Retrieval practice in an introductory statistics course: A randomised field trial

Priya Raman and Daniel Okafor

Department of Educational Psychology, University of Leeds

Author Note: Correspondence concerning this article should be addressed to Priya Raman, Department of Educational Psychology, University of Leeds. Email: priya.raman@example.com

Abstract

Objective: Retrieval practice reliably outperforms re-reading in laboratory studies, but evidence from full-scale, quantitatively demanding courses with delayed, high-stakes assessments remains scarce. This trial tested whether weekly low-stakes retrieval quizzes improve delayed examination performance in an authentic introductory statistics course. **Method:** First-year students (N = 118) enrolled in a single offering of the course were randomised to complete a short retrieval quiz (n = 59) or to spend an equivalent amount of timetabled time re-reading their notes (n = 59) in each of ten teaching weeks. Four weeks after the final teaching week, all students sat the same credit-bearing final examination; self-reported weekly study time was collected as a manipulation check on overall effort. **Results:** Students in the retrieval condition scored higher on the final examination (M = 74.2, SD = 10.1) than students in the re-reading condition (M = 66.8, SD = 11.9), $t(116) = 3.65$, $p < .001$, $d = 0.67$. Weekly study time did not differ between conditions, $p = .41$. **Conclusions:** A course-embedded programme of weekly retrieval quizzes produced a medium-to-large improvement in delayed, high-stakes examination performance at no cost in study time, supporting the classroom deployment of retrieval practice in quantitative disciplines.

Keywords: retrieval practice, testing effect, spaced practice, introductory statistics, study strategies, randomised field trial

Introduction

Students entering their first quantitative course meet a body of material — probability rules, sampling distributions, inference procedures — that is cumulative and unforgiving of early gaps. Concepts introduced in week two reappear in week eight wearing new notation, and procedures mastered in isolation must later be combined in multi-step problems. How students study that material therefore matters as much as how it is taught: a study strategy that consolidates each week's content as it is covered should compound across the term, whereas one that does not leaves students rebuilding the foundation at examination time.

A large body of evidence indicates that the two most common strategies differ sharply on this point. Re-reading is the strategy undergraduates report most, and it is also among the least effective: it improves familiarity with the material and a sense of fluency that learners mistake for knowledge, but its effects on delayed retention are weak (Dunlosky et al., 2013). Actively retrieving information from memory, by contrast, produces large and durable gains. Roediger & Karpicke (2006) showed that students who practised retrieval after initial study retained far more material after a delay of a week than students who spent the same time restudying — even though the restudying group felt better prepared and predicted better performance for themselves. The advantage of retrieval over restudying grows, rather than shrinks, as the retention interval lengthens, which is precisely the pattern that matters for a course with a delayed examination.

A second reason to target study strategy is that it is one of the few levers an instructor can pull without changing what is taught. Course content, assessment design and contact time are largely fixed by

programme requirements, and students arrive with the habits they already have; but the activities that fill the review time a course already contains can be specified, and the evidence indicates that students left to choose for themselves overwhelmingly choose poorly. Surveys reviewed by Dunlosky et al. (2013) place re-reading and cramming among the most frequently reported strategies, despite their weak support, and the mismatch between what students do and what works is stable across cohorts and institutions. A strategy that can be built into the timetable, then, reaches students who would never seek out advice on how to study, and it does so at the point in the term when each week's material is fresh.

The testing effect is also theoretically instructive, because the advantage of retrieval runs against learners' own judgements. In the classic demonstration, Roediger & Karpicke (2006) found not only that repeated testing beat repeated study after a delay, but that students predicted the opposite: restudy felt productive, so students who re-read expected to outperform those who had taken tests. This dissociation between felt fluency and actual retention is central to the reviews of learning techniques: strategies that make material feel familiar are mistaken for strategies that make it known, and the strategies that actually secure retention — retrieval, spacing — feel harder and are therefore rated as less useful by the people they would benefit most (Dunlosky et al., 2013). the classroom is the natural place to supply it.

Retrieval practice also outperforms study techniques that are more elaborative in character. Karpicke & Blunt (2011) compared retrieval practice with concept mapping, a technique that asks learners to build an explicit diagram of conceptual relationships. Although concept mapping is often recommended precisely because it appears to foster deep, structured understanding, students who practised retrieval outperformed those who built concept maps on delayed questions requiring inference and application. The implication for applied courses is uncomfortable for intuitive theories of study: what consolidates knowledge is not the elaborateness of the encoding activity but the requirement to reconstruct the material from memory.

The durability of these effects depends on how retrieval is distributed over time. Cepeda et al. (2006) synthesised several hundred spaced-practice studies and found that retention improves consistently when learning episodes are distributed across intervals rather than massed, with the optimal spacing increasing as the retention interval itself lengthens. For a course whose culminating assessment sits weeks after the final class, this implies that review should be spread across the term rather than concentrated in a pre-examination revision period. Weekly quizzes embedded in the course structure operationalise both mechanisms at once: each quiz is an act of retrieval, and the weekly schedule spaces those retrieval acts across the interval between learning and examination.

Despite this evidence base, most studies of retrieval practice have been conducted in laboratory settings, in single lectures or sessions, in content-light materials such as word lists and prose passages, or over retention intervals of minutes to days rather than weeks. Reviews have repeatedly called for research on classroom-embedded implementations in authentic, demanding subject areas, measured against the assessments that actually count (Dunlosky et al., 2013). Statistics is a particularly important test case. Its content is procedural and cumulative, its examinations stress application rather than recognition, student anxiety about quantitative material is high, and study habits that suffice for memorisation-heavy subjects may fail outright. If retrieval practice is to be recommended to statistics instructors on practical grounds, evidence from a full-scale course over a whole term — with a delayed, high-stakes outcome measure — is needed.

The present study reports a randomised field trial of retrieval practice in an introductory statistics course. First-year students were assigned, for the full ten weeks of teaching, to either complete weekly low-

stakes retrieval quizzes or to spend an equivalent amount of time re-reading their notes. Four weeks after the final teaching week, all students took the same final examination, which carried the usual credit weighting for the course. Because the two conditions occupied equivalent weekly time within the course structure, any difference in examination performance can be attributed to the character of the study activity rather than to the amount of it. We asked two questions. First, does a ten-week programme of low-stakes retrieval practice improve delayed examination performance relative to an equal-time re-reading condition? Second, is any effect attributable to differences in overall study time? We predicted better examination performance in the retrieval condition, with no difference in reported study time, consistent with the conclusion that the effect arises from the retrieval operations themselves rather than from additional effort (Dunlosky et al., 2013; Roediger & Karpicke, 2006).

Method

Participants

Participants were 118 first-year undergraduate students enrolled in a single offering of an introductory statistics course at a research university. All students enrolled in the course were eligible for the study; participation was embedded in the normal course structure, and all students received the same contact time and the same credit-bearing opportunities regardless of condition. Students provided informed consent for the use of their course data in research. Allocation was randomised within the course roster, with 59 students assigned to each condition. Randomisation at this scale balances observable and unobservable student characteristics between conditions in expectation, although chance imbalances remain possible and are considered in the limitations. No attrition occurred for either outcome measure: all 118 students sat the final examination, so the analysis includes every randomised participant. Because the intervention was delivered through normal course activities rather than through voluntary sign-up, the design avoids the self-selection that weakens most classroom studies of study strategies, in which the students who opt into a programme differ systematically from those who do not.

Design and conditions

The study used a two-group randomised design with condition as the between-subjects factor. Both conditions ran in parallel across the ten teaching weeks of the course, in the same weekly timetable slots, so that total contact time was identical across conditions; the manipulation concerned what students did with a fixed weekly block of class time devoted to consolidating the week's material. Ethical approval was obtained from the institutional research ethics committee; because both activities were plausibly beneficial study activities replacing unstructured review time, neither condition was disadvantaged relative to ordinary practice.

Retrieval practice condition. Students in this condition completed a short retrieval quiz each teaching week. Each quiz comprised a small number of closed- and open-response items drawn from the material covered that week and in preceding weeks, so that each act of retrieval revisited earlier content and the quizzes collectively spaced practice across the term; the questions required reconstructing definitions, procedures and applications from memory rather than recognising correct answers. Quizzes were low stakes: they carried a small fixed participation credit rather than marks scaled to performance, so that the incentive was to attempt retrieval rather than to avoid error, and feedback was provided after completion so that errors were corrected rather than consolidated. The quizzes took place in the timetabled slot, replacing an equivalent amount of unstructured note-review time.

This design follows directly from what the literature identifies as the boundary conditions of the testing effect. The quizzes drew on cumulative content because the benefit of retrieval is largest when each retrieval event reaches back beyond the most recent material, which turns a sequence of weekly quizzes into a spaced schedule rather than ten isolated checks of that week's content (Cepeda et al., 2006). The questions were constructed-response because the demonstration that retrieval outperforms concept mapping used free recall and inference demands rather than recognition formats, and because a statistics examination itself requires reconstructing and applying procedures rather than recognising them (Karpicke & Blunt, 2011). The stakes were kept low deliberately: a weekly assessment scaled to performance would measure students' ability under pressure rather than give them practice at retrieving, and would introduce the very anxiety that makes statistics courses aversive, while a fixed participation credit and post-quiz feedback preserve the essential ingredient — every student must attempt to retrieve — without the adverse incentives (Roediger & Karpicke, 2006).

Re-reading condition. Students in this condition spent the equivalent weekly time re-reading their own notes from the course — the study strategy the retrieval quizzes replaced, and the one that surveys identify as the most common among undergraduates (Dunlosky et al., 2013). They received no quizzes during the teaching weeks. The sessions were structured by the same timetable and supervised in the same way as the quiz sessions, so that supervision, setting and contact time were held constant and only the study activity itself differed. Materials, contact time and credit-bearing opportunities were otherwise identical across conditions.

Measures

Final examination performance. The outcome measure was score on the end-of-course examination, taken by all students four weeks after the final teaching week. The four-week interval is a demanding test of retention: it exceeds the intervals typical of laboratory studies by an order of magnitude and includes the examination revision period, during which both conditions were free to study as they chose. The examination covered the full course content, including procedural questions on inference and conceptual questions on sampling and probability, and was marked blind to condition using the course's standard rubric. Because the examination is the assessment that carries the students' course credit, the outcome is consequential rather than experimental: an improvement on this measure is an improvement in what the course exists to produce. Scores are reported as percentages.

Weekly study time. As a manipulation check on overall effort, students reported their typical weekly study time for the course at the end of the teaching period. This measure addresses the most obvious alternative explanation for a between-condition difference: that students in the retrieval condition simply worked harder than students in the re-reading condition. Because the timetabled components of the two conditions were equivalent by design, any difference in reported time would reflect differential voluntary study outside the class; the absence of such a difference supports the attribution of any examination effect to the study activity itself. Self-reported time is a coarse measure, and this is considered below, but a large difference in effort would be detectable even by such a measure.

Analysis

Group differences on the final examination were tested with an independent-samples t test, and the effect size is reported as Cohen's d, the difference in means expressed in pooled standard-deviation units. Weekly study time was compared between conditions with an independent-samples t test. The

significance level was .05, and exact p values are reported. The analysis was conducted on all randomised participants, as the design produced no missing outcome data.

Results

Descriptive statistics for both conditions are shown in Table 1. Students in the retrieval-practice condition scored higher on the final examination (M = 74.2, SD = 10.1) than students in the re-reading condition (M = 66.8, SD = 11.9). The difference was significant, t(116) = 3.65, p < .001, and corresponds to Cohen's d = 0.67, a medium-to-large effect by conventional benchmarks. In standard-deviation units, the retrieval condition outperformed the re-reading condition by roughly two thirds of a standard deviation on a high-stakes, delayed outcome measure — an advantage large enough to move a student from around the median of one condition to well within the upper half of the other.

Condition	n	M	SD
Retrieval practice	59	74.2	10.1
Re-reading	59	66.8	11.9

Table 1: Descriptive statistics for final examination scores (percentages) by condition

Weekly study time did not differ between conditions, t(116) < 1, p = .41. Students in the two conditions reported spending a comparable amount of time studying the course each week, so the examination advantage in the retrieval condition is unlikely to reflect additional time on task, whether timetabled or voluntary. Taken together, the two results support the interpretation that retrieval practice itself, distributed weekly across the term, produced better delayed performance than the same time spent re-reading notes — not that the quiz group studied more, but that they studied differently.

Discussion

The trial set out to test whether a course-embedded programme of weekly low-stakes retrieval quizzes improves delayed examination performance in an introductory statistics course, relative to an equal time spent re-reading notes. It does. Students who practised retrieval weekly for ten weeks scored about 7.4 percentage points higher on a final examination taken four weeks after the final teaching week, an effect of d = 0.67, with no difference in reported weekly study time between the conditions. The effect size is in the range that matters in practice: across a cohort, it implies a meaningful upward shift in the examination distribution rather than a marginal one, and it was obtained with a minimal change to the course's existing timetable rather than with additional instructional resource.

This finding extends the laboratory literature on the testing effect into a demanding, quantitatively oriented undergraduate course assessed by a delayed, high-stakes examination. It is consistent with the demonstrations by Roediger & Karpicke (2006) that retrieval practice produces superior delayed retention even when learners prefer restudying and predict the opposite, and by Karpicke & Blunt (2011) that retrieval outperforms structured non-retrieval study in material where application matters. It also answers, in one setting, the call for classroom research with authentic materials and consequential outcomes that reviews of learning techniques have made (Dunlosky et al., 2013). The weekly schedule additionally operationalised the distributed-practice conditions that the synthesis by Cepeda et al. (2006) identified as favourable to long-term retention, with spacing between retrieval events matched to a substantial retention interval.

The study-time result is the more practically important of the two findings. Because the conditions were time-equivalent by design and reported effort did not differ, the advantage of retrieval practice came not

from working harder but from studying differently. For statistics instructors, this is the strongest argument for the intervention: it requires no additional instructional time, no new staffing and no reduction in content coverage, and it replaces an activity — re-reading notes — that the evidence base identifies as among the least effective techniques available to students (Dunlosky et al., 2013). The low-stakes format matters for adoption as well. Quizzes carrying a small fixed participation credit and followed by feedback give students good reason to attempt retrieval and to learn from error, without the anxiety and gaming that high-stakes weekly testing can induce, and without the equity concerns that attach to assessments scaled to performance.

The findings also speak to the specific context of quantitative courses. Statistics content is cumulative, so retrieval that revisits earlier weeks compounds; the quizzes by design required students to reconstruct material from preceding weeks, which operationalises both the retrieval mechanism and the spacing mechanism in a single activity. For a subject in which student anxiety is high and perceived difficulty drives avoidance, a routine that guarantees regular, low-cost contact with the material across the term may confer benefits beyond those measured here, by reducing the pre-examination cramming that both conditions were otherwise free to engage in during the four-week interval.

Several limitations qualify these conclusions. The study was conducted in a single course at a single institution with one cohort of 118 students, and the outcome measure was a single examination; effects could vary with course content, examination format or institutional context, and replication across multiple courses and cohorts is needed before the size of the effect should be treated as typical. Assignment was randomised, but the intervention was delivered within a real course and could not be blinded, so instructor and student expectations are possible — though unquantified — influences. Self-reported study time is a coarse measure that may not capture differences in the quality or in the efficiency of the time spent, so the manipulation check rules out differences in quantity of effort rather than in how effort was deployed outside the timetabled sessions. Finally, the quizzes combined retrieval with feedback, so the present design cannot isolate the contribution of the feedback component, although the equivalence of contact time ensures that any combined effect is attributable to the activity the quiz group performed rather than to time on task. A further limitation concerns the manipulation check itself. Self-reported weekly study time is a single retrospective judgement made at the end of the teaching period, and such judgements are known to be imprecise; students may aggregate across weeks unevenly or estimate in round numbers. The measure therefore constrains what can be concluded about effort: it supports the claim that the conditions did not differ in reported overall time, but it cannot speak to how time was distributed across the term or how effortfully it was used. Because the two conditions occupied identical timetabled blocks, the quantity of guided study was fixed by design, so the comparison of interest is confined to voluntary study outside class; a daily diary measure would have characterised that component more finely. We note, too, that the equivalence of total time does not by itself establish that the quiz requirement displaced no other activity — students in the retrieval condition may have reduced unstructured review to compensate — though any such displacement would work against, rather than account for, the observed examination advantage.

Future work should replicate the design across multiple courses and institutions, vary the spacing of retrieval events against the retention interval as the synthesis by Cepeda et al. (2006) suggests matters, and examine whether the benefit is uniform across item types and across students of differing prior attainment and quantitative confidence. Even with those qualifications, the present results give statistics instructors direct grounds for a minimal change: ten weekly low-stakes quizzes yielded a medium-to-

large improvement in what ultimately counts — delayed, high-stakes examination performance — at no cost in study time. The result is deliberately unglamorous: no new technology, no restructuring of the course, only a change in what students were asked to do with time they already spent reviewing. That it produced a medium-to-large difference on the examination that carries the students' credit is precisely what makes the finding worth acting on.

References

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

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

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Roediger, H. L., & 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