

Mandatory retrieval test to incentivize retrieval practice of physics principles

Vegard Gjerde¹,[✉] Martino Marisaldi¹,[✉] Kjellmar Oksavik¹,[✉] Kjartan Olafsson¹,
Harald Spångberg²,[✉] and Bodil Holst¹

¹Department of Physics and Technology, *University of Bergen*, 5007 Bergen, Norway

²Department of Computer Science, *Electrical Engineering and Mathematical Sciences*,
5528 Haugesund, Norway



(Received 28 October 2024; accepted 10 February 2025; published 11 March 2025)

Students and educators often face time constraints, making it essential to develop interventions that are both easy to implement and have a substantial impact on learning and performance. One promising strategy is retrieval practice, particularly in the context of physics education, where many students struggle with basic knowledge. Prior research has shown that retrieval practice improves self-explanation and problem-solving performance, yet many students avoid it, even when resources are provided. To address this, we introduced a mandatory retrieval test with a minimum benchmark of 50% to encourage greater engagement with retrieval practice. The primary aim of this study was to explore students' experiences with the test and its impact on their use of retrieval practice. A secondary aim was to examine whether increased retrieval practice effort correlated with improved performance on factual, conceptual, and problem-solving exams. Interviews with 13 students revealed that about half crammed for the test, while the other half spaced their practice throughout the semester. Several students noted that the test motivated them to both memorize and understand the principles, with no complaints about the mandatory test. Of the 463 students, 179 achieved a perfect score, suggesting substantial engagement. Effect sizes on factual, conceptual, and problem-solving performance tests ranged from negligible to large, indicating substantial potential benefits with minimal risk. In conclusion, the mandatory retrieval test is an effective, low-effort intervention for educators, promoting greater student adoption of retrieval practice and yielding both qualitative and quantitative evidence of enhanced learning and performance.

DOI: [10.1103/PhysRevPhysEducRes.21.010119](https://doi.org/10.1103/PhysRevPhysEducRes.21.010119)

I. INTRODUCTION

Students find physics difficult [1], and both students and educators are hard-pressed for time. Therefore, there is a need for interventions that are simple to implement, ensure student adoption of effective learning strategies, and have a high benefit-to-effort ratio for both lecturers and students. Reaching mastery is especially difficult and time consuming for students with low prior knowledge [2]. Prior research has shown that students often lack foundational knowledge, like Newton's second law, even toward the end of the semester [3]. The same research showed that *retrieval practice*, the act of actively recalling information from memory rather than passively reviewing it, has very large effects on the memory accessibility of physics principles and definitions [3].

While retrieval practice has been extensively studied in controlled settings, its practical implementation in physics

education remains underexplored. This study addresses the critical gap in translating evidence-based strategies into scalable interventions that are both accessible and effective for diverse student populations in real-world physics courses.

A. Retrieval practice improves learning

Incorporating retrieval-based methods is essential for effective and meaningful learning in any educational domain [4–10]. Moreover, there is broad agreement among researchers that there is a need for more retrieval practice in education [11–14]. Retrieval, which involves actively recalling information, is more effective than *restudy*—where students review or reread materials—provided that some feedback is given when retrieval success rates are low [9]. The effect of testing and retrieval is consistently found to be around half a standard deviation for facts, problem-solving, and conceptual tests [9,10]. This aligns with previous findings on the effect of retrieval practice of physics principles on problem-solving tests [15]. To put this in perspective, some argue that, in real-world educational settings, a fifth of a standard deviation is a large effect [10]. Considering the ease of implementation for both lecturers and students, the potential benefit-to-effort ratio for retrieval practice is outstanding.

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Unlike more complex interventions designed to improve problem-solving or conceptual understanding, retrieval practice can be seamlessly integrated into existing teaching methods, such as Peer Instruction or problem-solving strategy instruction. By strengthening foundational knowledge, retrieval practice serves as a crucial scaffold for more advanced cognitive processes, enhancing the effectiveness of other interventions without adding significant complexity.

Retrieval practice can also yield more meaningful learning than elaborative study strategies, such as *concept mapping* [16], which involves connecting new information to existing knowledge by creating meaningful associations. The strong memories students gain from retrieval practice are especially important when the demand on working memory is high [17–19], whether due to task complexity or limitations in working memory.

One key advantage of retrieval practice is its simplicity and adaptability [13,14], allowing it to complement more extensive interventions. For instance, integrating a mandatory retrieval test into a physics course requires minimal resources but can amplify the benefits of more resource-intensive strategies [15,20,21], such as collaborative learning or inquiry-based problem-solving.

B. The direct effects of retrieval practice

The most obvious benefit of retrieval practice is the increased longevity of the retrieved knowledge [22], especially when students space their retrieval practice over time [23,24]. We argue that long-lived memories of physics principles should be a core goal in any physics course.

Problem-solving performance is also affected by *memory strength* [19], which refers to the durability and accessibility of a memory trace. Memory strength determines how quickly and effectively information can be retrieved, playing a critical role in solving complex problems. Retrieval practice of physics principles can directly affect the performance of physics students on problem-solving tests [15].

C. Indirect effects: Retrieval practice potentiates other learning strategies

Memory and understanding are tightly connected [25]. Having strong memories of core knowledge components makes it easier to learn new, more complex knowledge centered around the core knowledge components [18,19,26–28]. Physics principles are such core knowledge components. See the Hierarchical Principle Structure for Mechanics in Supplemental Material at [29] for examples of what we define as physics principles. Physics principles form the basis for understanding the domain. They are the core components for understanding worked solutions, and they are the core components in solving physics problems [3,30–33]. In fact, retrieval practice of physics principles has been shown to increase the quality of self-explanations when working through worked examples and to positively affect problem-solving performance [15]. We believe that

the indirect effect of retrieval practice—the potentiation of learning through other learning strategies, like elaborative encoding, self-explanation, and problem-solving—is the most important benefit of retrieval practice.

Knowledge retrieval also potentiates learning of new unrelated material [5,34]—a *forward effect*—and even when the material is complex [35]. Some proposed explanations for this somewhat strange phenomenon relate to the freeing of cognitive resources, improved encoding strategies, better isolation of the contexts of the learning episodes, and improved knowledge integration [36]. Crucially, retrieval of recently encoded knowledge potentiates the subsequent encoding of new knowledge [37]. Therefore, we encourage our students to retrieve the previous week’s physics principles, then elaboratively encode the current week’s principles, and, finally, do retrieval practice on the new principles. The forward effect is most useful for the first four items in a list of multiple items [38], which is the average number of principles and definitions per box in the Hierarchical Principle Structure for Mechanics.

D. Retrieval practice elevates lower-achieving students

Retrieval practice may be especially beneficial for lower-achieving students [3,10]. Students who tend to use poor study strategies will gain more by incorporating retrieval practice into their study routine [39], while higher-achieving students may already incorporate retrieval in their study strategies. Having strong memories of knowledge components is also more important for lower working memory students [18,19,40]. Hence, structured retrieval practice may be an instructional tool for leveling the playing field [40].

E. Retrieval practice should be distributed in time

Spacing—spreading learning episodes over time, rather than massing them together—and retrieval are important for learning [41,42]. Spaced retrieval practice improves the transfer of learning compared to regular study [43]. The effects of retrieval practice and spacing also seem to be additive [42]. Memory strength accumulation follows a power law, where the strength gained from each learning episode is additive and decays as a power function of time [24]. The memory strength gained reflects physical changes such as new receptor sites [44]. Distributing the retrieval practice in time reduces the decay rate for each learning episode [23].

The effect of increasing the amount of practice within days, especially beyond four trials, is small compared to increasing the amount of practice across days [45]. Successive relearning distributed in time has a large effect on the retention post exam [46]. Moreover, relearning speeds up over time, becoming more efficient. Finally, there is evidence that not only does the retrieval of specific facts speed up, but that central processes involved in retrieval are also speeding up when one practices fact retrieval [45].

F. Retrieval practice improves knowledge transfer

Retrieval enhances the transfer of learning compared to restudy [47]. Stronger memories are quicker to retrieve [24,48]; they are easier to hold in working memory and to manipulate [19,49], and they require less contextual cues to be retrieved from memory [50,51]. The help one gets from the contextual cues in the environment is limited and one cannot expect a full overlap between practice and test cues when needed. Hence, retrieval practice makes memories less context dependent, and may promote insight, inference, and generalization [50–52]. The extent to which retrieval practice improves transfer of knowledge is moderated by the retrieval performance and whether the memories have been elaboratively encoded [53].

Retrieval practice of physics principles transfers to self-explanation of worked examples by increasing principle-based explanations [15]. Including physics principles in self-explanations of worked examples improves transfer of knowledge to novel problems [15,31,54]. Retrieval practice also transfers to problem-solving performance [15]. Moreover, scores on a simple factual test have been shown to have a very high correlation with exam performance [3]. Finally, retrieval practice can affect the contribution of students during Peer Instruction [21].

G. Elaborative encoding before retrieval practice (and after)

Elaborative encoding, where students enrich new information by linking it with prior knowledge and creating meaningful associations, can make retrieval practice more effective [55]. The effect of retrieval practice is potentiated by well-formed knowledge structures [56]. Elaborative encoding before retrieval practice ensures that the student has some semantically meaningful intraitem associations to guide their memory reconstruction [20,57]. Active memory reconstruction—retrieval practice—strengthens both cue-item associations (e.g., cue: “Newton’s second law”; item: “ $F = ma$ ”) and intraitem associations (e.g., between force and acceleration) [58,59]. This strengthening of the intraitem associations speeds up the retrieval process of the item [24,45,59], making cognitive processes involving the memory more fluent. In contrast, merely studying an item does not entail reconstruction of a memory and the resulting strengthening of intra-item associations.

Retrieval practice can also make elaborative encoding more effective, especially when studying restricted topics [60], like physics principles. Retrieval practice improves the memory strength of the constituent knowledge pieces before elaborative encoding binds them together [26]. Even if the students fail to retrieve a memory, the retrieval attempt potentiates subsequent learning from restudy [59,61].

Retrieval practice and elaborative encoding can bootstrap the student from a state of negligible knowledge, with elaborative encoding potentiating subsequent retrieval

practice, and retrieval practice potentiating subsequent elaborative encoding.

H. A hierarchical visual structure may improve effects of retrieval practice

The instructional material we use for retrieval practice is a visual structure that meaningfully organizes the most important physics principles for each of the physics courses; see the Hierarchical Principle Structure and retrieval sheets for each of the courses in this study in Supplemental Material at [29]. This enables our students to store item-location combinations, giving them spatial cues for memory items in addition to the semantic cues. With practice, they can potentially bind all items together in a configuration and reduce *proactive interference*—when older memories interfere with the retrieval of newer, similar information—for other recently studied items [62]. Location and gaze direction can be an additional cue to memories when multiple memories compete [62–64]. We are better at remembering information when it is consistently presented in the same location [65], with the benefit coming from the spatial structure [64]. Moreover, visual working memory performance improves with *semantic structure* [66], the meaningful organization of information based on relationships between concepts. Hence, by using hierarchical principle sheets, we can capitalize on our exceptional ability to remember spatial information, similar to what you gain by using the Method of Loci [67] but with the added benefit that the placement can yield additional semantic information.

I. Students avoid retrieval practice

Regrettably, students tend to choose ineffective learning strategies [68,69]. They avoid mentally effortful learning strategies, both because of the mental cost and because of a misperception of effort as a cue to poor learning [70]. Many students avoid retrieval practice, even when they are told about its effectiveness and how to use it [71,72]. In fact, most interventions have little or no effect on students’ use of retrieval practice [73]. Researchers have started to promote multifaceted intervention, where students are informed of benefits, plan how to use it, and reflect on outcomes, as it is not enough to teach students how, when, and why to use retrieval practice [71,74]. Some argue that interventions to promote retrieval practice must reduce the perceived cost to succeed [73,74]. This aligns with the cognitive ACT-R theory of Anderson and Lebiere [51] and the expectancy-value-cost theory of motivation [75,76], where actions are chosen based on estimations of utility, probability of success, and cost.

J. The current study

Despite its proven effectiveness, many students avoid retrieval practice even when aware of its benefits and provided with the necessary materials [20]. Hence, we

introduced a mandatory retrieval test, requiring students to score at least 50% to be eligible for the final exam. The purpose of the test was to incentivize retrieval practice, preferably spaced throughout the semester. Our approach targets most of the considerations above. If students must complete a retrieval test, the perceived utility of retrieval practice would presumably increase and the cost of not practicing increases. Regular retrieval practice gives students immediate feedback on their improvements, presumably increasing their perceived probability of success or self-efficacy. Regular practice can also speed up the process and reduce the perceived cost of retrieval practice. Targeted elaborative encoding before retrieval practice of the physics principles can also reduce the mental cost. Moreover, having access to ready-made retrieval sheets reduces the cost of retrieval practice.

Unlike previous approaches, our intervention uniquely combines mandatory retrieval testing with a hierarchical principle structure to promote systematic and distributed retrieval practice. This dual approach not only incentivizes consistent engagement but also supports students in organizing and integrating core physics principles, bridging the gap between factual recall and conceptual understanding.

The study was guided by the following research questions:

1. Does the mandatory retrieval test increase students' use of retrieval practice?
2. How do students perceive and reflect on the mandatory retrieval test as part of their study process?
3. Is increased engagement in retrieval practice associated with improved performance on factual, conceptual, and problem-solving assessments?

The primary aim of this study was to explore students' experiences and reflections on the mandatory retrieval test and retrieval practice. Specifically, we sought to investigate whether the test incentivized retrieval practice and whether increased effort in retrieval practice correlated with higher performance on tests and exams. By addressing these questions, this study aims to provide insights into the effectiveness and scalability of retrieval practice interventions in physics education.

II. METHODS

A. Participants

This research was conducted over two years in three consecutive bachelor's-level physics courses at a large Norwegian university and an introductory mechanics course at a large Norwegian university of applied sciences. About half to two-thirds of each course cohort gave written consent. See Tables II and III in Sec. III—Quantitative results for the exact number of students from each course cohort. Thirteen students from the 2022 Mechanics 1 cohort consented to partake in one-on-one interviews. Mechanics 1 was chosen for qualitative data collection due to its large cohort size, which facilitated recruiting a

diverse group of students. This diversity allowed us to capture a wide range of experiences and perspectives, spanning the entire grade interval on the final exam and including students with varied prior results. The large cohort size also increased the likelihood of reaching thematic saturation in the qualitative analysis, ensuring that no new themes emerged with additional data collection.

B. Context

1. Research project

The courses were part of a research project where the aim was to integrate effective learning strategies into the instructional material and learning arenas. The lecturers decided which of the suggested instructional interventions they wanted to implement. However, three of the four courses implemented a standardized test of factual knowledge (Classical Mechanics did not), and all four courses implemented a standardized multiple-choice test and conceptual multiple-choice questions with written argumentation on the final exam (15%–30% of the exam score). Moreover, all course instructors designed, and gave the students access to, a principle structure that meaningfully organized the most important principles and definitions in the courses. Retrieval practice sheets were made based on the principle structures. See Supplemental Material at [29] for the principle structure and retrieval sheet for each course. Moreover, each course implemented a mandatory retrieval test, in which students had to score a minimum of 50% to be allowed to take the exam. The retrieval test was identical to the retrieval sheets.

2. Courses

Introduction in Astrophysics is a first-semester physics course for most students, with a typical enrollment of 40–50 students. The course covers a broad array of concepts relevant to astrophysics without going into depth. The course is a mix of weekly lectures, seminars, and a large project-based component where students work in groups for several weeks. One key goal of the project-based work is to give students practice in scientific writing, including peer-review of each other's writing. Of the courses involved, it has the lowest fit to the instructional interventions and is also the least changed course beyond the newly implemented tests, the principle structure, and the introduction of peer review of scientific writing. The course included some structured retrieval practice sessions during lectures in the final year of the project and minor additions of Peer Instruction.

Mechanics 1 is a second-semester physics course for most of the attending students, with a typical enrollment of 150–180 students. The course covers translational-, rotational-, and fluid mechanics. The lectures were almost completely student-active, with one hour of discussion problems, two hours of Peer Instruction, and one hour of

problem-solving strategy practice during the two weekly two-hour lectures. The lectures also included a 5–10-minute individual structured retrieval session. As preparation for the lectures, the students were encouraged to use the introductory videos created by the course instructor. These videos were intended to stimulate elaborative encoding of the physics principles and definitions from the hierarchical principle structure for mechanics [57]. The students were also encouraged to do structured retrieval practice on the retrieval sheets after engaging with the week's videos. Specifically, they were advised to first do retrieval practice of the principles from previous weeks, then watch the current week's videos, and finally do retrieval practice of the current week's principles. They were also advised to gradually drop old principles from their retrieval practice, as the first four re-learnings yield most of the benefits. Moreover, they were encouraged to kickstart all their study sessions with retrieval practice.

Mechanics 2 and Thermodynamics is a third-semester physics course for most of the attending students, with a typical enrollment of 50–70 students. Most of the students came from Mechanics 1 and were well-versed in retrieval practice and the other study practices taught. The course covers wave mechanics, gravitation, Hamiltonian and Lagrange formalism, and thermodynamics. In the treatment years, the lectures were mostly traditional with some Peer Instruction in the first two-thirds of the semester. The final third included substantial Peer Instruction. The course had structured retrieval practice in-lecture for the whole semester during the treatment years. This course was the most changed out of the courses in the study.

Classical mechanics is a second-semester physics course for engineering students at a Norwegian university of applied sciences. The typical enrollment is 30–40 students and covers translational and rotational mechanics. The main change made to the course was the addition of the mandatory retrieval test and structured retrieval practice in-lecture. The course already included Peer Instruction before the intervention.

C. Measures and data gathering

1. Retrieval tests

The retrieval tests were performed in class the week after all the course concepts had been covered. See Supplemental Material at [29] for the retrieval test for each course. The course instructors manually scored each retrieval test. The students received 1 point for formally correct responses, even if the equation was written in a different (but formally correct) way compared to the formulation in the principle structure. The students were required to score a minimum of 50% to be allowed to take the final exam. Of the 463 students, only three students failed the test on the first try.

2. Factual tests

The factual tests were performed on a web-based assessment platform the week after the retrieval test. The factual tests were developed by the course instructors, covering SI-units of concepts, important physics principles, and conditions of application of principles. See Supplemental Material at [29] for the factual tests for each course, and the corresponding test analysis. The factual test was not implemented in the Classical Mechanics course because the course already included three tests, so the lecturer at that time decided against adding a fourth to avoid overburdening the students. Both the Cronbach's alpha values (ranging from 0.81 to 0.85) and the mean inter-item correlations (ranging from 0.15 to 0.22) were in the ideal range [77,78]. All items had positive discrimination values, with almost all items having good to excellent item discrimination [79].

3. Conceptual multiple-choice tests

The conceptual multiple-choice tests were also performed on a web-based assessment platform, the same week as the factual test. We used established and pre-validated conceptual tests in all courses. The Astrophysics course used the Test of Astronomy Standards [80]. The Mechanics 1 and Classical Mechanics courses used the Mechanics Baseline Test [81]. The Mechanics 2 and Thermodynamics course used the Survey of Thermodynamic Processes and First and Second Laws [82].

4. Final exams

The exams in all the courses had the same basic structure. Most of the exam consisted of traditional text problems, counting for 70% of the final grade (85% in Classical Mechanics, CM). There were five (three in CM) conceptual multiple-choice problems, counting 10% (5% in CM) of the final grade. Finally, the students had to give a written argument for their choice, counting 20% (10% in CM) of the final grade. To receive full points for their argumentation, they had to argue with physics principles and be formally correct.

5. Interviews

The interviews lasted approximately one hour on average and were conducted digitally by the first author in the final weeks of the 2022 semester in Mechanics 1. The semi-structured interviews followed a guide with questions targeting students' experiences and reflections on different learning strategies and course materials. A subset of the questions was relevant to the current study. See Supplemental Material at [29] for the interview guide.

The interviewer strove to follow guidelines for qualitative interviews [83] which emphasize creating a safe and open environment for participants to share their experiences, asking open-ended and nonleading questions, and allowing flexibility in the interview structure to explore

emerging themes. This approach ensures that data collection is both rigorous and sensitive to the participants' perspectives, aligning with best practices in qualitative research.

D. Data analysis

1. Quantitative analysis

The quantitative analysis was performed in R [84]. Pearson's correlation assumes that the relationship is linear and may underestimate the effect if it is not [85]. Both the retrieval test and the factual test had ceiling effects, which also potentially lead to an underestimation of effect sizes. We also conducted a median split effect size analysis, in which we split the data so that high retrievers had zero or one wrong answer, and low retrievers had more than one wrong answer. This resulted in 220 high retrievers and 243 low retrievers in total (212 and 211 in the case of the factual test). A qualitative reason for doing this split is that to get a full score (or only a single mistake) demands much more retrieval practice beforehand than being somewhat close to a full score. However, some information is lost when performing a median split [86]. Therefore, the effect sizes should be interpreted in relation to the correlation coefficients.

2. Qualitative analysis

The interviews were transcribed intelligent verbatim by a professional transcription company, which specializes in research transcription and adheres to strict confidentiality agreements. The data were analyzed using thematic analysis, following the approach outlined by Braun and Clarke [87], and processed using Nvivo software [88]. To ensure anonymity, all participants were assigned pseudonyms.

The thematic analysis proceeded through several key stages. First, the first author listened to all audio recordings multiple times throughout the process to gain a deep familiarity with the data and to identify potential missed themes or inconsistencies. The transcribed interviews were then segmented into broad categories, followed by a detailed coding of specific utterances within the data.

The first author identified themes, and these themes were organized hierarchically to show relationships between them. Throughout the process, themes underwent continuous refinement, where they were modified, collapsed, or separated in response to issues of internal coherence (the consistency within a theme) and external distinguishability (the differences between themes) [87]. Finally, the qualitative findings were written up and then reviewed and scrutinized by the other authors.

While thematic analysis can provide rich insights, it is important to acknowledge potential bias stemming from the familiarity between the first author (who conducted the interviews) and the participants, as the first author was also the course instructor. Personal involvement in data collection may have influenced the dynamics of the interviews.

Specifically, while the students appeared to be open and critical during the discussions, there is a possibility that positive students became more enthusiastic, whereas neutral students may have underplayed their actual feelings, potentially being more negative. This is an important consideration when interpreting the findings.

Furthermore, as Firestone [89] has pointed out, it is inappropriate to generalize qualitative findings from a specific sample to an entire population. However, despite this limitation, we believe that the identified themes accurately reflect the experiences and reflections of the students in this study. The themes provide a meaningful representation of their perspectives, and while not generalizable in the traditional sense, they can serve as working hypotheses that may be tested or explored in other educational settings [90].

Finally, as Lincoln and Guba [91] emphasize, case-to-case generalization requires readers to carefully consider whether the findings from this study apply to their specific context. We invite readers to reflect on the relevance of the identified themes for their situation and to approach the transfer of these findings with appropriate caution.

III. QUANTITATIVE RESULTS

The quantitative results in this section paint a broad picture of trends and correlations. It is particularly valuable in tandem with the qualitative findings. In accordance with established guidelines, Cohen's d values of 0.2, 0.5, and 0.8 are typically considered to represent small, medium, and large effect sizes, respectively. Similarly, correlations are generally categorized as small when $r < 0.2$, medium when $0.2 < r < 0.3$, and large when $r > 0.3$ [92,93].

Table I shows descriptive statistics for the scores on the mandatory retrieval test for each of the course cohorts. It shows that most of the students were close to a 100% score on the retrieval test in most of the course cohorts. In all but one course cohort, the proportion of students who scored 100% was above 25%, with one course cohort having more than half of the students at 100% score. Classical Mechanics stood

TABLE I. Mean score, median score, and the proportion of students who scored 100% on the mandatory retrieval test in each course.

Course	Mean	Median	Proportion full score
Astrophysics 2022	90%	95%	27%
Astrophysics 2023	90%	94%	35%
Mechanics 1 2022	92%	97%	34%
Mechanics 1 2023	93%	98%	44%
Mechanics 1 2024	94%	100%	55%
Mechanics 2 & TD 2022	91%	98%	31%
Mechanics 2 & TD 2023	95%	97%	33%
Classical Mechanics 2023	79%	77%	6%
Classical Mechanics 2024	82%	86%	26%

TABLE II. Correlation and median split effect size for the factual test, conceptual multiple-choice test, and exam for each course, with a continuously cumulating meta-analysis (CCMA). The median split effect size compared the students with 0–1 wrong answers on the retrieval test to the students with 2+ wrong answers. Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ^a $p < 0.10$. The effect sizes in bold are statistically significant.

Test type:		Facts		Conceptual		Exam	
Course:	n	Correlation	Cohen's d	Correlation	Cohen's d	Correlation	Cohen's d
Astrophysics 2022	22	$r = 0.35$	$d = \mathbf{1.09}^*$	$r = -0.07$	$d = -0.23$	$r = 0.35$	$d = 0.34$
Astrophysics 2023	20	$r = \mathbf{0.61}^{**}$	$d = 0.63$	$r = -0.05$	$d = 0.31$	$r = 0.33$	$d = 0.25$
Mechanics 1 2022	115	$r = \mathbf{0.35}^{***}$	$d = \mathbf{0.58}^{**}$	$r = 0.04$	$d = 0.16$	$r = \mathbf{0.34}^{***}$	$d = \mathbf{0.46}^*$
Mechanics 1 2023	108	$r = \mathbf{0.40}^{***}$	$d = \mathbf{0.77}^{***}$	$r = 0.16^a$	$d = 0.12$	$r = \mathbf{0.48}^{***}$	$d = \mathbf{0.82}^{***}$
Mechanics 1 2024	92	$r = \mathbf{0.29}^{**}$	$d = \mathbf{0.63}^{**}$	$r = 0.17^a$	$d = 0.29$	$r = \mathbf{0.35}^{***}$	$d = \mathbf{0.56}^{**}$
Mechanics 2 & TD 2022	36	$r = \mathbf{0.52}^{**}$	$d = \mathbf{1.14}^{**}$	$r = 0.16$	$d = 0.43$	$r = 0.23$	$d = 0.32$
Mechanics 2 & TD 2023	30	$r = \mathbf{0.38}^*$	$d = 0.46$	$r = \mathbf{0.48}^{**}$	$d = \mathbf{0.93}^*$	$r = \mathbf{0.63}^{***}$	$d = 0.63^a$
Classical Mechanics 2023	17	...	...	$r = 0.13$	$d = 0.66$	$r = 0.66$	$d = 0.94$
Classical Mechanics 2024	23	...	...	$r = 0.01$	$d = -0.18$	$r = 0.18$	$d = 0.16$
CCMA	463 (423)	$r = \mathbf{0.40}^{***}$	$d = \mathbf{0.71}^{***}$	$r = \mathbf{0.13}^{**}$	$d = \mathbf{0.23}^*$	$r = \mathbf{0.39}^{***}$	$d = \mathbf{0.54}^{***}$

out with the lowest score and low proportion of students at 100% score.

Table II shows the correlation between the score on the mandatory retrieval test and the factual test, conceptual multiple-choice test, and exam scores, and the corresponding median split effect size. Note that the factual test was not implemented in Classical Mechanics. Table III shows the correlation between the score on the mandatory retrieval test and the text problems, multiple-choice problems, and argumentation problems from the exam, and the corresponding median split effect size.

Nonsignificant results from smaller sample sizes may provide additional evidence of effects [94,95]. Therefore, we performed a continuously cumulating meta-analysis (CCMA) on the effect sizes of every sample. The CCMA on Cohen's d effect sizes was performed using a spreadsheet from Sibley [96], which provides utilities for calculating simple meta-analytic averages. This tool is designed to facilitate the analysis and visualization of meta-analytic

data, making it accessible for researchers to compute and interpret aggregated effect sizes efficiently. The CCMA on the correlation coefficients was performed with the meta package in R [97]. The CCMA showed a significant effect of retrieval test scores on factual test performance, $r = 0.40$, $p < 0.001$ and $d = 0.71$, $p < 0.001$; conceptual test performance, $r = 0.13$, $p < 0.01$ and $d = 0.23$, $p < 0.05$; and exam performance, $r = 0.39$, $p < 0.001$ and $d = 0.54$, $p < 0.001$. For the subtypes of problems on the exam, the CCMA showed a significant effect of retrieval test score on text problems, $r = 0.40$, $p < 0.001$ and $d = 0.55$, $p < 0.001$; conceptual multiple-choice questions, $r = 0.18$, $p < 0.001$ and $d = 0.22$, $p < 0.05$; and argumentation, $r = 0.29$, $p < 0.001$ and $d = 0.37$, $p < 0.001$.

IV. QUALITATIVE FINDINGS AND DISCUSSION

We present the findings from the qualitative analysis as a narrative, integrating discussion throughout and highlighting

TABLE III. Correlation and median split effect size for text problems, multiple-choice problems, and argumentation problems on the exam for each course, with a continuously cumulating meta-analysis (CCMA). The median split effect size compared the students with 0–1 wrong answers on the retrieval test to the students with 2+ wrong answers. Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ^a $p < 0.10$. The effect sizes in bold are statistically significant.

Problem type:	n	Text problems		Multiple-choice		Argumentation	
		Correlation	Cohen's d	Correlation	Cohen's d	Correlation	Cohen's d
Astrophysics 2022	22	$r = 0.38^a$	$d = 0.35$	$r = 0.19$	$d = 0.24$	$r = 0.03$	$d = 0.08$
Astrophysics 2023	20	$r = \mathbf{0.46}^*$	$d = 0.45$	$r = 0.04$	$d = -0.07$	$r = -0.09$	$d = -0.32$
Mechanics 1 2022	115	$r = \mathbf{0.35}^{***}$	$d = \mathbf{0.51}^*$	$r = 0.07$	$d = 0.12$	$r = \mathbf{0.23}^*$	$d = 0.24$
Mechanics 1 2023	108	$r = \mathbf{0.49}^{***}$	$d = \mathbf{0.82}^{***}$	$r = \mathbf{0.20}^*$	$d = 0.34^a$	$r = \mathbf{0.36}^{***}$	$d = \mathbf{0.61}^{**}$
Mechanics 1 2024	92	$r = \mathbf{0.32}^{**}$	$d = \mathbf{0.47}^*$	$r = 0.16$	$d = \mathbf{0.45}^*$	$r = \mathbf{0.42}^{***}$	$d = \mathbf{0.73}^{***}$
Mechanics 2 & TD 2022	36	$r = 0.22$	$d = 0.35$	$r = 0.09$	$d = 0.11$	$r = 0.17$	$d = 0.04$
Mechanics 2 & TD 2023	30	$r = \mathbf{0.58}^{***}$	$d = 0.61$	$r = \mathbf{0.60}^{***}$	$d = 0.60$	$r = \mathbf{0.51}^{**}$	$d = 0.43$
Classical mechanics 2023	17	$r = \mathbf{0.64}^{**}$	$d = 0.89$	$r = \mathbf{0.53}^*$	$d = 0.66$	$r = 0.46^a$	$d = 1.16$
Classical mechanics 2024	23	$r = 0.19$	$d = 0.18$	$r = 0.06$	$d = -0.10$	$r = 0.01$	$d = -0.22$
CCMA	463	$r = \mathbf{0.40}^{***}$	$d = \mathbf{0.55}^{***}$	$r = \mathbf{0.18}^{***}$	$d = \mathbf{0.22}^*$	$r = \mathbf{0.29}^{***}$	$d = \mathbf{0.37}^{***}$

TABLE IV. Summary of qualitative findings.

Themes	Subthemes
When they practiced	Seven practiced regularly; five crammed before the test; one did not practice at all.
Benefits of retrieval practice	Improved study habits, better problem-solving, reduced effort during tasks.
How they practiced	Methods included Anki, retrieval sheets (paper/tablet); approaches ranged from superficial to deep.
Attitudes toward the test	Seven positive, four neutral, and two thought it was unnecessary for themselves but useful for others.
Experience of the test	Neutral to positive overall; initial stress relieved by test transparency.
Benefits of the test	Motivated retrieval practice, improved understanding, and inspired peer competition.

key insights with illustrative quotes from the students. A summary of the findings can be seen in Table IV. This section explores various themes, such as students’ retrieval practice habits, the benefits they experienced, and their attitudes toward the mandatory retrieval test. Through a combination of student reflections and analysis, we examine when and how students practiced retrieval, the benefits they experienced, and their attitudes toward and experiences with the mandatory retrieval test. The discussion is enriched with direct quotes that capture the students’ diverse experiences, offering a comprehensive view of how retrieval practice impacted their learning process.

A. When did they practice? crammers versus regulars

One student did no retrieval practice, as they felt certain that they would pass the test anyway.

Five students did no retrieval practice until a few days or weeks before the mandatory retrieval test. Two of these students reported that they noticed the benefits of retrieval practice after the test. Their reasons for not doing regular retrieval practice were diverse, with no reason repeating among students.

Seven students did regular retrieval practice throughout the semester. Some of these students made an extra effort in the days before the test. Two of the regulars reported using retrieval practice as a kickstart of study sessions.

Starting with retrieval practice has been really helpful for getting into problem-solving. In the morning, when you’re tired and having coffee, it’s easy to put on a podcast or something, and suddenly 10–15–20 minutes slip away. So, the retrieval practice has been great for getting going in the mornings.

Six of the regulars reported doing a retrieval practice session right after having watched the videos on the current week’s physics principles, following the advice of the course lecturer.

B. Benefits of regular retrieval practice

The regulars reported multiple benefits of doing regular retrieval practice, with all the benefits being reported by at

least two students. First, retrieval practice made it easier to start study sessions. They also felt better prepared for lectures and problem-solving,

It has helped me a lot. I wasn’t very interested in doing it at the beginning of the year. I thought it wasn’t that important, but I’ve since realized that the more I drill and practice retrieval training, the more the principles stick. This way, I don’t have to think so much about the tasks. I know that it’s that principle. I probably save some time by actually studying and learning them, so I can use the knowledge later when working on tasks.

They reported remembering principles when needed and recognizing principles while studying. One student said: “I’m sitting in the lectures, trying to guess which principle you used.”

Finally, some students found retrieval practice enjoyable

In the beginning, I thought it seemed incredibly stressful to have to memorize all these principles and need 50% to pass. But once I actually started working on it, my perspective completely changed. I actually found it quite fun to work with the principles.

C. How did they practice?

One student used the free spaced retrieval software Anki, while seven used the retrieval sheets on paper and four used a tablet. The tablet users were particularly satisfied with the process.

Three students focused on superficial aspects of the principles during retrieval, such as recognizing symbols or counting the number of symbols. The other students engaged in deeper thought processes, like using analogical similarities between the equations, the meaning of the symbols, how the principle should logically be constructed to make sense, and by referring to their experience from problem-solving.

I think about what all the different symbols mean and what this principle is actually supposed to give me.

Four students talked about how they needed to understand each principle before they started retrieval practice, with one noting that the retrieval practice also served as a test of understanding. Four students talked about how they did the retrieval practice first, building an understanding afterward.

At first, I just practiced retrieval 20 times in a row, writing the same principle each time. The principle stuck—or rather, I had just memorized the term. But then I tried to understand why that condition leads to that outcome. It was actually quite obvious.

All but one student said that doing retrieval practice feels intuitive, and that the student did not feel a need for advice on how to do it.

D. What were their attitudes toward the retrieval test?

Crucially, none of the students had a negative attitude toward the test. Two of the students said that the test was unnecessary for them, but that it was probably necessary for many of the other students. Seven students were positive about the retrieval test being mandatory, one being especially enthusiastic:

I love it. I think it's fantastic. I really appreciated that you made us learn all of them through the retrieval test. I probably wouldn't have learned them as well if it hadn't been for that.

The rest were neutral toward the test, with no specific feelings either way.

E. How did they experience the retrieval test?

Three students talked about the test producing some feelings of stress during the semester and two students mentioned that they knew others who were stressed about the test. However, almost all the interviewees had a neutral or positive experience during the test, a typical response being:

It went just fine. Yeah, it was really just... I don't have many thoughts about it, to be honest. It was just a matter of sitting down, getting it done, and then I was finished with it.

Four students talked about how the transparency of the test—that it was identical to the retrieval sheets—was helpful for relieving anxiety. One student suggested increasing the benchmark from 50% to 80%

F. What benefits did the test yield?

Seven students talked about how the test motivated them to do retrieval practice, either throughout the semester or cramming before the test.

I think it's been really great to have something that motivated me to do retrieval practice. I don't think I would have been as diligent with it if I didn't know there was a test. So, I believe it has helped and pushed me to engage in retrieval practice because we have that test.

Four students said that the test motivated them to also understand the principles. Two students said that having the retrieval test improved their study structure, with more regular studying than before. Two students mentioned how the students in their study program started a competition:

A little competition started within our physics class, where people were like, 'Sure, we only need 50%, but we're all going for 100%.'

Finally, two students from the student-teacher program talked about how they had transferred retrieval practice to their concurrent math classes.

Yeah, I've applied it to other subjects as well. And it's worked really well, in my opinion. I'm really happy with it.

V. GENERAL DISCUSSION

The findings from both the quantitative and qualitative data in this study provide valuable insights into the effects of mandatory retrieval practice on student performance across physics courses. The results suggest that retrieval practice has a meaningful impact on factual recall, problem-solving, principle-based argumentation, and, to a lesser extent, conceptual understanding.

A. Retrieval practice enhances performance

The quantitative results demonstrate that students who engaged more with retrieval practice performed better on factual tests, problem-solving tasks, and principle-based argumentation, consistent with prior research [3,15], with a crucial assumption being that no student would score 100% on the retrieval test without doing retrieval practice. The effect sizes for these areas were substantial, with large correlations and corresponding Cohen's d values nearing or exceeding the medium to large effect size benchmarks ($r = 0.40$, $d = 0.71$ for factual tests; $r = 0.40$, $d = 0.55$ for text-based problems, and $r = 0.29$, $d = 0.37$ for principle-based argumentation). The strong correlation between retrieval test scores and performance on text-based and argumentation problems underscores that retrieval practice does more than promote rote memorization. These findings suggest that students who consistently engage in retrieval practice develop a more robust understanding of physics principles, which they can effectively apply to problem-solving and reasoning tasks.

Interestingly, the impact of retrieval practice was small for conceptual multiple-choice tests and multiple-choice problems on the exams, indicating that retrieval practice may not enhance all types of knowledge equally. This finding suggests that retrieval practice is particularly beneficial for strengthening the memory of factual information and improving problem-solving skills but has a more limited effect on multiple-choice conceptual understanding. This nuanced result suggests that stronger students are not the only ones who tend to use retrieval practice when encouraged. Students across the performance spectrum use it and gain from it. Moreover, the findings from this study align with prior experimental research on retrieval practice, with effect sizes comparable to the experimental findings from [3,15]. Similar to the current results, Gjerde and Paulsen [15] found a smaller effect of retrieval practice on a multiple-choice test than on a problem-solving test.

The qualitative findings provide valuable context for the observed limited, but significant, effect of retrieval practice on conceptual multiple-choice tests. While no interviewees specifically addressed multiple-choice tests, several students noted that their retrieval practice helped them with problem-solving. This alignment with problem-solving rather than multiple-choice conceptual tasks may partly explain the smaller observed effects. Additionally, the Mechanics Baseline Test [81], used as the conceptual measure in the two mechanics courses, has historically shown resistance to various interventions compared to other measures, perhaps due to its emphasis on concepts commonly encountered in high school. In contrast, Mechanics 2 in 2023, which used the STPFaSL conceptual test [82], exhibited a larger effect.

Moreover, the timing of the conceptual tests, administered only a week after the retrieval test, likely limited the potential for crammers to benefit. These students had minimal opportunity to translate their stronger memory of principles into deeper conceptual understanding.

The interviews also highlighted differences in how students engaged with retrieval practice. Some focused superficially on symbols or rote memorization, which may help explain the smaller gains in conceptual understanding observed in the quantitative data. However, the large effect observed in principle-based argumentation indicates that many students did achieve a deeper conceptual understanding. This discrepancy raises questions about the extent to which conceptual multiple-choice tests accurately measure conceptual understanding.

While direct research on the retrieval practice of domain principles in STEM fields is limited beyond references [3,15], laboratory studies on retrieval practice offer valuable insights. These studies often focus on strengthening cue-item associations, with the most substantial effects occurring when the associations are semantically meaningful [9,36]. Physics principles, such as Newton's laws or conservation laws, exhibit a highly interconnected semantic structure,

making them particularly well-suited for retrieval practice interventions. This study's findings suggest that retrieval practice effectively leverages these semantic connections, enhancing the accessibility and durability of principles for application in problem-solving and argumentation. Furthermore, our results align with broader research on retrieval practice in educational settings, which consistently demonstrates its benefits for recall and learning transfer across diverse subjects [10,22].

B. Mandatory testing promotes regular retrieval practice

The mandatory retrieval test appears to have been one of the most effective interventions to date for encouraging students to consistently engage in retrieval practice. Up to 50% of students achieved full scores on the test, suggesting that substantial effort was put into retrieval practice. Qualitative feedback also highlighted that students were motivated by the existence of the test, with many acknowledging that they would not have practiced retrieval as diligently without it.

C. Considerations for raising the bar

While the 50% benchmark for passing the retrieval test ensured a baseline level of engagement, the data suggests that some students could meet this criterion without fully committing to retrieval practice. Increasing the passing threshold—perhaps to 80% or even 100%—might encourage more consistent and thorough practice, especially in introductory physics courses where students often arrive with varying levels of prior knowledge.

There is also potential for implementing spaced, expanding retrieval tests or integrating peer review processes, where students score each other on the retrieval tests, to further enhance the depth and retention of knowledge. These approaches could increase engagement and help ensure that retrieval practice is not just a one-time effort but a continuous learning process throughout the semester.

D. Acceptance and transferability of the approach

Mandatory work often imposes external pressure on students and can sometimes generate resistance or animosity. Therefore, the overall positive or neutral reception of the mandatory retrieval test among students is an important indicator of the intervention's success. Acceptance suggests that the intervention was not only tolerated but appreciated by many students, which is crucial for its sustainability and potential scalability. As indicated in the qualitative results, two students transferred their retrieval practice habits to concurrent mathematics courses. This observation highlights the potential for broader applications of retrieval training beyond physics, showcasing its value in improving study habits across disciplines.

Moreover, the retrieval practice method is easy for both educators and students to implement, with clear goals and transparent processes. The up-front investment of creating hierarchical principle structures is minimal compared to the potential long-term benefits. Notably, students have even begun requesting similar principle structures in other courses, further demonstrating their appeal and utility.

VII. CONCLUSIONS

This study confirms the substantial benefits of retrieval practice for enhancing factual recall, problem-solving, principle-based argumentation, and to a lesser extent, conceptual understanding. The mandatory retrieval test may be the most effective intervention to date for increasing the use of retrieval practice. Furthermore, the practice has shown potential for transfer to other academic domains, such as mathematics, suggesting that it could serve as a promising approach for improving study habits more broadly. By raising the bar for retrieval test benchmarks and exploring additional methods like spaced testing or peer review, educators can further optimize this powerful learning strategy.

A. Strengths and limitations of this study

There were ceiling effects on the retrieval tests and factual tests, which may suppress the effect sizes. A more substantial weakness is that the design of this study makes it impossible to rule out the influence of third variables, for example, that it was just the stronger students who did retrieval practice. However, the cross-checking between quantitative results and qualitative findings adds strength to the study, as students reported retrieval practice helped them remember principles, prepared them for problem-solving and lectures, and improved their study habits.

B. Open questions and future research

While this study provides strong evidence of the benefits of retrieval practice in physics education, several important questions remain open for further investigation.

One key area for future research is how to encourage more students to begin retrieval practice earlier in the semester, engage with it consistently, and focus more on understanding the underlying principles rather than solely memorizing. One potential approach is to implement spaced retrieval tests with progressively expanding scope throughout the semester. This could help reinforce knowledge over time and promote deeper comprehension, while also preventing cramming behavior right before exams or a final mandatory retrieval test.

Another critical avenue for future research involves conducting experimental studies to establish the causal relationship between retrieval practice and improved academic performance. While this study's findings align with previous experimental research, additional controlled experiments would further validate the causality of the observed effects. By randomly assigning students to different retrieval practice conditions, researchers can better isolate the specific factors that contribute to enhanced learning outcomes.

Finally, expanding research beyond physics into other subject areas, such as mathematics or social sciences, could reveal the broader applicability of retrieval practice across disciplines. Testing whether retrieval practice can be effectively transferred to different educational contexts would provide valuable insights into its potential as a universal learning strategy.

ACKNOWLEDGMENTS

This work was supported by The University of Bergen. We would like to thank HKDIR/DIKU for funding the research project Aktiv-2019/10122.

-
- [1] F. Örnek, W. R. Robinson, and M. Haugan, What makes physics difficult, *Int. J. Environ. Sci. Educ.* **3**, 30 (2008), https://www.researchgate.net/publication/228689133_What_makes_physics_difficult.
 - [2] K. R. Koedinger, P. F. Carvalho, R. Liu, and E. A. McLaughlin, An astonishing regularity in student learning rate, *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2221311120 (2023).
 - [3] V. Gjerde, B. Holst, and S. D. Kolstø, Retrieval practice of a hierarchical principle structure in university introductory physics: Making stronger students, *Phys. Rev. Phys. Educ. Res.* **16**, 013103 (2020).
 - [4] J. D. Karpicke, Retrieval-based learning: Active retrieval promotes meaningful learning, *Curr. Dir. Psychol. Sci.* **21**, 157 (2012).
 - [5] B. Pastotter and K. H. Bauml, Retrieval practice enhances new learning: The forward effect of testing, *Front. Psychol.* **5**, 286 (2014).
 - [6] T. Zu, J. Munsell, and N. S. Rebello, Comparing retrieval-based practice and peer instruction in physics learning, *Phys. Rev. Phys. Educ. Res.* **15**, 010105 (2019).
 - [7] N. Kornell, M. J. Hays, and R. A. Bjork, Unsuccessful retrieval attempts enhance subsequent learning, *J. Exp. Psychol. Learn. Mem. Cogn.* **35**, 989 (2009).
 - [8] V. X. Yan, L. G. Eglington, and M. A. Garcia, Learning better, learning more: The benefits of expanded retrieval practice, *J. Appl. Res. Mem. Cognit.* **9**, 204 (2020).
 - [9] C. A. Rowland, The effect of testing versus restudy on retention: A meta-analytic review of the testing effect, *Psychol. Bull.* **140**, 1432 (2014).

- [10] C. L. Yang, L. Luo, M. A. Vadillo, R. Yu, and D. R. Shanks, Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review, *Psychol. Bull.* **147**, 399 (2021).
- [11] H. L. Roediger and J. D. Karpicke, The power of testing memory: Basic research and implications for educational practice, *Perspect. Psychol. Sci.* **1**, 181 (2006).
- [12] T. C. Rickard and S. C. Pan, A dual memory theory of the testing effect, *Psychon. Bull. Rev.* **25**, 847 (2018).
- [13] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham, Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology, *Psychol. Sci. Publ. Interest* **14**, 4 (2013).
- [14] P. K. Agarwal, P. M. Bain, and R. W. Chamberlain, The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist, *Educ. Psychol. Rev.* **24**, 437 (2012).
- [15] V. Gjerde and H. Paulsen, Problem solving in basic physics: Effective self-explanations based on four elements with support from retrieval practice, *Phys. Rev. Phys. Educ. Res.* **18**, 010136 (2022).
- [16] J. D. Karpicke and J. R. Blunt, Retrieval practice produces more learning than elaborative studying with concept mapping, *Science* **331**, 772 (2011).
- [17] J. R. Anderson, L. M. Reder, and C. Lebiere, Working memory: Activation limitations on retrieval, *Cogn. Psychol.* **30**, 221 (1996).
- [18] M. C. Lovett, L. Z. Daily, and L. M. Reder, A source activation theory of working memory: Cross-task prediction of performance in ACT-R, *Cognit. Syst. Res.* **1**, 99 (2000).
- [19] Z. Shen, V. Popov, A. B. Delahay, and L. M. Reder, Item strength affects working memory capacity, *Mem. Cogn.* **46**, 204 (2018).
- [20] V. Gjerde, B. Holst, and S. D. Kolstø, Integrating effective learning strategies in basic physics lectures: A thematic analysis, *Phys. Rev. Phys. Educ. Res.* **17**, 010124 (2021).
- [21] V. Gjerde and S. Hagane, Enhancing peer instruction in physics: Understanding cognitive processes and refining rules, *Phys. Rev. Phys. Educ. Res.* **20**, 010134 (2024).
- [22] D. S. Lindquist, B. E. Sparrow, and J. M. Lindquist, Spaced recall reduces forgetting of fundamental mathematical concepts in a post high school precalculus course, *Instr. Sci.* **52**, 859 (2024).
- [23] P. I. Pavlik and J. R. Anderson, Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect, *Cogn. Sci.* **29**, 559 (2005).
- [24] J. R. Anderson, J. M. Fincham, and S. Douglass, Practice and retention: A unifying analysis, *J. Exp. Psychol. Learn. Mem. Cogn.* **25**, 1120 (1999).
- [25] N. Entwistle and D. Entwistle, Preparing for examinations: The interplay of memorising and understanding, and the development of knowledge objects, *Higher Educ. Res. Dev.* **22**, 19 (2003).
- [26] L. M. Reder, X. L. Liu, A. Keinath, and V. Popov, Building knowledge requires bricks, not sand: The critical role of familiar constituents in learning, *Psychon. Bull. Rev.* **23**, 271 (2016).
- [27] B. Bellana, R. Mansour, N. Ladyka-Wojcik, C. L. Grady, and M. Moscovitch, The influence of prior knowledge on the formation of detailed and durable memories, *J. Mem. Lang.* **121**, 104264 (2021).
- [28] N. L. Varga, L. Cronin-Golomb, and P. J. Bauer, Self-derivation of new knowledge through memory integration varies as a function of prior knowledge, *Memory* **30**, 971 (2022).
- [29] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.21.010119> for the factual tests with test analyses, principle structures, retrieval tests, and interview guide.
- [30] J. Larkin, Enriching formal knowledge: A model for learning to solve textbook physics problems, in *Cognitive Skills and Their Acquisition*, edited by J. R. Anderson (Erlbaum, Hillsdale, NJ, 1981).
- [31] M. T. H. Chi *et al.*, Self-explanations—How students study and use examples in learning to solve problems, *Cogn. Sci.* **13**, 145 (1989).
- [32] M. T. H. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices, *Cogn. Sci.* **5**, 121 (1981).
- [33] J. P. Mestre, R. J. Dufresne, W. J. Gerace, P. T. Hardiman, and J. S. Touger, Promoting skilled problem-solving behavior among beginning physics students, *J. Res. Sci. Teach.* **30**, 303 (1993).
- [34] H. S. Lee and D. Ahn, Testing prepares students to learn better: The forward effect of testing in category learning, *J. Educ. Psychol.* **110**, 203 (2018).
- [35] K. T. Wissman, K. A. Rawson, and M. A. Pyc, The interim test effect: Testing prior material can facilitate the learning of new material, *Psychon. Bull. Rev.* **18**, 1140 (2011).
- [36] J. C. K. Chan, C. A. Meissner, and S. D. Davis, Retrieval potentiates new learning: A theoretical and meta-analytic review, *Psychon. Bull. Rev.* **144**, 1111 (2018).
- [37] B. Pastötter, S. Schicker, J. Niedernhuber, and K.-H. T. Bäuml, Retrieval during learning facilitates subsequent memory encoding, *J. Exp. Psychol. Learn. Mem. Cogn.* **37**, 287 (2011).
- [38] B. Pastötter, M. Engel, and C. Frings, The forward effect of testing: Behavioral evidence for the reset-of-encoding hypothesis using serial position analysis, *Front. Psychol.* **9**, 1197 (2018).
- [39] A. Robey, The benefits of testing: Individual differences based on student factors, *J. Mem. Lang.* **108**, 104029 (2019).
- [40] P. K. Agarwal *et al.*, Benefits from retrieval practice are greater for students with lower working memory capacity, *Memory* **25**, 764 (2017).
- [41] S. K. Carpenter, S. C. Pan, and A. C. Butler, The science of effective learning with spacing and retrieval practice, *Nat. Rev. Psychol.* **1**, 496 (2022).
- [42] A. Latimier, H. Peyre, and F. Ramus, A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention, *Educ. Psychol. Rev.* **33**, 959 (2020).
- [43] A. C. Butler, Repeated testing produces superior transfer of learning relative to repeated studying, *J. Exp. Psychol. Learn. Mem. Cogn.* **36**, 1118 (2010).

- [44] J. R. Anderson, D. Bothell, C. Lebiere, and M. Matessa, An integrated theory of list memory, *J. Mem. Lang.* **38**, 341 (1998).
- [45] P. L. Pirolli and J. R. Anderson, The role of practice in fact retrieval, *J. Exp. Psychol. Learn. Mem. Cogn.* **11**, 136 (1985).
- [46] K. A. Rawson and J. Dunlosky, Successive relearning: An underexplored but potent technique for obtaining and maintaining knowledge, *Curr. Dir. Psychol. Sci.* **31**, 362 (2022).
- [47] D. Rohrer, K. Taylor, and B. Sholar, Tests enhance the transfer of learning, *J. Exp. Psychol. Learn. Mem. Cogn.* **36**, 233 (2010).
- [48] B. Kowialiewski, B. Lemaire, and S. Portrat, How does semantic knowledge impact working memory maintenance? Computational and behavioral investigations, *J. Mem. Lang.* **117**, 104208 (2021).
- [49] G. Jones, L. V. Justice, F. Cabiddu, B. J. Lee, L.-S. Iao, N. Harrison, and B. Macken, Does short-term memory develop?, *Cognition* **198**, 104200 (2020).
- [50] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, An integrated theory of the mind, *Psychol. Rev.* **111**, 1036 (2004).
- [51] J. R. Anderson and C. Lebiere, *The Atomic Components of Thought* (Lawrence Erlbaum Associates, Mahwah, NJ, 1998), xi, p. 490.
- [52] J. W. Antony, C. S. Ferreira, K. A. Norman, and M. Wimber, Retrieval as a fast route to memory consolidation, *Trends Cognit. Sci.* **21**, 573 (2017).
- [53] S. C. Pan and T. C. Rickard, Transfer of test-enhanced learning: Meta-analytic review and synthesis, *Psychol. Bull.* **144**, 710 (2018).
- [54] A. Renkl, Learning from worked-out examples: A study on individual differences, *Cogn. Sci.* **21**, 1 (1997).
- [55] M. Racsmány, Á. Szöllösi, and M. Marián, Reversing the testing effect by feedback is a matter of performance criterion at practice, *Mem. Cogn.* **48**, 1161 (2020).
- [56] J. Roelle and M. Nückles, Generative learning versus retrieval practice in learning from text: The cohesion and elaboration of the text matters, *J. Educ. Psychol.* **111**, 1341 (2019).
- [57] V. Gjerde, Short introductory videos for elaborative encoding of physics principles, *Nordic J. STEM Educ.* (to be published).
- [58] W. J. Hopper and D. E. Huber, Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects, *J. Mem. Lang.* **102**, 1 (2018).
- [59] W. J. Hopper and D. E. Huber, Testing the primary and convergent retrieval model of recall: Recall practice produces faster recall success but also faster recall failure, *Mem. Cogn.* **47**, 816 (2019).
- [60] J. Roelle, L. Froese, R. Krebs, N. Obergassel, and J. Waldeyer, Sequence matters! Retrieval practice before generative learning is more effective than the reverse order, *Learn. Instr.* **80**, 101634 (2022).
- [61] N. Kornell, P. J. Klein, and K. A. Rawson, Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter, *J. Exp. Psychol. Learn. Mem. Cogn.* **41**, 283 (2015).
- [62] A. D. Endress, Memory and proactive interference for spatially distributed items, *Mem. Cogn.* **50**, 782 (2022).
- [63] R. Johansson and M. Johansson, Gaze position regulates memory accessibility during competitive memory retrieval, *Cognition* **197**, 104169 (2020).
- [64] S. R. Yousif, M. D. Rosenberg, and F. C. Keil, Using space to remember: Short-term spatial structure spontaneously improves working memory, *Cognition* **214**, 104748 (2021).
- [65] S. Darling and J. Havelka, Visuospatial bootstrapping: Evidence for binding of verbal and spatial information in working memory, *Q. J. Exp. Psychol.* **63**, 239 (2010).
- [66] R. Y. Hu and R. A. Jacobs, Semantic influence on visual working memory of object identity and location, *Cognition* **217**, 104891 (2021).
- [67] C. Twomey and M. Kroneisen, The effectiveness of the loci method as a mnemonic device: Meta-analysis, *Q. J. Exp. Psychol.* **74**, 1317 (2021).
- [68] R. A. Bjork, J. Dunlosky, and N. Kornell, Self-regulated learning: Beliefs, techniques, and illusions, *Annu. Rev. Psychol.* **64**, 417 (2013).
- [69] S. K. Carpenter, A. E. Witherby, and S. K. Tauber, On students' (mis)judgments of learning and teaching effectiveness, *J. Appl. Res. Mem. Cognit.* **9**, 137 (2020).
- [70] A. Kirk-Johnson, B. M. Galla, and S. H. Fraundorf, Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice, *Cogn. Psychol.* **115**, 101237 (2019).
- [71] S. K. Carpenter, Encouraging students to use retrieval practice: A review of emerging research from five types of interventions, *Educ. Psychol. Rev.* **35**, 96 (2023).
- [72] M. A. McDaniel and G. O. Einstein, Training learning strategies to promote self-regulation and transfer: The knowledge, belief, commitment, and planning framework, *Perspect. Psychol. Sci.* **15**, 1363 (2020).
- [73] L. S. Wang, K. Muenks, and V. X. Yan, Interventions to promote retrieval practice: Strategy knowledge predicts intent, but perceived cost predicts usage, *J. Educ. Psychol.* **115**, 1070 (2023).
- [74] S. D. Rea *et al.*, Students can (mostly) recognize effective learning, so why do they not do it?, *J. Intell.* **10**, 127 (2022).
- [75] K. Barron and C. Hulleman, Expectancy-value-cost model of motivation in *International Encyclopedia of Social & Behavioral Sciences* (Oxford, Oxford, 2014), pp. 503–509.
- [76] J. S. Eccles and A. Wigfield, Motivational beliefs, values, and goals, *Annu. Rev. Psychol.* **53**, 109 (2002).
- [77] L. A. Clark and D. Watson, Constructing validity: Basic issues in objective scale development, *Psychol. Assess.* **7**, 309 (1995).
- [78] P. Kline, *The Handbook of Psychological Testing* (Routledge, 2000).
- [79] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory*, 3rd ed., *McGraw-Hill Series in Psychology* (McGraw-Hill, New York, 1994), xxiv, p. 752.
- [80] S. Slater, The development and validation of the Test Of Astronomy Standards (TOAST), *J. Astron. Earth Sci. Educ.* **1**, 1 (2015).
- [81] D. Hestenes and M. Wells, A mechanics baseline test, *Phys. Teach.* **30**, 159 (1992).

- [82] B. Brown and C. Singh, Development and validation of a conceptual survey instrument to evaluate students' understanding of thermodynamics, *Phys. Rev. Phys. Educ. Res.* **17**, 010104 (2021).
- [83] S. Kvale, *Interviews: An Introduction to Qualitative Research Interviewing* (Sage Publications, Thousand Oaks, CA, 1996), xvii, p. 326.
- [84] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2022.
- [85] A. P. Field, J. Miles, and Z. Field, *Discovering statistics using R* (Sage, Thousand Oaks, CA, London, 2012), xxxiv, p. 957.
- [86] J. Cohen, The cost of dichotomization, *Appl. Psychol. Meas.* **7**, 249 (1983).
- [87] V. Braun and V. Clarke, Using thematic analysis in psychology, *Qual. Res. Psychol.* **3**, 77 (2006).
- [88] QSR International, NVivo 12 Pro, QSR International, 2018.
- [89] W. A. Firestone, Alternative arguments for generalizing from data as applied to qualitative research, *Educ. Res. J.* **22**, 16 (1993).
- [90] L. J. Cronbach, Beyond the two disciplines of scientific psychology, *Am. Psychol.* **30**, 116 (1975).
- [91] Y. S. Lincoln and E. G. Guba, *Naturalistic Inquiry* (Sage Publications, Beverly Hills, CA, 1985), p. 416.
- [92] G. E. Gignac and E. T. Szodorai, Effect size guidelines for individual differences researchers, *Pers. Individ. Diff.* **102**, 74 (2016).
- [93] J. F. Hemphill, Interpreting the magnitudes of correlation coefficients, *Am. Psychol.* **58**, 78 (2003).
- [94] S. L. Braver, F. J. Thoemmes, and R. Rosenthal, Continuously cumulating meta-analysis and replicability, *Perspect. Psychol. Sci.* **9**, 333 (2014).
- [95] R. Rosenthal, Replication in behavioral research, *J. Soc. Behav. Pers.* **5**, 1 (1990).
- [96] C. G. Sibley, Utilities for examining simple meta-analytic averages [computer software]. University of Auckland, 2008.
- [97] S. Balduzzi, G. Rücker, and G. Schwarzer, How to perform a meta-analysis with R: A practical tutorial, *Evidence-Based Ment. Health* **22**, 153 (2019).