts on information they are having trouble remembering, which is thought to increase learning efficiency.

Higham et al. (2022) and transfer of learning

In this section, we address the issue of *transfer*, which is the focus of the present research. Does spaced retrieval practice, whether in the specialised form of successive relearning or not, support the transfer of learning to new problems or settings, or does it merely promote rote memorisation? Higham et al. (2022) addressed the question of transfer in an 11-week classroom study that used successive relearning. After each weekly (Monday) lecture, students in an Introductory Psychology course practiced relearning the lecture content during three learning practice sessions (LPSs) spaced apart by 2–3 days. In each LPS, students either practiced retrieving answers to 20 fill-in-the-blank questions (e.g., “_____ wrote *Principles of Psychology*”; correct answer: *James*) about that week's lecture or they restudied 20 comparable statements (e.g., *James wrote Principles of Psychology*). This process continued throughout the 11 weeks of the semester, with students answering/restudying a new set of questions/statements to accommodate new weekly lecture content. Finally, participants completed two, different, cumulative, criterial fill-in-the-blank memory tests at the end of the semester. These tests were administered, on average, 42 and 68 days after the material was first taught.

Higham et al. (2022) distinguished between three different question types in the criterial memory tests:

repeated, transfer, and new questions. Repeated questions were identical to those presented in the LPSs. Transfer questions were used to examine whether the benefits of successive relearning generalised to a different part of a question that was presented earlier in the LPSs. For example, if the question that appeared during the LPSs was “Attitudes are composed of _____, cognitive, and behavioral components” (correct answer: *affective*), the corresponding transfer question that appeared on one of the final tests was “Attitudes are composed of affective, _____, and behavioral components” (correct answer: *cognitive*). Finally, new questions were entirely novel questions that had not been presented in the LPSs. For repeated questions, Higham et al. observed a successive relearning advantage: information that participants practiced retrieving over spaced LPSs was recalled better than information that had been restudied over spaced LPSs. Critically, for items that had been practiced with spaced retrieval, transfer was also observed: transfer questions were answered much better than new questions on the criterial tests, although not as well as repeated questions.²

Although Higham et al. (2022) observed a transfer effect, some may argue that generalisation from one part of a question to a different part of the same question is a weak, overly-near form of transfer that has few educational implications. Did participants just learn the statements (or questions with the correct answers) by rote during the LPSs, or was there some deeper learning? The type of transfer that occurred in Higham et al.’s (2022) study could potentially be quite superficial, something that could be exacerbated by repetition. That is, the fact that statements/questions were presented to students as many as three times in each LPS, with three LPSs per week, might have encouraged rote memorisation (e.g., answering the questions without retrieving any broader information related to the topic being queried).

The issue of transfer is an important one in the context of learning. If spaced retrieval practice (with or without the added procedural features associated with successive relearning) merely encourages rote learning, its value is limited. For this reason, we again investigated transfer in a spaced retrieval paradigm in the current study, but with a modified procedure we refer to as the *related-questions paradigm*. Instead of asking participants to provide a different element of the same sentence, here we presented a different question altogether: one that followed the same grammatical structure as the original but whose content was different and where the correct answer was one of the lures for the original question. The related-questions paradigm is potentially more conducive to uncovering deeper forms of transfer than the paradigm used by Higham et al. (2022) in ways that we explain next. Moreover, it also has the potential to reveal *negative* forms of transfer.

Transfer in the related-questions paradigm

The related-questions paradigm was developed by Little, Bjork, and colleagues (Little et al., 2012, 2019; Little & Bjork, 2015) to investigate whether retrieval practice with multiple-choice (MC) questions could produce positive transfer on a later cued-recall test. Specifically, they were interested in whether retrieval processes evoked to answer MC practice questions might later benefit retrieval performance with related (but untested) final cued-recall test questions, which would constitute a form of positive transfer.

To understand why positive transfer might occur in this paradigm, it is important to understand the structure of the questions. The final test of this paradigm includes repeated, related, and new questions. Repeated questions are those that were previously presented and have the same correct answer, whereas new questions are entirely novel and had not undergone retrieval practice. Critically, a *related* final-test question is one that did not undergo initial retrieval practice but queried a similar topic to another question that did. In particular, the correct answer for these related final-test questions was a *lure* (i.e., an incorrect option) on the earlier MC question. To give a concrete example, consider a student who initially answered the practice question “Who was the founder of Structuralism? (a) **Wundt**, (b) James, (c) Skinner, (d) Freud” (boldface indicates the correct answer). To answer this question, the student might engage in retrieval processes to reject one or more of the lures. For example, they may recall that “James” cannot be the correct answer because he is the founder of Functionalism, a school of thought developed in the United States at approximately the same time that Structuralism was developed in Germany. Later, when presented with the cued recall question “Who is the founder of Functionalism?” on the final test, the student’s recall accuracy will be facilitated by the retrieval processes evoked to reject the lure “James” on the earlier MC practice question. In other words, the retrieval processes that allowed “James” to be rejected on the MC practice test should facilitate later cued recall performance when “James” is the target.

Across multiple studies, Little, Bjork and colleagues demonstrated positive transfer in the related-questions paradigm. That is, related cued recall questions on a final test were consistently associated with higher recall accuracy than new, previously untested questions presented for the first time (e.g., Little et al., 2012, 2019; Little & Bjork, 2015). Moreover, several findings have converged on the retrieval interpretation just described. For example, only MC questions with competitive lures (i.e., difficult questions) tend to produce positive transfer (Little et al., 2012; Little & Bjork, 2015). If the lures are not competitive (i.e., easy questions), there is no need to engage in much processing of the lures because there is only one plausible (correct) answer. Similarly, test formats that encourage lure processing tend to produce

more positive transfer than test formats that do not. For example, confidence-weighted tests, where examinees are required to assign relative confidence to the MC options (Sparck et al., 2016), and elimination testing, where examinees are required to provide reasons for rejecting lures (e.g., Alamri & Higham, 2022, Experiment 3; Little et al., 2019), both show substantial positive transfer.

Although these positive transfer effects are noteworthy, other research has indicated that *negative* transfer (related < new) can also occur in the related-questions paradigm. One factor determining whether positive or negative transfer will be observed is the final test format. For example, Alamri and Higham (2022, Experiment 3) found that if the final test was cued recall, positive transfer was observed (mean final test accuracy: related = .49; new = .37). However, if the final test format was MC, negative transfer was observed instead (related = .36; new = .49). This negative transfer was obtained even though Alamri and Higham used elimination testing on the practice test, which (as noted above) typically facilitates positive transfer. Alamri and Higham reasoned that when the final test is in MC format, the presence of matching answer alternatives on the practice and final tests for related questions creates a strong sense of familiarity for related questions on the final test. This misleading familiarity (which is not present to the same degree with cued recall final test questions because of the missing alternatives) creates the false belief that the questions are the same. This belief, in turn, seduces participants into choosing the option that was the corrective feedback during practice (which is now incorrect) perhaps without even properly processing the question stem.

This research shows that final test format is important in determining whether positive or negative transfer will be observed when MC practice tests are used. However, sometimes even MC final tests are associated with positive transfer if the circumstances are right. One moderating factor is the context in which the test is taken. For example, Bjork et al. (2014) observed positive transfer in an educational setting, even though an MC format was used for both practice and final testing. In the same vein, Alamri and Higham (2025) found negative transfer using an MC final test if educational (SAT) materials were used in a non-educational environment, but positive transfer if the context shifted to a genuine university classroom. The likely explanation for this pattern of results is that in an authentic educational context, participants are motivated to more deeply process the question stems and alternatives. Deeper processing, in turn, should reduce false recognition of related questions, and clear the way for earlier retrieval practice to produce positive transfer. Thus, the picture that is emerging is that positive transfer will occur by answering some types of MC questions at practice (those with competitive lures and/or a cued recall final test). However, these positive transfer effects can be overridden by misleading feelings of familiarity

for the corrective feedback if the final test is MC and the testing environment is non-educational.

Overview of the study

Given the equivocal results obtained with the related-questions paradigm when both the practice and final tests are in MC format, we implemented a classroom intervention similar to that of Higham et al. (2022) to investigate the matter further. While the procedure shared many features with that study, we employed the related-questions paradigm because it has the potential to provide evidence for both a deeper form of positive transfer and negative transfer. The primary purpose of the intervention was to determine whether knowledge acquired during learning practice completed throughout the semester would transfer to a final MC test, as in Higham et al., while assessing transfer using the related-questions paradigm.

We were particularly interested in whether, if transfer occurred, it would be positive or negative. Based on the literature discussed above, there are reasons to expect both forms of transfer. On the one hand, the fact that the study occurred in a genuine educational context (as in Higham et al., 2022) would suggest that positive transfer would occur because the authentic educational context should encourage students to fully process the questions and answer options. In turn, students should be less susceptible to confusing related questions with repeated on the final test (e.g., Alamri & Higham, 2025; Bjork et al., 2014). The current study also required students to answer the same questions repeatedly during the LPSs. To our knowledge, transfer in the related-questions paradigm has not been explored with spaced repetitions. However, spaced repetitions provide multiple opportunities to fully process the lures, at spaced intervals, potentially increasing positive transfer. On the other hand, these spaced repetitions may also increase the familiarity of the corrective feedback, potentially increasing negative transfer on a later test when the feedback is no longer correct. As a result, there are several procedural differences between this study and previous experiments using related questions to make the nature of transfer uncertain (e.g., delayed final test, although see Bjork et al., 2014).

Following Higham et al. (2022), we included both retrieval and restudy trials during the LPSs (see below). Importantly, the task that participants performed during the LPSs (retrieval or restudy) could potentially moderate these transfer effects. For example, familiarity might increment with each repetition equally for restudied and retrieved items, thereby producing equal amounts of negative transfer for the two conditions. Conversely, lure processing may be more complete under retrieval (vs. restudy) conditions because there may be retrieval of information specifically used to reject the lures, something that is not required in the restudy condition. If that is the

case, positive transfer might be greater in the retrieval condition than the restudy condition.

Figure 1 provides an overview of the protocol. The procedure for the intervention was similar to that used in Higham et al. (2022) in that students engaged in three spaced LPSs following each weekly lecture in an 11-week Introductory Psychology course. Specifically, students attended each week's lecture on the Monday, and then completed an LPS on Wednesday, Friday, and the following Monday. These days were selected to broadly match Higham et al.'s protocol, which followed the same timings except students in their study completed the final LPS on Sunday rather than Monday. We placed the final LPS on the Monday of the following week to avoid forcing students to study on the weekend, when many have other commitments. Therefore, this spacing meant that the LPSs were spaced 2–3 days apart. This spacing interval between each LPS was primarily based on what was achievable within the module's constraints and not on a theoretically optimal spacing interval. However, it is worth noting that three repetitions with a spacing gap of 2–3 days is considered an efficient successive relearning/spaced retrieval practice schedule that produces good retention while limiting diminishing returns (e.g., Rawson & Dunlosky, 2011).

Unlike Higham et al. (2022), the procedural features specific to successive relearning (e.g., mastery, dropout) were not used here. Although the LPSs consisted of retrieval and restudy trials in both studies, the tasks performed in these conditions differed somewhat. For retrieval, students in the current study practiced learning eight facts/concepts by ranking the correctness of four options presented with a question stem, after which immediate corrective feedback was provided.³ For restudy, students practiced learning eight additional facts/concepts by studying a question stem alongside four options, with the correct answer highlighted. Regardless of condition, each item was practiced across two passes per LPS. This change in the protocol had the advantage of allowing us

to carefully match students' exposure to the facts/concepts across the two conditions.

After completing the second pass of each LPS, participants rated their subjective experience. Specifically, participants completed four Visual Analogue Scales (VASs; Koo & Yang, 2025) on anxiety, motivation, mastery, and attentional control as well as answering a question about their learning strategy. We obtained these measures so that we could examine the effects of spaced retrieval practice and restudy on aspects of the student learner experience, beyond final test memory performance.

Prior research has found that retrieval practice can improve both memory and students' subjective learning experiences. For example, practice quizzes have been shown to increase motivation to continue studying (Abel & Bäuml, 2020), while reducing mind-wandering (Szpunar et al., 2013) and test anxiety (for a meta-analytic review, see Yang et al., 2023). In the context of successive relearning, Higham et al. (2022) found that students' self-reported experiences differed across the LPSs. Specifically, in the first LPS, students reported greater anxiety, less mastery, and similar attentional control after retrieval practice than after restudying. By the third LPS, however, the pattern had reversed, with students reporting less anxiety and greater mastery and attentional control after retrieval practice than restudying – a pattern the authors attributed to improvements in recall over the LPSs. In the current work, we aimed to test whether these findings would replicate in the context of spaced retrieval practice. If so, it would suggest that the learner-experience benefits observed by Higham et al. are not dependent on participants achieving mastery through successive relearning.

At the end of the course, students completed a final 200-item MC criterial test comprising five question types: (1) repeated-restudy: 40 MC questions identical to those that had been restudied during the LPSs; (2) repeated-retrieval: 40 MC questions identical to those that had been practiced via retrieval during the LPSs; (3) related-

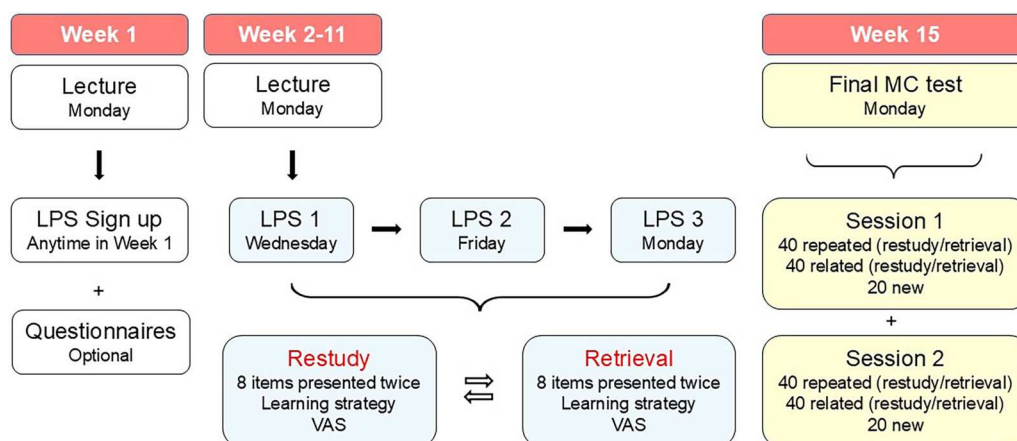


Figure 1. A schematic representation of the study protocol.

Note: LPS: learning practice session; VAS: visual analogue scale. A diagram of a flowchart.

restudy: 40 MC questions that shared the same structure and answer options as questions that had been restudied during the LPSs but a different correct answer; (4) related-retrieval: 40 MC questions that shared the same structure as questions that had been practiced via retrieval during the LPSs but a different correct answer; and (5) new: 40 new MC questions assessing lecture content that had not been practiced during the LPSs.

Research questions

We preregistered one primary research question (RQ1) and five secondary RQs (RQ2–6). RQs 5 and 6 are presented in the Supplemental Materials as they are not directly related to the investigation of transfer and retrieval practice that this paper mainly focuses on. RQ1–4 are paraphrased from the preregistration as follows.

RQ1: For the criterial test, will related questions (that were either restudied and retrieved during the LPSs) be answered better or worse than new, control questions? We deliberately did not predict a direction for the hypothesis associated with this RQ because both facilitated and impaired accuracy on related questions has been observed in prior research when the practice and final tests are MC, as in the current study (e.g., see Alamri & Higham, 2025, who obtained both results).

RQ2: Is there an effect of practice type (restudy vs. retrieval) on final-test accuracy and, if so, is it moderated by question type (repeated vs. related)? We predicted a retrieval practice effect such that participants would correctly answer more questions that they practiced retrieving during the LPSs than questions they restudied. We also predicted the effect of practice type to be qualified by an interaction with the question type such that we would observe a larger retrieval practice effect for repeated questions than related questions.

RQ3: Is there an effect of question type (repeated vs. related) on final-test accuracy? We predicted participants would correctly answer more questions in the repeated condition than the related condition. We predicted this result because participants had prior exposure to the questions and corrective feedback in the repeated condition but not in the related condition.

RQ4: Is there an effect of practice type (restudy vs. retrieval) on learners' subjective weekly experiences? We predicted a main effect of practice type (retrieval > restudy) for mastery, motivation and attentional control, and the opposite pattern for anxiety (restudy > retrieval). Although not preregistered, we predicted a main effect of LPS ($3 > 2 > 1$) for mastery, motivation and attentional control, and the opposite pattern for anxiety ($1 > 2 > 3$). Finally, we predicted the effect of practice type to be qualified by an interaction with the LPS. This interaction prediction was in line with Higham et al.'s (2022) study where participants' mastery and attentional control scores were lower (higher for anxiety) in the retrieval condition compared with the restudy condition on the first LPS, but this pattern reversed by the time participants were on the third LPS.

Method

Transparency and openness

The research questions, method, and analyses were preregistered. The preregistration (<https://osf.io/k5j64/>

overview), data, analytic code, and the Supplemental Materials (https://osf.io/t79zx/overview?view_only=4b776d48b3da4c70a5fe7c44909f4363) for this study are publicly archived on the Open Science Framework. We report all deviations from the preregistration, the rationale for the sample size used in the study, and all data exclusions, manipulations, and methods. Data were analysed using R, version 4.5.0 (R Core Team, 2025). This study received ethical approval from the Psychology Ethics Committee at the University of Southampton (ID: 45089.A3).

Participants

The participants were 230 first-year undergraduate students who were enrolled in an "Introduction to Psychology" module in 2023–2024. As preregistered, the sample size was determined by the number of students who enrolled on the course and elected to participate in the study.

As part of the course assessment, students were required to engage in 30 LPSs, which were spread evenly throughout the semester (three LPSs per week for 10 weeks), and two final tests (not the main examination for the module). As explained in more detail later, both the LPSs and the final tests contained MC questions about the course material. If students completed an LPS and scored at least 25% correct on the retrieval trials, they received 0.5% toward their overall module mark (15% in total if all LPSs were completed). No students lost credit due to low LPS accuracy. The students received an additional 5% for completing the final tests (with no minimum score applied) for a possible total of 20% towards their final module mark.

Additionally, students were able to sign up for a voluntary research component associated with the module in exchange for nine research credits. This component entailed students completing five pre-learning questionnaires in Week 1 before starting the LPSs from Week 2, and four VASs in each LPS to indicate how they felt during learning.

Participant and data exclusions

We preregistered that we would exclude participants who did not consent to having their LPS and final test data used for research purposes. From the 323 students who originally signed up for the LPSs, two students were excluded on this basis. Twelve students did not reach the end of a final test session. These participants were therefore also excluded. Additionally, given the longitudinal nature of this study, several participants provided incomplete data. We therefore introduced a non-preregistered exclusion criterion to exclude participants who either did not complete any LPSs or the final tests. Nineteen participants completed the LPSs but not the final tests, and seven participants completed the final test but no LPSs. These students were excluded.

The remaining 283 participants provided data from 6,374 LPSs. In 136 cases, participants restarted an LPS and thus might have experienced some trials multiple times. Data from these LPSs were excluded. Out of the remaining LPSs, data were excluded if participants indicated they had used their lecture notes while completing the LPS (179 LPSs) or if participants had been shown a question either less or more than twice due to technical difficulties (227 LPSs). Additionally, to control for missing LPSs, we decided to only include data from weeks where a participant completed all three LPSs on that topic. If a participant, for example, completed the first and last LPS of the third week and all three LPSs of the fourth, only data relating to the fourth week were included in the analyses. This preregistered criterion excluded data from a further 1941 LPSs.

After performing the above-mentioned exclusions, 244 participants remained. Fourteen further participants were excluded under the preregistered criterion of excluding participants who experienced significant technical issues. Specifically, 10 participants had two subject IDs in the initial weeks of practice, meaning they were assigned to two different item counterbalancing groups. These participants were excluded. Additionally, one participant completed an LPS twice due to a technical error, and three participants experienced a technical glitch during the final test (see Footnote 4 under Design and Materials). These participants were also excluded.

In total, data from 3,693 LPSs were included. Our final sample comprised 230 participants (female = 203; male = 23; non-binary = 3; others = 1) and 1,231 weeks of data; these participants provided complete LPS (completed all three LPSs in the weeks included) and final test data and had ages ranging between 17 and 43 years ($M = 18.92$, $SD = 2.52$). Of the 230 participants, five participants who did not consent to the learning strategy/VAS measures administered during the LPSs were excluded from the analyses pertaining to RQ4.

Design and materials

There were 11 weeks of teaching for the course, with a single 90-min lecture delivered on Monday of each week. Following the lecture in Week 1, students signed up for the intervention and were introduced to the online system that would deliver the LPSs in Week 1. Students did not complete formal LPSs in the first week. Following each Monday lecture in Weeks 2–11, students were invited to take part in three LPSs over the remainder of the week, one each on Wednesday, Friday, and the following Monday. Questions from the lecture content in Week 1 were included in the LPSs in Week 2 along with the content covered in Week 2. Each LPS consisted of 16 MC question stems with four options related to that week's lecture topic. Eight questions were answered as MC questions (retrieval items), whereas the other eight appeared as MC questions with the correct answer

highlighted in green, which students restudied (restudy items).

Items of the same practice type (retrieval vs. restudy) were presented in a blocked design. Specifically, all eight items of the same practice type were presented in a randomised order once before they were freshly randomised and presented again for a second pass. The order of retrieval versus restudy blocks was freshly randomised for each participant for each LPS. Overall, this meant that each LPS consisted of a total of 32 practice trials (two passes over eight blocked restudy items and two passes over eight blocked retrieval items).

In addition to practice type (restudy vs. retrieval), we also manipulated question type with three levels: repeated, related, and new. To achieve this manipulation, we created a set of 20 related MC question pairs on the content of each week's lecture (10 weeks $\times$ 20 question pairs = 200 question pairs in total). Each question from a given pair was on the same topic, had similar question stems, and the same set of four answer options. One option for each question was the correct answer and the remaining three options were lures. Critically, the two questions in each pair never had the same correct answer. In other words, the correct answer for one question was a lure for the other question. For example, the question "Who founded Structuralism?" with the answer options **Wundt**, James, Freud, and Skinner (boldface indicates the correct answer), was paired with the question "Who founded Functionalism?" with the same answer options (correct answer: **James**).

To eliminate item effects, 10 item counterbalancing groups were created for each LPS, each consisting of a set of four restudy items to be repeated at final test (restudy-repeated condition), four retrieval items to be repeated at final test (retrieval-repeated condition), four restudy items whose related pairs were presented at final test (restudy-related condition), four retrieval items whose related pairs were presented at final test (retrieval-related condition), and four new items that were on the final test, but neither they nor their related pairs were presented during the LPSs (new condition).⁴ Item counterbalancing groups were created such that each of the forty items (from 20 pairs) was assigned equally often to each of the five within-subject test conditions (restudy-repeated, retrieval-repeated, restudy-related, retrieval-related, and new) across participants. Each participant was assigned to an item counterbalancing group at random before the start of the LPSs. Participants in the same group practiced the same items in the same way in each LPS for a given week.

Four weeks after the last lecture (three weeks after the last LPS), students were invited to take part in a final MC test that was framed as a mock exam. For each of the 10 LPS weeks (Weeks 2–11), 20 questions were included on the test (four each for the five experimental conditions described earlier), for a total of two hundred final test items. Questions were included on the test regardless of

whether the student participated in the corresponding LPSs. (Given the longitudinal nature of the study, some students did not complete every session; see earlier Participant and Data Exclusions section). Thus, all participants answered all 200 final test items. Counterbalancing ensured that half of the participants answered the first question from each question pair (with 40 questions each serving in each of the five experimental conditions), whereas the other half of participants answered the other half (also with 40 questions per experimental condition).

Students were permitted to take a short break during testing. The break was achieved by randomly dividing the 200 questions into two sets of 100, with the constraint that each set of 100 questions represented all five experimental conditions (i.e., 20 items each in the restudy-repeated, retrieval-repeated, restudy-related, retrieval-related, and new). Furthermore, for each half of the test, two items in each set of 20 questions per experimental condition represented each of the 10 weeks of lecture content. This meant that participants answered an equal number of items for each topic in each final test session. The order of presentation of the questions within each final test session was randomised individually for each participant. Note that while during the LPSs, participants did not experience questions from four of the item pairs (so that these questions could be used as “new” questions on the final test), one question from each of these item pairs was presented on the final test. However, each item pair was assigned to a different final test condition based on a participant’s item counterbalancing group.

Pre-learning questionnaires

For students who volunteered for research credit, the following questionnaires were administered in the first week of the module, before the LPSs began in Week 2: General Worry (Kelly, 2004), Attentional Control (Derryberry & Reed, 2002), Student Self-Efficacy (Rowbotham & Schmitz, 2013), Test Anxiety (Putwain et al., 2021), and Approach to Learning (Biggs et al., 2001). More detail about each scale and the analyses pertaining to the pre-learning questionnaires are included in the Supplemental Materials (see Tables S2 and S3).

LPS measures (learning strategy and VAS)

After each set of eight retrieval or restudy items had been presented twice within an LPS, participants judged the extent to which they had considered all response options on the previous set of trials using a 7-point Likert scale (“To help my learning: I only considered one answer vs. I considered all options”). After the strategy question, participants responded to a visual analogue scale (VAS) to measure participants’ learning experiences across four constructs, including mastery, intrinsic motivation, anxiety, and attentional control. Participants provided ratings on a scale from 0 (“not at all”) to 100 (“Extremely”) to a word (e.g., “ANXIOUS”, “WORRIED”,

“CONFIDENT”) or phrase (e.g., “I WILL MASTER THIS TASK”) by moving a slider to indicate how they felt at the moment. Each scale comprised three items, with “I WILL MASTER THIS TASK”, “MOTIVATED” and “CONFIDENT” creating the mastery scale; “HAPPY”, “SATISFIED” and “INTERESTED” creating the intrinsic motivation scale; “ANXIOUS”, “WORRIED”, “RELAXED” creating the anxiety scale; and “FOCUSED”, “CONCENTRATED” and “DISTRACTED” creating the scale for attentional control.⁵ All constructs showed good internal consistency at all measurement time points (Cronbach’s α s > .70). Higher scores indicated an increased sense of mastery, motivation, attentional control, and anxiety.

Procedure

The study was hosted on JATOS (Lange et al., 2015). The procedure consisted of an initial sign-up session, thirty LPSs (three LPSs for each of 10 weekly topics), and one final test. There was one study link created for each LPS. The section involving the strategy question and VAS measures was automatically skipped for participants who had not provided consent to these measures.

Initial sign-up session

After students attended the Week 1 introductory lecture to the course, they received an invitation to sign up for the LPSs that accompanied the course. During this session, they were also asked to give consent for their data to be used in a research project and opt into additional questionnaires that would be completed during the initial session and alongside the LPSs for additional research credits. Participants then provided their name, email address, current age in years, gender identity, and native language. If participants opted to take part in the additional questionnaires, they then completed the pre-learning questionnaires (see Design and Materials section). All students then completed a practice session where they were introduced to the main trials and received four restudy and four retrieval trials as practice for the main LPSs.

LPSs

On Wednesday, Friday, and the following Monday after each Monday lecture, students were invited to complete an LPS online. LPSs were accessible between 0900 and 2100 on practice days. About 30 min before each session went live, students were sent a personalised link to the session. Every session started with the instructions and asked participants to state whether they had attended the Monday lecture. Participants then completed a restudy block and a retrieval block (in randomised order). Each block consisted of two passes over the eight restudy or retrieval trials. If participants had opted in to obtaining research credits during sign up, they were then presented with a question querying their learning strategy during the just-completed block of trials (either

restudy or retrieval trials). Specifically, they were asked to rate the degree to which they had considered just a single answer option or all the answer options to help their learning ("To help my learning: I only considered one answer – I considered all the options"). This rating was made on a 7-point scale ranging from 0 ("I considered only one answer") to 6 ("I considered all the options"). After the strategy rating, we measured students' subjective learning experience across four VASs. Thus, students completed the strategy rating and the four VASs twice per LPS, once after the two passes through the eight retrieval trials and once after the two passes through the eight restudy trials (see Figure 1).

During the LPSs, each trial started with a 500 ms blank screen. Then, participants were presented with a prompt in the form of a question or a statement where a phrase was replaced with a blank line (e.g., "Light is measured in _____"; correct answer: *nanometres*). Under the question, four answer options (*Nanometres*, *Decibels*, *Hertz*, and *Ampere*) were presented in a random order. On retrieval trials, participants were asked to rank each candidate answer according to the likelihood of it being correct with "1" (top rank) representing the most likely answer and "4" (bottom rank) the least. At the beginning of the trial, a question mark was displayed next to each answer option. By clicking on it, a drop-down menu opened, and participants were able to select the question mark or a ranking number from one to four that was not yet assigned to another answer option for this question. Students had a total of 22 s to rank the answer options. During this time, they were able to change their selections as often as they wanted. The instructions encouraged them to "rank each of the four options" and not just the most likely. After 22 s elapsed, students received corrective feedback, with green highlighting appearing around the correct answer. This feedback remained on the screen for 3 s before the trial moved on to a blank screen (total trial time = 25 s).

During restudy trials, both the question stems and answer options were presented as in the retrieval condition. However, instead of question marks and the expectation that participants would rank the alternatives themselves, the numbers "1" to "4" appeared next to the answer options with the "1" placed next to the correct answer and values "2", "3", and "4" randomly assigned to the other answer options. Additionally, the area around the correct answer was highlighted in green, similar to how the correct answer was identified in the retrieval condition (except there was no retrieval attempt). After 25 s elapsed, the trial moved on to a blank screen which marked the beginning of the next trial (total trial time = 25 s).

After completing both passes through the retrieval and restudy blocks (including questionnaires if participants consented to them), participants were given two final questions. First, they were asked whether they used their notes during the LPS. Second, they were given the

chance to leave any other helpful comments, for example, if they encountered technical issues.

Final test

The final test took place on the Monday of Week 15, four weeks after the last lecture and three weeks after the last LPS. Again, participants were given a 12-hour window to participate. The final test consisted of two sessions which could be completed in separate sittings as long as both were completed within the 12-hour window and the first session was completed before the second began. Both sessions had the same structure: participants were introduced to the question format and informed that they had 35 s to answer each question, but they could move on before that time elapsed by clicking a button. Participants were also advised not to refer to their lecture notes during the sessions and instead rely on their memory. Additionally, they were encouraged to answer every question, even if they were unsure about the answer, as there would be no penalty for answering incorrectly. Lastly, participants were told that there would not be a break within a given session, but a break was permitted between the sessions.

Each trial began with a 500 ms presentation of a blank screen. Afterwards, the question stem was presented on the screen alongside the four answer options (in randomised order). Unlike the retrieval practice trials during the LPSs, the format of the final test trials followed a standard MC format. That is, participants simply selected one of the candidate answers instead of ranking them. Underneath the answers, a "continue" button was displayed which participants could use to move to the next trial. If participants did not progress within 35 s, the trial moved on automatically. Unlike the LPSs, only items in their question format (used for the retrieval condition during the LPSs) were included on the final test; no statements (used for the restudy condition during the LPSs) were included.

Results

We first focus on reporting our preregistered analyses pertaining to final test performance and then report some preregistered and non-preregistered exploratory analyses. Alongside frequentist analyses, we report Bayes Factors (BF_{10}) to evaluate the evidence for both the null and alternative hypotheses. We interpreted the Bayes Factors according to the evidence categories proposed by Jeffreys (1961), which were subsequently interpreted by Lee and Wagenmakers (2013; see Table 1).

Final test performance

Our preregistered research questions, associated analyses, and a summary of findings are shown in Table 2.

Transfer (RQ1 – preregistered)

To test RQ1, we conducted a preregistered one-way within-subjects Analysis of Variance (ANOVA) with three

Table 1. Bayes factor evidence categories and corresponding interpretations.

BF_{10}	Interpretation
> 100	Extreme evidence for H1
30–100	Very strong evidence for H1
10–30	Strong evidence for H1
3–10	Moderate evidence for H1
1–3	Anecdotal evidence for H1
1	No evidence
0.33–1	Anecdotal evidence for H0
0.10–0.33	Moderate evidence for H0
0.03–0.10	Strong evidence for H0
0.01–0.03	Very strong evidence for H0
<0.01	Extreme evidence for H0

Note: H1 refers to the alternative hypothesis, whereas H0 refers to the null hypothesis.

Table 2. Preregistered research questions (RQ1–4), associated analyses and summary of findings.

RQ	Analysis	Finding
RQ1: Will restudy-related and retrieval-related questions be answered better or worse than new questions?	RQ1: One-way ANOVA on final test accuracy with question type as IV	No effect of question type
RQ2: Will retrieved items be answered better or worse than restudied items? Will this be qualified by an interaction with question type (repeated/related)? RQ3: Will repeated questions be answered better than related questions?	RQ2 & RQ3: A 2 (practice type: restudy/retrieval) $\times$ 2 (question type: repeated/related) ANOVA on final-test accuracy	RQ2 & RQ3: Main effects of practice type (retrieval > restudy) and question type (repeated > related), and an interaction: There was a retrieval practice effect (retrieval > restudy) on repeated questions but not on related questions
RQ4: Will there be an effect of practice type on VAS (mastery, motivation, anxiety and attention) scores?	RQ4: Four 2 (practice type: restudy/retrieval) $\times$ 3 (LPS: 1/2/3) ANOVAs on VAS scores	RQ4: Main effects of practice type and LPS were obtained on most measures, which were qualified by interactions (larger difference between the retrieval and restudy conditions on later LPSs) for all measures other than attention

Note: RQ: research question; ANOVA: analysis of variance; IV: independent variable.

conditions (restudy-related, retrieval-related, and new items) with final-test accuracy as the dependent variable. We did not find a significant overall effect of condition, $F(2, 458) = 0.00$, $p = .99$, $\eta^2_G = 0.00$, $BF_{10} = 0.02$. Accuracy was, therefore, comparable amongst restudy-related items ($M = 62.32\%$, $SD = 18.14\%$), retrieval-related items ($M = 62.42\%$, $SD = 16.64\%$) and new items ($M = 62.32\%$, $SD = 17.45\%$).

Effects of question and practice type (RQ2 and RQ3 – preregistered)

Mean recall performance by question type and practice type is shown in Figure 2. To test RQ2 and RQ3, we

conducted a preregistered 2 (question type: repeated/related) $\times$ 2 (practice type: restudy/retrieval) within-subjects ANOVA with final-test accuracy as the dependent variable. We found a significant main effect of question type, $F(1, 229) = 219.21$, $p < .001$, $\eta^2_G = 0.13$, $BF_{10} = 5.85 \times 10^{35}$. Accuracy was significantly higher for repeated items ($M = 75.13\%$, $SD = 13.05\%$) than related items ($M = 62.37\%$, $SD = 14.15\%$). We also found a significant main effect of practice type, $F(1, 229) = 30.35$, $p < .001$, $\eta^2_G = 0.02$, $BF_{10} = 1.42 \times 10^4$. Accuracy was significantly higher for retrieved items ($M = 71.31\%$, $SD = 12.30\%$) than restudied items ($M = 66.20\%$, $SD = 15.22\%$).

There was also a significant interaction between question type and practice type, $F(1, 229) = 35.73$, $p < .001$, $\eta^2_G = 0.02$, $BF_{10} = 5.65 \times 10^5$. For repeated items, a paired-samples t -test showed that accuracy was significantly higher in the retrieval condition ($M = 80.15\%$, $SD = 13.49\%$) than the restudy condition ($M = 70.07\%$, $SD = 17.57\%$), $t(229) = 8.85$, $p < .001$, $d_z = 0.58$, $BF_{10} = 2.19 \times 10^{13}$. For related items, by contrast, accuracy was comparable between the retrieval condition ($M = 62.42\%$, $SD = 16.64\%$) and restudy condition ($M = 62.32\%$, $SD = 18.14\%$), $t(229) = 0.33$, $p = .94$, $d_z = 0.01$, $BF_{10} = 0.074$. Thus, there was a retrieval practice effect in the repeated condition, but not in the related condition.

VAS scores (RQ4 – preregistered)

Mean VAS scores are shown in Figure 3. To test RQ4, we conducted four separate preregistered 2 (practice type: restudy/retrieval) $\times$ 3 (LPS: 1/2/3) repeated measures ANOVAs for each of the VAS measures (Mastery, Motivation, Anxiety, and Attention) as the dependent variables (collapsed across weeks).⁶ We found a significant main effect of practice type on all four VASs. Mean ratings were higher in the retrieval condition than the restudy condition for mastery, $F(1, 224) = 63.43$, $p < .001$, $\eta^2_G = 0.01$, $BF_{10} = 3.06 \times 10^{17}$, motivation, $F(1, 224) = 55.55$, $p < .001$, $\eta^2_G = 0.01$, $BF_{10} = 5.85 \times 10^{13}$, and attention $F(1, 224) = 88.95$, $p < .001$, $\eta^2_G = 0.03$, $BF_{10} = 1.99 \times 10^{32}$, and lower in the retrieval condition for anxiety, $F(1, 224) = 4.64$, $p = .03$, $\eta^2_G = 0.00$, $BF_{10} = 0.47$.

We also found a significant main effect of LPS on mastery, $F(2, 448) = 24.79$, $p < .001$, $\eta^2_G = 0.01$, $BF_{10} = 1.94 \times 10^9$, anxiety, $F(2, 448) = 47.90$, $p < .001$, $\eta^2_G = 0.02$, $BF_{10} = 4.89 \times 10^{28}$, and the effect of LPS on motivation approached significance, $F(2, 448) = 2.78$, $p = .06$, $\eta^2_G = 0.00$, $BF_{10} = 0.27$. We did not find a significant main effect of LPS on attention, $F(2, 448) = 1.26$, $p = .29$, $\eta^2_G = 0.00$, $BF_{10} = 0.02$.

Finally, we found significant practice type $\times$ LPS interactions for mastery, $F(2, 448) = 4.08$, $p = .02$, $\eta^2_G = 0.00$, $BF_{10} = 0.13$, motivation, $F(2, 448) = 7.07$, $p = .001$, $\eta^2_G = 0.00$, $BF_{10} = 0.42$, and anxiety, $F(2, 448) = 7.46$, $p = .001$, $\eta^2_G = 0.00$, $BF_{10} = 0.22$. We did not find a significant interaction on attention, $F(2, 448) = 0.11$, $p = .89$, $\eta^2_G = 0.00$,

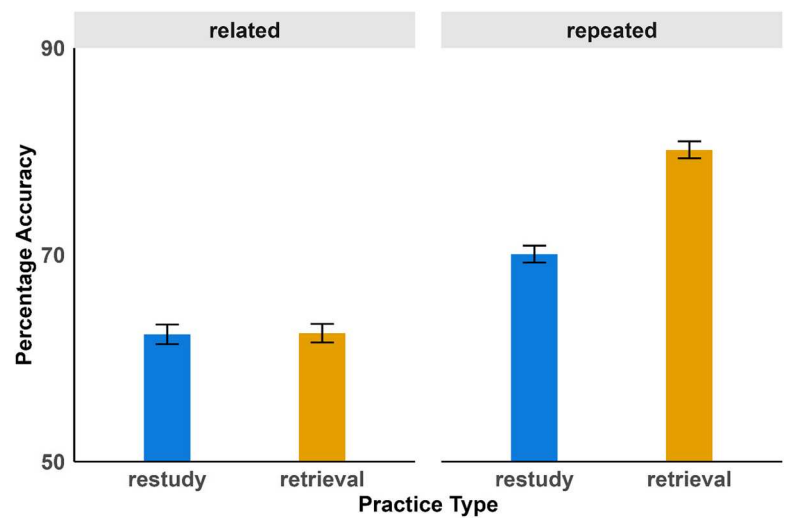


Figure 2. Mean accuracy on the multiple-choice final test by question type and practice type.
Notes: Error bars reflect standard error of the mean. The data were collapsed over the two parts of the final test. Error bars in this paper reflect within-subjects adjusted (Morey, 2008) standard errors of the mean. However, error bars reflecting week-level standard error estimates in Figures 4 and 7 must be interpreted considering that we included data only from the weeks in which a participant completed all three LPSs, as noted in the Participant and Data Exclusions section.

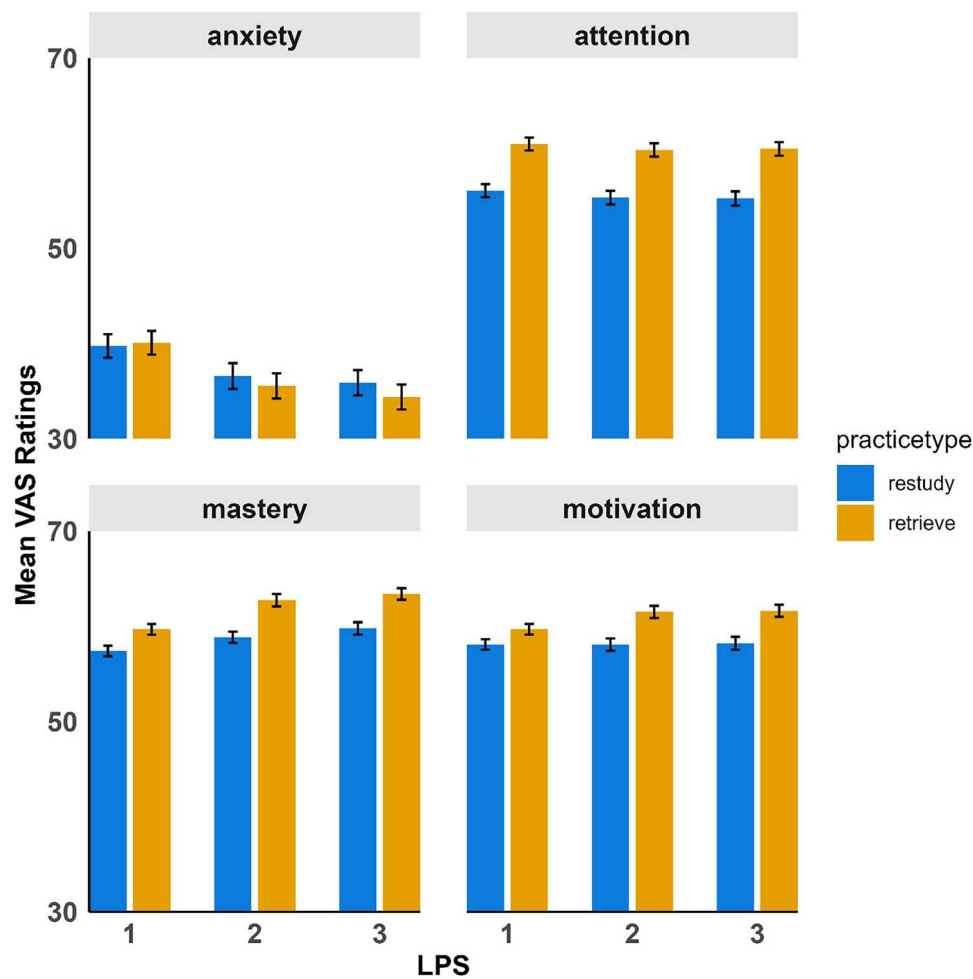


Figure 3. Mean visual analogue scale (VAS) scores by practice type and LPS.
Note: VAS: visual analogue scale. Error bars reflect standard error of the mean. LPSs are collapsed across weeks 2–11.

Table 3. Results from the paired-samples *t*-tests on visual analogue scale (VAS) responses.

Simple retrieval vs. restudy comparisons	<i>df</i>	<i>t</i>	<i>p</i>	<i>d_z</i>	<i>BF</i> ₁₀
<i>Mastery</i>					
LPS 1 (retrieval > restudy)	224	4.64	<.001	0.31	1941.00
LPS 2 (retrieval > restudy)	224	7.87	<.001	0.52	4.05×10^{10}
LPS 3 (retrieval > restudy)	224	5.81	<.001	0.39	4.15×10^5
<i>Motivation</i>					
LPS 1 (retrieval > restudy)	224	3.54	<.001	0.24	30.86
LPS 2 (retrieval > restudy)	224	7.55	<.001	0.50	5.84×10^9
LPS 3 (retrieval > restudy)	224	5.92	<.001	0.39	7.22×10^5
<i>Anxiety</i>					
LPS 1 (retrieval = restudy)	224	0.73	.46	0.05	0.10
LPS 2 (restudy > retrieval)	224	2.57	.01	0.17	1.85
LPS 3 (restudy > retrieval)	224	3.27	.001	0.22	13.10

Note: LPS: learning practice session.

$BF_{10} = 0.02$. Note, however, that the Bayes Factors supported the null for each interaction.

Paired *t*-tests were used to follow up the interactions for mastery, motivation, and anxiety and are shown in Table 3. For mastery, ratings tended to be higher in the retrieval condition than the restudy condition on all LPSs, but the difference between the two conditions was larger in the second and third LPSs (particularly the second) compared with the first LPS, thereby resulting in the interaction. The same pattern was observed for motivation, with mean ratings tending to be higher in the retrieval condition than the restudy condition, and the difference between the two conditions being higher on the later LPSs. Finally, anxiety scores tended to reduce over the LPSs in both the retrieval and restudy conditions. However, anxiety ratings were higher in the restudy condition than the retrieval condition on the later LPSs, although there was no difference between the two conditions on the first LPS, thereby producing the interaction.

Exploratory analyses

We also conducted several exploratory analyses, some of which were preregistered as exploratory and some of which were not preregistered. The research questions, associated analyses, and a summary of findings are shown in Table 4.

Learning practice performance (EQ1 – preregistered)

We mentioned in the preregistration that we would potentially examine LPS performance in our exploratory analyses. Hence, to address EQ1, we computed recall accuracy during the LPSs using data from the first pass through the questions within each LPS in the retrieval condition. Recall performance as a function of LPS and week is shown in Figure 4. We collapsed the data over weeks and conducted a one-way, within-subjects ANOVA with LPS (1/2/3) as the independent variable and recall accuracy as the dependent variable. We found a main effect, $F(2, 458) = 940.88$, $p < .001$, $\eta^2_G = 0.57$, $BF_{10} = 8.70 \times 10^{162}$. Paired-samples *t*-tests revealed that accuracy was lower on the first LPS ($M = 57.46\%$, $SD = 14.63\%$) compared to the second LPS ($M = 86.94\%$, $SD = 12.37\%$), $t(229) = 33.21$, $p < .001$, $d_z = 2.19$, $BF_{10} = 6.07 \times 10^{85}$, and the third LPS ($M = 91.22\%$, $SD = 12.27\%$), $t(229) = 33.19$, $p < .001$, $d_z = 2.19$, $BF_{10} = 5.3 \times 10^{85}$. Accuracy was also significantly lower on the second LPS than the third LPS, $t(229) = 7.43$, $p < .001$, $d_z = 0.49$, $BF_{10} = 3.16 \times 10^9$. Thus, recall performance improved with spaced retrieval practice.

Exploring transfer effects further (non-preregistered)

Following our preregistered analyses involving the comparison between related and new items with respect to final test performance, we explored the data further to potentially uncover transfer effects. The following

Table 4. Exploratory research questions, associated analyses, and summary of findings.

Research question	Analysis	Finding
EQ1 (preregistered): Will memory performance on LPSs 2 and 3 be higher than in LPS1 in the retrieval condition, as measured on the first pass of each LPS?	A one-way ANOVA on practice recall accuracy with the three LPSs (1/2/3) as IV	Main effect of LPS ($3 > 2 > 1$)
EQ2 (non-preregistered): Will participants be more likely to select the feedback lure in the related versus new condition (negative transfer) when the correct answer trials are excluded?	A one-way ANOVA on selection rate of feedback lure with the three conditions (retrieval-related/restudy-related and new) as IV	Main effect of condition (retrieval-related > restudy-related > new)
EQ3 (non-preregistered): Will participants be more likely to select the correct answer option in the related versus new condition (positive transfer) when the trials in which the feedback lure was chosen are excluded?	A one-way ANOVA on final-test recall accuracy with condition as IV	Main effect of condition (retrieval-related > restudy-related = new)
EQ4a (preregistered): Will there be an association between mean learning strategy in the retrieval condition and the percentage difference in feedback lure selection between the retrieval-related and new conditions?	Correlational analysis	The correlation was negligible
EQ4b (preregistered): Will there be an association between mean learning strategy in the restudy condition and the percentage difference in feedback lure selection between the restudy-related and new conditions?	Correlational analysis	The correlation was negligible
EQ5 (preregistered): Will retention interval (the interval between the last LPS of a week and the final test) be a significant predictor of final-test accuracy?	A multiple linear regression model on mean final-test accuracy with the linear and quadratic terms of retention interval as the predictor variables	Both the linear and quadratic terms of retention interval were significant predictors

Note: EQ: exploratory question; feedback lure: the lure on a related test question that was the corrective feedback during the LPSs. For analyses including new items, the same option that served as the feedback lure in the related conditions was used as the point of comparison.

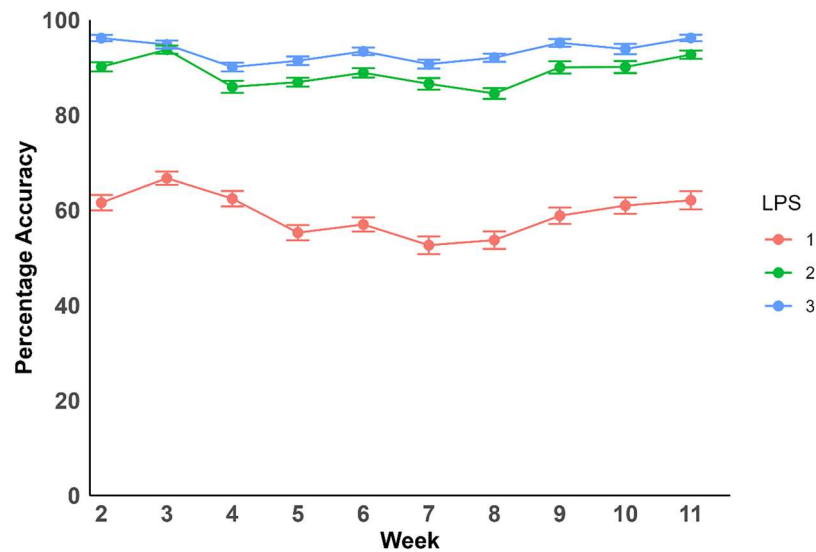


Figure 4. Mean accuracy in the retrieval condition on pass 1 for each week and LPS.

Note: Error bars reflect standard error of the mean.

exploratory analyses on potential transfer effects were not preregistered. Although we observed no overall transfer effect, it is possible that both positive and negative transfer effects occurred, but because the effects opposed each other, they cancelled each other out.

To investigate this possibility, we ran two non-preregistered, exploratory analyses using purer measures of negative and positive transfer effects, respectively. To isolate negative transfer and address EQ2, we estimated how likely students were to select the feedback lure amongst all the incorrect alternatives on error trials. To facilitate the comparison of the related (retrieval and restudy) and new conditions, which was complicated by the fact that there was no feedback lure for new items, we took advantage of the counterbalanced design. That is, each item serves in both the new condition and the two related conditions. This design feature allowed us to compare the probability of choosing the specific option that serves as the feedback lure in both the related conditions (e.g., *James*) with the probability of choosing that same option (*James*) when the question is new. By conditioning this analysis on errors, the effect of positive transfer (leading to correct responses) is removed.

Feedback lure selection (EQ2). We calculated the percentage of feedback lure selections (i.e., previously correct answers in the LPSs) in the final test as follows: for each condition, we divided the number of trials in which participants selected the feedback lure by the total number of trials in which they made an error. The equation was:

$$\text{Percentage of feedback-lure selection} = \frac{\text{Number of feedback-lure errors}}{\text{Total number of errors}} \times 100\%$$

Percentage feedback lure selection given an error response in the related and new conditions is shown in

Figure 5. We conducted a one-way within-subjects ANOVA with condition (restudy-related, retrieval-related, and new items) as the independent variable and lure selection as the dependent variable. We found a significant main effect of condition, $F(2, 432) = 23.02, p < .001, \eta^2_G = 0.07, BF_{10} = 1.53 \times 10^7$. Paired-samples *t*-tests showed that, when participants selected an incorrect option, they were significantly more likely to select the feedback lure in the retrieval-related condition ($M = 62.06\%$, $SD = 23.55\%$) than the restudy-related condition ($M = 55.42\%$, $SD = 24.72\%$), $t(218) = 2.87, p = .004, d_z = 0.19, BF_{10} = 4.14$, and the new condition ($M = 47.82\%$, $SD = 22.89\%$), $t(222) = 7.18, p < .001, d_z = 0.48, BF_{10} = 6.55 \times 10^8$. The percentage of feedback lure selections was also significantly higher in the restudy-related condition compared with the new condition, $t(218) = 3.30, p = .001, d_z = 0.22, BF_{10} = 14.47$.

Correct answer selection without the feedback lure (EQ3). To investigate the potential of positive transfer once the opposing influence of negative transfer was removed and to address RQ3, we examined the probability of choosing the target after eliminating trials in which the feedback lure was chosen. In other words, given that the feedback lure was not chosen, what was the likelihood of making a correct response? If that rate is higher in the (retrieval and restudy) related conditions compared to the new condition, then evidence of positive transfer would be obtained. The equation was:

$$\text{Percentage of correct-answer selections} = \frac{\text{Correct}}{\text{Correct} + \text{Other Wrong Answers}} \times 100\%$$

Percentage correct answer selection excluding the feedback lure in the related and new conditions is shown in **Figure 6.** We conducted a one-way within-subjects ANOVA with condition (restudy-related, retrieval-related,

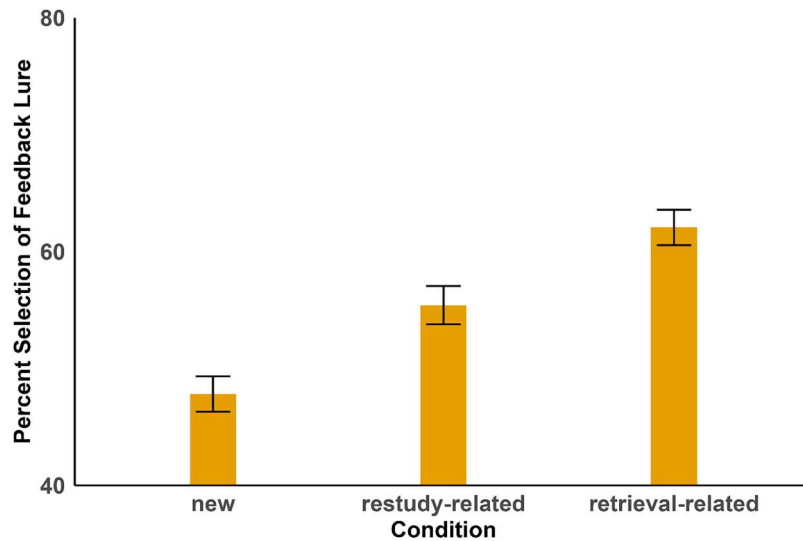


Figure 5. Percent feedback lure selection given an error response on the multiple-choice final test in the related and new conditions.
Note: Error bars reflect standard error of the mean.

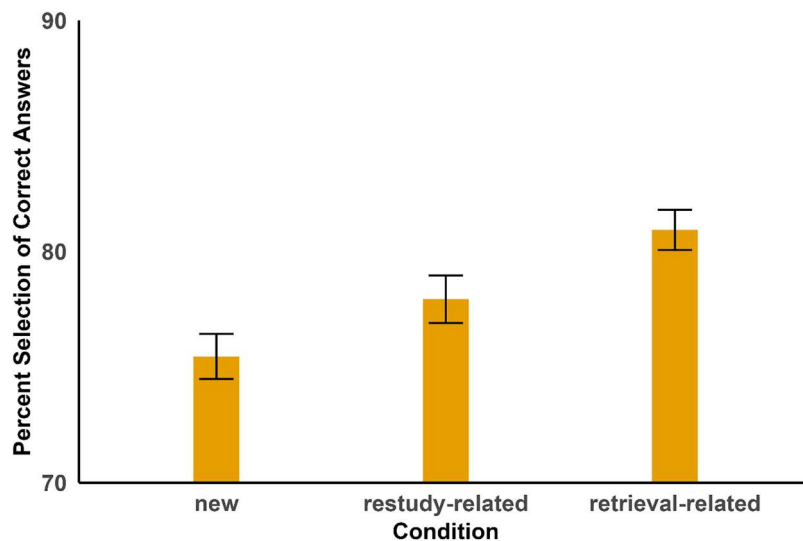


Figure 6. Correct answer selection on the multiple-choice final test in the related and new conditions when excluding the feedback lure.
Note: Error bars reflect standard error of the mean.

and new) as the independent variable and correct answer selection as the dependent variable. We found a significant effect of condition, $F(2, 458) = 8.17$, $p < .001$, $\eta^2_G = 0.02$, $BF_{10} = 40.39$. Paired-samples t -tests showed that, on trials in which students did not select the feedback lure, they were significantly more likely to select the correct answer option in the retrieval-related condition ($M = 80.94\%$, $SD = 14.76\%$) than the restudy-related ($M = 77.94\%$, $SD = 17.82\%$), $t(229) = 2.25$, $p = .02$, $d_z = 0.15$, $BF_{10} = 0.87$ and the new condition ($M = 75.46\%$, $SD = 16.84\%$), $t(229) = 4.38$, $p = .001$, $d_z = 0.29$, $BF_{10} = 659.78$. We did not find a significant difference between the restudy-related and new conditions, $t(229) = 1.68$, $p = .09$, $d_z = 0.11$, $BF_{10} = 0.29$.

Association between learning strategy and transfer in terms of feedback lure selection (EQ4a and EQ4b – preregistered)

We indicated in the preregistration that we would use learning strategy (the degree to which students considered just a single answer option or all the answer options while ranking alternatives on retrieval trials or while restudying the question options during learning practice) in our potential exploratory analyses. To address EQ4a, we examined the association between mean learning strategy in the retrieval condition and the percentage difference in feedback lure selection given an error response between the retrieval-related and new conditions. We conducted this analysis as we predicted that

considering more options while ranking alternatives during learning practice would lead participants to be less seduced by the feedback lure in the retrieval-related condition (a form of positive transfer), thereby choosing the feedback lure less often. Therefore, we expected an inverse correlation between mean learning strategy and the percentage difference between the retrieval-related and new conditions in terms of feedback lure selection. However, we found only a negligible negative correlation between the two variables $r(215) = -.05$, $R^2 = 0.003$.

To address EQ4b, we examined the association between mean learning strategy in the restudy condition and the percentage difference in feedback lure selection given an error response between the restudy-related and new conditions. As participants did not engage in active retrieval of information on study trials during learning practice, we expected that there would not be an association between mean learning strategy and the percentage difference between the restudy-related and new conditions in terms of feedback lure selection. Again, we found only a negligible negative correlation between the two variables $r(212) = -.04$, $R^2 = 0.002$.

Effect of retention interval (EQ5 – preregistered)

We indicated in the preregistration that we would also examine retention interval in our potential exploratory analyses. To address EQ5, we fitted a multiple linear regression model on mean final-test accuracy with the linear and quadratic effects of retention interval as the predictor variables. Final-test accuracy as a function of week is shown in Figure 7. We found that the model was significant, $F(2, 1228) = 40.34$, $p < .001$, $R^2 = 0.06$. Both the linear term, $\beta = 0.01$, $SE = 0.00$, $t(1228) = 3.26$, $p < .001$, and the quadratic term,⁷ $\beta = 0.04$, $SE = 0.00$,

$t(1228) = 7.86$, $p < .001$, were significant predictors of final-test accuracy.

Discussion

The primary objective of this preregistered intervention study, which took place in a real educational environment with a large cohort of students (total $N = 230$), was to examine whether spaced retrieval practice would facilitate positive or negative transfer of knowledge to related questions in a final MC test. We wanted to specifically examine the role of repetition over the course of a semester in a real undergraduate course to discover whether multiple spaced repetitions would facilitate deep learning and positive transfer or increase misleading feelings of familiarity and negative transfer.

We built on Higham et al.'s (2022) study that used successive relearning, a specialised form of spaced retrieval practice, in a semester-long Introduction to Psychology module. Although our study used the same basic procedure, it differed in two important ways. First, we did not adopt the components that were specific to successive relearning (mastery learning and dropout methodology). That is, items retrieved correctly on the first pass within an LPS were not dropped from practice on the second pass. Second, although both studies examined transfer of knowledge from learning practice to a final test, Higham et al. examined near transfer effects by determining if the benefits of successive relearning generalised to a different part of the same question versus new (untested) questions. By contrast, we examined a deeper form of transfer using the related-questions paradigm, whereby a related (previously untested) final-test question queried a similar topic as an earlier MC practice question but, critically, had a different correct answer. We were specifically interested in investigating the nature of transfer when

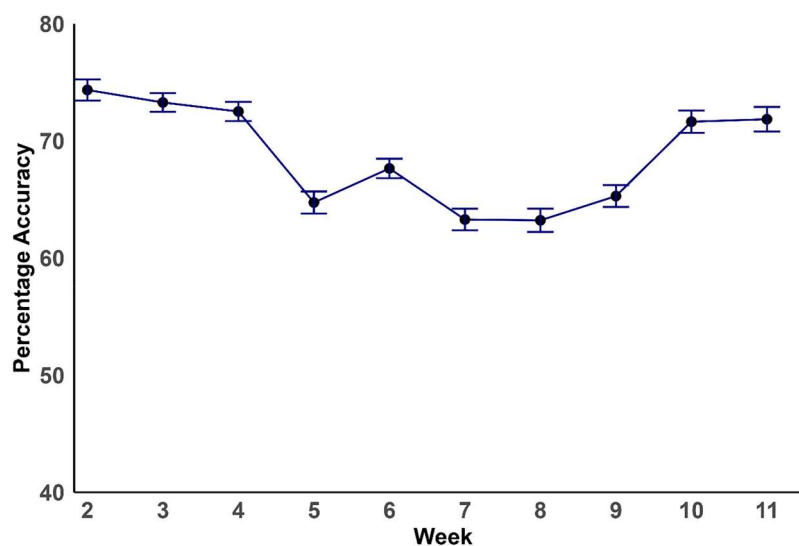


Figure 7. Mean final-test accuracy for items from each week.

Note: Error bars reflect standard error of the mean.

both the practice and final tests were MC, because past studies have yielded mixed results with this design. In our work, we observed a retrieval practice effect for repeated questions, but no overall transfer effect for related questions. From an educational standpoint, the insights from the current study are important in the sense that ours was the first study to integrate spaced retrieval practice and the related-questions paradigm to examine learning transfer.

Transfer effects

Although we observed a retrieval practice effect for repeated questions, no overall transfer effect for related questions was observed. The lack of a retrieval practice effect for related questions (retrieval = restudy), and the lack of overall transfer effects suggest that repeated, spaced retrieval practice (and spaced restudying) may have led participants to simply memorise the correct answer during the LPSs rather than encouraging deep processing of the correct and incorrect alternatives. From this perspective, spaced retrieval practice and spaced restudy – at least as implemented in the current study – appear to have offered limited benefit. Despite extensive student effort and study time over an 11-week course, spaced retrieval practice and restudy did not improve learning beyond the specific material that had been tested or restudied.

Unlike in our study, Higham et al. (2022) observed positive transfer effects (performance on related questions > new questions). One notable difference between these studies is that Higham et al. examined a form of near-transfer, while we used the related-questions paradigm to assess a deeper form of transfer. Specifically, in the final test of Higham et al.'s study, participants were required to recall a different aspect of a fact they had learned through spaced retrieval practice. In the current study, by contrast, the related questions on the final test had the same format and answer options to questions in the LPSs, but critically required a different correct answer. One implication of these findings is that spaced retrieval practice may facilitate relatively superficial form of transfer, potentially supported by rote memorisation, but not deeper forms of transfer that require flexible application of knowledge to novel situations.

Considering that we did not observe overall transfer effects, it is worth revisiting two previous studies that did observe transfer using both the related-questions paradigm and educational environments similar to that used in the current study. First, Alamri and Higham (2025) observed positive transfer effects using materials from a similar Introduction to Psychology module. However, unlike in our study, the authors did not implement spaced retrieval practice throughout the semester, and they tested transfer in a single end-of-the-semester tutorial in which participants completed both learning practice and a final test. Spaced retrieval practice, in our

study, likely increased participants' familiarity with both the questions and answer options. Because the final-test questions closely resembled those presented during retrieval practice and had the same answer options, this increased familiarity may have biased participants toward selecting the previously correct – yet now incorrect – answer.

When Alamri and Higham (2025) tested transfer in the absence of repeated questions on the final test, they observed positive transfer effects on both related questions that were presented immediately after the other question of the pair and related questions that were presented after a delay. However, in the presence of repeated questions, positive transfer was seen only when the related question was presented immediately after the other question of the pair, whereas negative transfer was observed on delayed related questions. Furthermore, participants were more likely to choose the feedback lure on such delayed related questions. The findings from Alamri and Higham highlight an important nuance: the presence (or absence) of repeated questions on the final test plays a critical role in participants' ability to successfully discriminate between two related questions that have different correct answers. In the current study, repeated and related questions were intermixed, which may have reduced the likelihood of observing positive transfer.

Second, Bjork et al. (2014) observed positive transfer in a study in which students completed MC practice quizzes throughout a research methods course, typically 3–4 days after the weekly Monday lecture. An MC final exam was administered at the end of the course. Memory performance on both repeated and related questions was better than new questions on the final exam. Hence, unlike in the current study, Bjork et al. observed positive transfer effects on an MC final exam when students engaged in retrieval practice in the form of practice quizzes across the course.

While students in Bjork et al. (2014) engaged in learning practice of the material from the preceding lecture, as in the current study, there were no spaced repetitions of the items appearing on the practice quizzes. That is, unlike our study, any given practice quiz item in Bjork et al. was presented only once and neither repeated within a session nor across multiple sessions within a week. The obvious question to ask here is why we did not observe transfer despite making students engage in not just retrieval practice but retrieval practice that was spaced both within and across LPSs over the semester. In Bjork et al., practicing each item just once may have led to an optimal balance of familiarity and deeper encoding. This familiarity could have facilitated the recognition of question stems and answer options in the repeated and related conditions on the final test, including the feedback lure. However, the deeper encoding involved with retrieving information to reject lures during the practice test may have helped to keep the misleading familiarity of the feedback lure in check, leading to positive transfer effects. On

the other hand, spaced repetitions within and across the LPSs within each week in the current study may have increased the misleading familiarity of the feedback lures more than it encouraged deeper lure processing, resulting in little positive transfer in our study. If repetition had this type of differential effect on familiarity versus deeper lure processing, the likely result would be enhanced recall of repeated items that were retrieved during practice but no obvious positive transfer effects.

This notion that repetition can have opposing, ironic effects on memory was investigated by Jacoby (1999). He demonstrated both positive and detrimental effects of repetition in an experiment where participants were required to respond “yes” to words that had previously been spoken, and “no” to words that had previously been read. For younger adults, false alarms to earlier-read words decreased with repetition, while for older adults, the opposite effect was found. The beneficial effect of repetition on younger adults’ memory could be attributed to their use of repetitions to consciously recollect the words they read and those they heard. Though younger adults may have experienced enhanced familiarity of repeated written words, they were not seduced by that familiarity because they were able to use source recollection to keep that familiarity in check. On the other hand, the ironic effect of repetitions in older adults could be attributed to a deficit in recollection. That is, the familiarity of information increased as information was repeated just as it did for younger adults. However, because recollection of source was impaired in elderly individuals, that familiarity went unchecked, causing elderly participants to erroneously respond “yes” to repeated written words.

Whenever opposing influences are present in a design, there is a risk that the two influences nullify each other, leading to the conclusion that there was a null result when actually there are two, opposing effects. We investigated this possibility in the current intervention and found evidence for both positive and negative transfer effects. To isolate these opposing transfer effects, we conducted two exploratory analyses (EQs2 and 3). The first analysis, which aimed to isolate the negative transfer effect, was limited to error trials. By eliminating the correct trials, we removed trials where positive transfer may have controlled the response, leaving negative transfer free to express itself. Specifically, if the feedback lure is chosen with greater likelihood when the question is related than when it is new, then negative transfer has occurred. This analysis indicated that students were more likely to choose the feedback lure than the other incorrect alternatives in both the retrieval-related and restudy-related conditions compared with the new condition, showing evidence of negative transfer. This evidence of negative transfer is in line with Alamri and Higham’s (2022, 2025) finding that participants were more likely to endorse the feedback lure on related questions compared to new questions on a final MC test. As

explained, the tendency to select the feedback lure more often on related questions could be attributed to the sense of familiarity that arises when participants see the same question stem and answer options.

The second analysis, which aimed to isolate a positive transfer effect, examined the likelihood of a correct response when the feedback lure was not selected. In other words, trials where the feedback lure was selected were removed, thereby eliminating all trials where negative transfer may have controlled the response. We then examined the likelihood of choosing the target on the remaining trials. We found that students were more likely to choose the correct answer option in the retrieval-related condition compared with the new condition, showing evidence of positive transfer. This evidence of positive transfer, on trials where negative transfer could have controlled the participant’s response, is encouraging in the sense that participants might have been able to better discriminate between the two questions of the pair (the one encountered at practice and the one encountered on the final test) on those trials. Therefore, eliminating the feedback lure option or finding ways to reduce the influence of the feedback lure on the final test could be an important step in aiding positive transfer effects to emerge. It is worth noting that students who write multiple MC tests throughout a course in the form of quizzes, learning practice tests or final tests may encounter the same options on different potentially related questions. For example, names like “Skinner” or terms like “positive reinforcement” likely appear as both correct and incorrect alternatives across questions. In other words, negative transfer might be a real concern rather than an issue specific to the related-questions paradigm. Together these results indicate that both negative and positive transfer effects occurred in this study despite overall test performance suggesting that neither effect occurred. As suggested earlier, one plausible account for this null effect on overall performance is that the two opposing types of transfer effects cancelled each other out.

Given the evidence of both positive and negative transfer, can we recommend spaced retrieval practice in real educational environments? As indicated, removing the feedback lure in the final test may be one way to increase the likelihood of obtaining positive transfer. We predict that, in such a procedure, participants would be better able to discriminate between the repeated and related questions, thereby boosting positive transfer. Another option would be to reduce the number of LPSs within each week to mitigate the chances of misleading familiarity (resulting from multiple spaced repetitions) seducing students into choosing the feedback lure on the final test. Future work could also consider adding a source recollection question before some, if not all, questions on the final test. Doing so should help students to consciously recollect whether a final test question is repeated, related, or new.

Retrieval practice effect on repeated questions

To answer RQ2 and RQ3, we examined the effects of question type (repeated/related) and practice type (restudy/retrieval) on final-test performance. As predicted, test accuracy was higher on items that students practiced retrieving compared to the items they restudied during learning practice (RQ2). We also found that test accuracy was higher on items answered in the repeated condition compared with those answered in the related condition (RQ3). These main effects were qualified by a significant interaction, which was due to there being a retrieval practice effect (retrieval > restudy) on repeated items, but not on related items (retrieval = restudy). The interaction indicates that, although related questions queried a similar topic to an earlier MC practice question, that similarity was not enough for related items to benefit from retrieval practice. We may observe a retrieval practice effect on related items if the questions encountered during the LPSs were made more difficult, as more difficult questions are thought to promote transfer by encouraging lure elaboration (Little et al., 2012; Little & Bjork, 2015). Additionally, future studies could provide students with guides comprising examples of the type of related questions they can expect to encounter on the final test to boost the chances of observing a testing effect on related items as well.

There are different, competing theories that could potentially explain the testing effect on repeated questions. For example, the episodic context account (Hong et al., 2019; Karpicke et al., 2014) suggests that the way in which retrieval practice boosts retention is by reinstating a prior learning context and updating the context representation in memory with the current recall context. This updated context representation, comprising both a previous study context and the current recall context, provides more retrieval cues at the time of later recall compared with restudying. In the current study, engaging in repeated retrieval practice across three LPSs within each week may have performed a similar function. On the first LPS of a week, participants may have been reminded of the context in which they studied for the first LPS. On subsequent LPSs, they may have been reminded of retrieval practice on the previous LPS(s) and the corrective feedback associated with each question. Hence, the initial study context along with the updated context representation from the multiple episodes of retrieval practice within each week would have provided participants with more retrieval cues in the retrieval-repeated condition compared with the restudy-repeated on the final test.

Benefits to subjective learner experience

As part of RQ4, we found a statistically significant interaction between practice type and LPS on mastery, motivation and anxiety. (Retrieval practice led to increased attentional control compared with restudying, with a

consistent difference between the two conditions on all LPSs). However, the corresponding Bayes Factors supported the null hypothesis, indicating that the evidence for these interactions was weak and should be interpreted cautiously. As such, the patterns described below are best viewed as tentative.

Repeated retrieval practice over the week led to a stronger sense of mastery, motivation and reduced anxious affect on later LPSs compared with restudying. Higham et al. (2022) found that mean mastery ratings were lower (higher for anxiety) in the retrieval condition compared with the restudy condition on the first LPS, but that this pattern reversed by the third LPS. In the current study, we observed that mean mastery ratings were higher in the retrieval condition than the restudy condition on all the LPSs, but the difference between the two conditions was larger on the later LPSs. As for anxiety, although we did not find a significant difference between the retrieval and restudy conditions on the first LPS, mean ratings were higher in the restudy condition than the retrieval condition on the second and third LPSs. These results, therefore, suggest that spaced retrieval practice may improve subjective learner experience, despite having no overall effect on transfer, although we make this suggestion cautiously given the Bayesian analyses. As expected, students' recall accuracy on retrieval trials improved as the LPSs progressed over a given week. This improvement may underpin their reports of improved learner experience (Higham et al., 2022).

The benefits of retrieval practice for subjective perceptions of mastery, motivation and attentional control in the current study are encouraging in the sense that effortful learning strategies are often perceived negatively by students. Students frequently judge fluent learning strategies, such as restudying and massed learning, to be more effective than more effortful approaches such as retrieval practice. Nonetheless, desirably difficult strategies typically confer greater longer-term memory benefits than more fluent strategies (e.g., Bjork, 1994; Deslauriers et al., 2019; Kornell & Bjork, 2008). As such, these findings indicate that retrieval practice may still provide meaningful benefits by positively influencing the learner experience.

Effect of repeated retrieval practice on learning practice performance

As part of EQ1, we found a significant main effect of LPS on practice response accuracy in the retrieval condition. Practice accuracy was the highest on the third LPS and lowest on the first LPS. This finding illustrates the benefits of engaging in repeated practice on recall performance.

Association between learning strategy and feedback lure selection

As part of EQ4a, we found only a negligible negative correlation between learning strategy in the retrieval

condition and the percentage difference in feedback lure selection given an error response between the retrieval-related and new conditions. Although the correlation was in the direction we predicted, a weak correlation could imply that participants' learning strategy while ranking alternatives during the LPSs did not affect their likelihood of selecting the feedback lure on retrieval-related items compared with new items on the final test.

As part of EQ4b, we again found a negligible negative correlation between learning strategy in the restudy condition and the percentage difference in feedback lure selection given an error response between the restudy-related and new conditions. The weak correlation between the two variables could be attributed to the fact that students were presented with the full statement with the correct answer highlighted beneath the statement on study trials. Hence, as participants did not engage in active lure processing on these trials during the LPSs, the number of alternatives they may have considered on these trials may not have affected their likelihood of feedback lure selection on restudy-related items compared with new items.

Effect of retention interval on final-test accuracy

A multiple linear regression model revealed that both the linear and quadratic effects of retention interval were significant predictors of final-test accuracy. Both the linear and quadratic estimates were positive suggesting that longer retention intervals were associated with higher accuracy (linear effect), with items from the initial and later weeks associated with higher accuracy compared with the middle weeks (quadratic effect). We expected accuracy from the later weeks to be higher, considering the shorter retention intervals between the LPSs and the final test for those weeks. Final-test accuracy was higher on items from the first three weeks of learning practice compared with items from the middle weeks. We may have obtained this pattern of result because participants either found items from the initial weeks to be easier or they may have spent longer revising these items before the final test, considering the longer interval between the practice of these items and the final test.

Limitations

As with any longitudinal study, we had limited control over attrition and how students revised for the LPSs after each lecture, or their level of engagement during the LPSs, especially on restudy trials when they had to simply read the statements. Additionally, we did not have control over how students prepared for the final test after the last LPS. Furthermore, participants could use their notes during the LPSs and final tests (although they were encouraged not to). That said, final test performance was not at ceiling, suggesting that they did not consistently use their notes.

Conclusions

Overall, we found evidence of both positive and negative transfer effects in the related-questions paradigm. These effects may sometimes cancel each other out, as suggested by the current study. There have been studies demonstrating the beneficial effects of successive relearning (Higham et al., 2022) and retrieval learning practice (Bjork et al., 2014) on the transfer of knowledge on a final test. However, the findings and insights from the current intervention are particularly important as ours was the first study to integrate the investigation of transfer in the related-questions paradigm with spaced retrieval practice in an educational environment. Future work might focus on developing ways to boost transfer in an educational environment in a spaced retrieval paradigm by ensuring deeper learning during practice and by helping students discriminate better between repeated and related questions on the final test.

Notes

1. The retrieval practice effect is also known as the *testing effect* or *backward testing effect*.
2. Since no part of the questions was removed during the LPSs in the restudy condition (i.e., the statements were presented intact), there was no logical difference between the related and repeated conditions for restudied items. The distinction between related and repeated items only existed for items in the retrieval condition.
3. Ranking was used in the retrieval condition because Alamri and Higham (2022, 2025) found that, unlike standard MC format, ranking encourages lure processing and transfer. Although other test formats such as elimination testing also encourage lure processing (e.g., Little et al., 2019), participant compliance can be a problem (e.g., many participants find it difficult to provide reasons for rejecting the lures on every question).
4. All participants should have ideally experienced 20 items from each week's lecture content on the final test. However, there was a technical glitch between the server and browser during the final test that resulted in the total exclusion of three participants, as noted in the Participant and Data Exclusions section. Of the 1,231 weeks of data that were included for analyses, there were 135 weeks (involving 94 participants) from which the participant experienced slightly less than 20 items owing to this technical glitch. Of those 135 weeks, there were 128 weeks (involving 91 participants) from which the participant experienced 19 items and seven weeks (involving seven participants) from which the participant experienced 18 items. The final test data pertaining to these weeks have not been excluded for those participants.
5. The item "RELAXED" within the anxiety scale, and the item "Distracted" within the attention control scale were reverse-scored (i.e., 100 – participant's rating).
6. We made a minor error when preregistering our predictions for this analysis. Specifically, we predicted a main effect of practice type (retrieve > restudy) for each of the four mean learner subjective weekly experiences. However, we only expected scores for mastery, motivation, and attention to be higher in the retrieve condition than the restudy condition; we expected the opposite pattern for anxiety. Furthermore, we did not preregister any predictions regarding the main effect of LPS or the practice type × LPS interaction. However,

it is reasonable to expect that scores would increase (decrease for anxiety) across the LPs, as recall improved, and that this pattern would be larger in the retrieval condition than the restudy condition, given Higham et al.'s (2022) results.

7. A significant quadratic effect with a positive β , in this case, suggests that the relationship between retention interval and final-test accuracy was not just linear, but was U-shaped. Accuracy was higher on items from the initial and later weeks, and reduced for items from the middle weeks, thereby showing the pattern of a serial position curve.

Acknowledgements

We would like to thank Julie A. Hadwin for her contribution to the development of methodology.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Our work was funded by the UKRI Economic and Social Research Council [grant number: ES/T013664/1].

Data availability statement

The research questions, method, and analyses were preregistered. The preregistration, data, analytic code, and the Supplemental Materials for this study are publicly archived on the Open Science Framework.

ORCID

Varun Ramgopal  <http://orcid.org/0009-0009-4366-2390>
 Greta M. Fastrich  <http://orcid.org/0000-0001-7105-9091>
 Philip A. Higham  <http://orcid.org/0000-0001-6087-7224>
 Ariana Modirrousta-Galian  <http://orcid.org/0000-0003-2925-2976>
 Mansi Pattni  <http://orcid.org/0009-0003-7512-4788>
 Rosalind Potts  <http://orcid.org/0000-0003-4042-2059>
 Tina Seabrooke  <http://orcid.org/0000-0002-4119-8389>

References

- Abel, M., & Bäuml, K. H. T. (2020). Would you like to learn more? Retrieval practice plus feedback can increase motivation to keep on studying. *Cognition*, 201, 104316. <https://doi.org/10.1016/j.cognition.2020.104316>
- 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>
- Alamri, A., & Higham, P. A. (2022). The dark side of corrective feedback: Controlled and automatic influences of retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(5), 752–768. <https://doi.org/10.1037/xlm0001138>
- Alamri, A., & Higham, P. A. (2025). Multiple-choice testing: Controlled and automatic influences of retrieval practice in an educational context. *Quarterly Journal of Experimental Psychology*, 78(10), 2065–2090. <https://doi.org/10.1177/17470218251327158>
- Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior*, 8(4), 463–470. [https://doi.org/10.1016/S0022-5371\(69\)80090-3](https://doi.org/10.1016/S0022-5371(69)80090-3)
- Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, 108(3), 296–308. <https://doi.org/10.1037/0096-3445.108.3.296>
- Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, 4(5), 316–321. <https://doi.org/10.1111/j.1467-9280.1993.tb00571.x>
- Biggs, J., Kember, D., & Leung, D. Y. P. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, 71(1), 133–149. <https://doi.org/10.1348/000709901158433>
- Biwer, F., De Bruin, A. B. H., Schreurs, S., & Oude Egbrink, M. G. A. (2020). Future steps in teaching desirably difficult learning strategies: Reflections from the study smart program. *Journal of Applied Research in Memory and Cognition*, 9(4), 439–446. <https://doi.org/10.1016/j.jarmac.2020.07.006>
- 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(3), 165–170. <https://doi.org/10.1016/j.jarmac.2014.03.002>
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). The MIT Press.
- Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, 19(5), 619–636. <https://doi.org/10.1002/acp.1101>
- Carpenter, S. K., & Pan, S. C. (2024). Spacing effects on learning and memory. In L. Mickes & J. T. Wixted (Eds.), *Learning and memory: A comprehensive reference* (3rd Edition). Elsevier. <https://doi.org/10.1016/B978-0-443-15754-7.00020-1>
- Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, 20(6), 633–642. <https://doi.org/10.3758/BF03202713>
- Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, 79(2), 162–170. <https://doi.org/10.1037/0022-0663.79.2.162>
- Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology*, 111(2), 225–236. <https://doi.org/10.1037/0021-843X.111.2.225>
- Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39), 19251–19257. <https://doi.org/10.1073/pnas.1821936116>
- 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*, 14(1), 4–58. <https://doi.org/10.1177/1529100612453266>
- Ebbinghaus, H. (1885). *Memory*. Leipzig, Germany: Duncker.
- Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (Trans. H. A. Ruger, C. E. Bussenius, & E. R. Hilgar). Dover Publications.
- Glover, J. A. (1989). The “testing” phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, 81(3), 392–399. <https://doi.org/10.1037/0022-0663.81.3.392>
- Higham, P. A., Fastrich, G. M., Potts, R., Murayama, K., Pickering, J. S., & Hadwin, J. A. (2023). Spaced retrieval practice: Can restudying Trump retrieval? *Educational Psychology Review*, 35(4), 98. <https://doi.org/10.1007/s10648-023-09809-2>
- Higham, P. A., Zengel, B., Bartlett, L. K., & Hadwin, J. A. (2022). The benefits of successive relearning on multiple learning outcomes. *Journal of Educational Psychology*, 114(5), 928–944. <https://doi.org/10.1037/edu0000693>

- Hong, M. K., Polyn, S. M., & Fazio, L. K. (2019). Examining the episodic context account: Does retrieval practice enhance memory for context? *Cognitive Research: Principles and Implications*, 4(1), 46. <https://doi.org/10.1186/s41235-019-0202-3>
- Hopkins, R. F., Lyle, K. B., Hieb, J. L., & Ralston, P. A. S. (2016). Spaced retrieval practice increases college students' short- and long-term retention of mathematics knowledge. *Educational Psychology Review*, 28(4), 853–873. <https://doi.org/10.1007/s10648-015-9349-8>
- Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, 17(6), 649–667. [https://doi.org/10.1016/S0022-5371\(78\)90393-6](https://doi.org/10.1016/S0022-5371(78)90393-6)
- Jacoby, L. L. (1999). Ironic effects of repetition: Measuring age-related differences in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(1), 3–22. <https://doi.org/10.1037/0278-7393.25.1.3>
- Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(6), 1441–1451. <https://doi.org/10.1037/a0020636>
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.
- Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 237–284). Elsevier Academic Press. <https://doi.org/10.1016/B978-0-12-800283-4.00007-1>
- Kelly, W. E. (2004). A brief measure of general worry: The three item worry index. *North American Journal of Psychology*, 6(2), 219–226.
- Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, 72(1), 269–319. <https://doi.org/10.1111/lang.12479>
- Koo, M., & Yang, S.-W. (2025). Visual analogue scale. *Encyclopedia*, 5(4), 190. <https://doi.org/10.3390/encyclopedia5040190>
- Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the “enemy of induction?”. *Psychological Science*, 19(6), 585–592. <https://doi.org/10.1111/j.1467-9280.2008.02127.x>
- Lange, K., Kühn, S., & Filevich, E. (2015). “Just another tool for online studies” (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS One*, 10(7), e0134073. <https://doi.org/10.1371/journal.pone.0134073>
- Latimier, A., Peyre, H., & Ramus, F. (2021). A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educational Psychology Review*, 33(3), 959–987. <https://doi.org/10.1007/s10648-020-09572-8>
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139087759>
- Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, 43(1), 14–26. <https://doi.org/10.3758/s13421-014-0452-8>
- Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, 23(11), 1337–1344. <https://doi.org/10.1177/0956797612443370>
- Little, J. L., Frickey, E. A., & Fung, A. K. (2019). The role of retrieval in answering multiple-choice questions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(8), 1473–1485. <https://doi.org/10.1037/xlm0000638>
- Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. S. (2020). How the amount and spacing of retrieval practice affect the short- and long-term retention of mathematics knowledge. *Educational Psychology Review*, 32(1), 277–295. <https://doi.org/10.1007/s10648-019-09489-x>
- Lyle, K. B., Bego, C. R., Ralston, P. A. S., & Immekus, J. C. (2022). Spaced retrieval practice imposes desirable difficulty in calculus learning. *Educational Psychology Review*, 34(3), 1799–1812. <https://doi.org/10.1007/s10648-022-09677-2>
- McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. I. I. (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>
- McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology*, 72(1), 609–633. <https://doi.org/10.1146/annurev-psych-010419-051019>
- Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, 4(2), 61–64. <https://doi.org/10.20982/tqmp.04.2.p061>
- Putwain, D. W., Von Der Embse, N. P., Rainbird, E. C., & West, G. (2021). The development and validation of a new multidimensional test anxiety scale (MTAS). *European Journal of Psychological Assessment*, 37(3), 236–246. <https://doi.org/10.1027/1015-5759/a000604>
- Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002), 335. <https://doi.org/10.1126/science.1191465>
- Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, 140(3), 283–302. <https://doi.org/10.1037/a0023956>
- Rawson, K. A., & Dunlosky, J. (2022). Successive relearning: An under-explored but potent technique for obtaining and maintaining knowledge. *Current Directions in Psychological Science*, 31(4), 362–368. <https://doi.org/10.1177/09637214221100484>
- Rawson, K. A., Vaughn, K. E., Walsh, M., & Dunlosky, J. (2018). Investigating and explaining the effects of successive relearning on long-term retention. *Journal of Experimental Psychology: Applied*, 24(1), 57–71. <https://doi.org/10.1037/xap0000146>
- R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Roediger, H. L., III., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Rowbotham, M., & Schmitz, G. S. (2013). Development and validation of a student self-efficacy scale. *Journal of Nursing & Care*, 02(01|1), 126. <https://doi.org/10.4172/2167-1168.1000126>
- Russo, R., & Mammarella, N. (2002). Spacing effects in recognition memory: When meaning matters. *European Journal of Cognitive Psychology*, 14(1), 49–59. <https://doi.org/10.1080/09541440042000133>
- Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, 25(5), 763–767. <https://doi.org/10.1002/acp.1747>
- Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). On the learning benefits of confidence-weighted testing. *Cognitive Research: Principles and Implications*, 1(1), 3. <https://doi.org/10.1186/s41235-016-0003-x>
- Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, 110(16), 6313–6317. <https://doi.org/10.1073/pnas.1221764110>
- Thios, S. J., & D’Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior*, 15(5), 529–536. [https://doi.org/10.1016/0022-5371\(76\)90047-5](https://doi.org/10.1016/0022-5371(76)90047-5)
- Toppino, T. C., & Bloom, L. C. (2002). The spacing effect, free recall, and two-process theory: A closer look. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3), 437–444. <https://doi.org/10.1037/0278-7393.28.3.437>
- Yang, C., Li, J., Zhao, W., Luo, L., & Shanks, D. R. (2023). Do practice tests (quizzes) reduce or provoke test anxiety? A meta-analytic review. *Educational Psychology Review*, 35(3), 87. <https://doi.org/10.1007/s10648-023-09801-w>
- Yang, C., Luo, L., Vadiello, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, 147(4), 399–435. <https://doi.org/10.1037/bul0000309>