Problem solving in basic physics: Effective self-explanations based on four elements with support from retrieval practice

Vegard Gjerde, Vegard Havre Paulsen, Bodil Holst, and Stein Dankert Kolstø
Department of Physics and Technology, University of Bergen, 5007 Bergen, Norway

(Received 20 December 2020; accepted 7 March 2022; published 2 May 2022)

Self-explanation, a learning strategy where students explain to themselves the steps taken in a worked example, is an effective learning strategy in early cognitive skill acquisition. However, many physics students produce self-explanations of low quality. There is also a lack of guidelines for what students should seek to explain when studying worked examples. Therefore, the overarching purpose of this article is to investigate how we can improve students' self-explanations of worked examples. We pursue the following two general research questions: (1) What knowledge representations should students seek while self-explaining worked examples to maximize their learning? (2) Can retrieval practice of physics principles and their conditions of application potentiate students' learning from self-explaining worked examples? In two studies ($n = 18$ and $N = 54$), we qualitatively categorized and quantified the students' written self-explanations. Our results indicate that to produce useful knowledge, self-explanations of the physics model in worked examples should explain what principle is used, how the principle is set up, and how the conditions of application are met for the principle, while explanations of the mathematical procedures should contain action descriptions, goals, and conditions ($r = 0.30$–0.50). Through a quasiexperimental ($N = 57$) and an experimental ($N = 54$) test, we found evidence that retrieval practice of physics principles and their conditions of application before self-explanation can have a medium-sized effect on post-test problem-solving scores and that it can increase the quality of students' self-explanations. Using retrieval practice to potentiate learning from more complex learning strategies is a novel and promising approach to improve physics students' learning.

DOI: 10.1103/PhysRevPhysEducRes.18.010136

I. INTRODUCTION

Learning a new cognitive skill is difficult. Basic physics is considered particularly difficult by students required to take it as a part of their degree [1,2]. It has been shown that a mix of worked examples and problem solving speeds up cognitive skill acquisition in mathematical domains and physics [3–7] and reduces the effort required from novice students [8]. The strong evidence for examples being essential in learning cognitive skills has triggered much research into how we learn from examples, how to structure the examples and instruction, and what the learning mechanisms are. *Self-explanation*—a learning strategy where students explain to themselves the steps taken in a worked example—has emerged as an effective strategy from this research [7,9–11]. However, the effects found from research on self-explanation are still diverging [12–14], which signals a persistent need for knowledge about how and when self-explanation is effective.

Research has been done on what students actually self-explain and how this affects later performance [9,15]. In a landmark paper on self-explanation, Chi *et al.* [9] found that the best self-explainers refine and expand conditions for actions, explicate and infer additional consequences for actions, impose goals for and purpose of actions, and give meaning to quantitative expressions. Renkl [15] further explored individual differences in self-explanation of worked examples in a statistics context and found that there were two effective styles of study: Principle-based explainers typically explicated goals and principles in their explanations ($r = 0.38$ with post-test); anticipative reasoners engaged in anticipative calculation of solution steps before checking the solution ($r = 0.49$ with post-test). Research has also been done on issues such as how the worked examples should be structured [16–18]; whether eliciting self-explanation is effective [12,19,20]; and how instruction with worked examples should be structured [3,14,21], factors students have little control over. Despite this extensive body of work, there is still a lack in the literature of prescriptions for what the students themselves should do when self-explaining worked examples. Specifically, what are the *knowledge representations* (cognitive representations of the environment, i.e., of problem solutions) students should seek to build when engaging in self-explanation, and how does this depend on circumstance and type of test? The answers to these questions can help educators facilitate effective learning. Furthermore, improving students' strategic knowledge for studying worked examples can potentiate their knowledge acquisition [22]. To achieve this, students need explicit instruction in when, why, and how to self-explain worked examples and ample opportunities to put the instruction into practice [23].

This ties in with other research [24–26], which suggests that the knowledge generated during self-explanation is dependent on the student's access to relevant prior knowledge. Research also shows that having stronger memories for basic knowledge improves working memory capacity and makes it easier to acquire more complex knowledge [27–30]. There is also research suggesting that distributed retrieval practice of physics principles may affect exam results [31]. These studies inspired us to investigate whether retrieval practice—a proven learning strategy for improving access to knowledge [32,33]—can potentiate the learning effects from self-explanation.

The overarching purpose of this article is to investigate how we can improve students' self-explanations of worked examples in physics so that students learn more and learn deeper. We pursue two general research questions: (1) What knowledge representations should students seek while self-explaining worked examples to maximize their learning? (2) Can retrieval practice of physics principles and their conditions of application potentiate students' learning from self-explaining worked examples? We addressed these questions through two studies.

Study 1 was an exploratory analysis of data from two prior experiments—a self-explanation experiment and a retrieval practice experiment—done one week apart towards the end of the semester. The first research question for study 1 was whether retrieval practice of relevant physics principles affected post-test results; see Sec. I. C for the theoretical background. The second research question for study 1 was whether principles, goals, and conditions of application—see Secs. I. B. 2 and I. B. 3 for detailed explanations—are important elements to include in self-explanations as measured by their predictiveness for post-test results.

Study 2 was done to experimentally test the effect of retrieval practice on post-test performance (randomized controlled trial) and see whether it affected how students self-explained and solved problems (hypotheses 1–3 in study 2, see Sec. III. A. 1). We also further investigated how the different categories of self-explanations (coded and categorized according to which self-explanation elements were included, see Secs. II. A. 4 and III. A. 6) predicted post-test performance. Study 2 differed from study 1 in that it included more participants (from the subsequent year's cohort) doing self-explanation, involved new physics concepts, and that the explanations of the *physics model*—consisting of physics principles that describe the situation or predict changes in the system—were separated from the explanations of the mathematical procedures (hypotheses 4–5 in study 2, see Sec. III. A. 2). See Supplemental Material [34] for examples of separating the physics model from mathematical procedures.

This article draws on empirical results from the literature, but we base most of our hypotheses and analysis of the written self-explanations on Anderson's adaptive control of thought-rational, ACT-R [35–37]. A short description of this theory is therefore included in Sec. I. A.

A. The ACT-R theory

ACT-R models the human cognitive architecture using basic psychological research. It provides precise predictions for the performance and learning of both declarative and procedural knowledge. More recent versions of ACT-R have a strong focus on learning cognitive skills from examples [35,38,39]. Some researchers have already based much of their analyses of the self-explanation effect on ACT-R, e.g., [40]. ACT-R has also been used to investigate the effects of retrieval practice on memory and fluency [32,41].

According to ACT-R, the two types of knowledge are *declarative knowledge* and *procedural knowledge* [37]. The unit of declarative knowledge is the *chunk* (e.g., $A \times B = C$). The unit of procedural knowledge, and hence cognitive skill, is the *production rule* (e.g., IF the goal is to multiply $A \times B$, and there is a fact stating that $A \times B = C$, THEN set the answer to C). Both units are very small [37] due to our inherent working memory limitations, but larger declarative units can be formed through hierarchical relationships [42]. Declarative knowledge is flexible, accessible, and verbalizable, but it is slow for solving problems except when the solution is directly retrievable [39]. Procedural knowledge is specialized, efficient, and nonverbalizable. Self-explanation of worked examples can produce declarative knowledge of procedural actions. However, this is not the same as procedural knowledge. It must be retrieved and interpreted during problem solving, slowly being turned into procedural knowledge with problem-solving practice [37].

We want to figure out what exactly the students should try to learn when studying worked examples, or, in other words, what types of explanations we should teach them [43]. Some initial ideas come from the structure of the production rule in ACT-R. It consists of a goal, an action, and conditions, and it usually entails retrieval of one or more declarative knowledge chunks based on the conditions [36,37]. In physics, these chunks will often be physics principles of various forms. Hence, we tentatively propose four basic elements for inclusion in self-explanation (SE elements) of a solution step: Actions, goals, principles, and conditions. The rest of this introduction elaborates on the theoretical and empirical basis for each SE element, followed by a discussion of retrieval practice as a promising adjunct to self-explanation.

B. Four elemental components of self-explanations

1. SE-Element 1: Actions

We define an *action* as some procedural step taken in a worked example. Procedural skill is intricately tied to actions, both cognitive (e.g., calculating) and motoric (e.g., writing the answer). Worked examples mostly consist of action steps, with some examples also including instructional explanations. To learn declarative knowledge of the procedural actions, one must necessarily encode these actions (e.g., that mass has been removed on both sides of the equation) because further explanation of a solution step centers around the action [9]. Descriptions (of actions) are better for learning details, while explanations are better for discovering broad patterns [24], potentially synergizing during self-explanation. Mere descriptions of actions are the shallowest explanations and lack direction and context if not attached to goals and conditions.

2. SE-Element 2: Goals

The goal is a central part of procedural skill in ACT-R. Production rules—actions we subconsciously choose to perform—are selected based on the current goal state [37]. Goals provide direction for problem-solving behavior by directing attention to the components of the goal. For example, if the goal is to calculate the net force on an object, the given acceleration and mass of the object can be components in the goal.

The *goal hierarchy* is the glue that ties together knowledge components and enables coherent behavior, providing purpose, organization, and seriality to cognition [37]. Goal hierarchies are essential when the problem involves several steps, and the learner must work backward from the goal [44]. The main goal of the problem, usually finding an unknown, gets progressively decomposed into subgoals until the student reaches an immediately achievable goal [42]. Experts in physics tend to have a more forward working character in their problem solving [45]. This might be because they can use previously constructed goal hierarchies based on their problem categorization. However, when experts encounter difficult problems, they too have to revert to backward working through a slow construction of hierarchical goal structures [44].

In sum, goals seem to be essential for learning to solve problems through SE. There is also empirical support for including goals in self-explanations. The categories of good explanations in Chi *et al.* [9] included "impose a goal or purpose for an action." Two of the other categories, "explicate or infer additional consequences of an action" and "give meaning to a set of quantitative expressions" [9], also seem to be goal related when considering the examples given. In Renkl [15], one of the self-explanation categories most highly correlated with performance was goal-operator combinations ($r = 0.37$ with post-test). Although explicating goals seems to be most important for explaining mathematical procedures, explicating goals is also relevant for understanding the physics modeling part of the solutions. The physics model must describe the situation accurately and be sufficient for reaching the main goal of the problem.

3. SE-Element 3: Principles and SE-element 4: Conditions

We present these two SE elements together because they are intimately connected. Many physics principles have conditions for application (e.g., the condition for conservation of momentum is that the sum of external forces on the system is zero). It is probably not enough for the students to recognize the physics principles used for solving worked examples; they must also learn to recognize when they apply.

In ACT-R, the typical sequence for a production rule is goal match, then declarative retrieval, followed by a goal transformation [37]. The planning stage of problem solving is especially characterized by large amounts of retrieval activity [46], which when solving physics problems often means retrieving physics principles. A large part of solving problems in physics is to model the problem by finding physics principles that apply to the problem and are sufficient for solving it [47]. Hence, the explication of principles and their conditions of application during self-explanation should improve the physics modeling capabilities of our students. Since physics modeling is arguably the most difficult part of solving a physics problem, the explication of physics principles and their application conditions should substantially affect our students' problem-solving ability (hypothesis 4 in study 2). We also believe that this should substantially affect our students' conceptual understanding (hypothesis 5 in study 2). Conceptual understanding is usually tested with multiple-choice conceptual problems, which are often designed to remove the need for mathematical procedures. Hence, the students must find the solution through recognizing which physics principles apply and then using the principles to analyze the situation qualitatively.

There are also strong empirical reasons for including principles and conditions in self-explanations. Chi *et al.* [9] found that good self-explainers tended to relate example statements to physics principles and explicate conditions for actions in general. Renkl [15] also found that principle-based explanations were effective ($r = 0.38$ with post-test). However, he did not code it as such if the student failed also to explicate the principle's conditions of application. Berthold and Renkl [48] found that principle-based explanations were effective for developing conceptual knowledge ($r = 0.39$ with post-test). We believe that students can also gain from explicating principles without explicating their conditions for application, regardless of whether they are unable to explicate them or choose not to (hypotheses 4 and 5 in study 2). In addition, we believe that students can gain from explicating conditions for procedural actions, without explicit reference to principles (hypothesis 4 in study 2), in line with the results of Chi *et al.* [9].

C. Retrieval practice

Little research has been done on retrieval practice in physics, with reference [49] as a rare exception apart from our research. However, the retrieval-based learning of lecture content in the article by Zu, Munsell, and Rebello [49]—where the students went through a molar treatment package including categorizing problems, recalling definitions, and problem solving based on the preceding lecture—is different from our retrieval practice of physics principles and their conditions. Gjerde, Holst, and Kolstø [31] found that retrieval practice of physics principles and their conditions of application had a large effect on the students' performance on a declarative factual test and that participation in structured and regular retrieval practice sessions correlated with better exam scores. They have also investigated students' experiences and reflections regarding retrieval practice of a hierarchical principle structure [50].

We suggest that retrieval practice of physics principles and their conditions of application can facilitate performance on problem solving and conceptual tests either (i) indirectly through improved self-explanation quality, (ii) indirectly through potentiating other learning strategies, or (iii) by directly affecting problem-solving performance.

1. Retrieval practice can improve self-explanation quality

Researchers agree that self-explanations are based on prior knowledge [24,25]. Wong *et al.* [26] found that the self-explanation effect was mediated by knowledge generation, which in turn was predicted by knowledge access and prior knowledge. In Sec. I. B. 3, we suggest that physics principles and their conditions of application are the critical knowledge components students must access to produce high-quality self-explanations. However, self-explanation is especially effective in the early phase of skill acquisition, when many students cannot access their memories for principles and conditions [31]. Therefore, retrieval practice of physics principles and their conditions of application can be a simple intervention to facilitate high-quality self-explanations through increased access to the relevant knowledge (hypothesis 1 and 2 in study 2). We are not aware of any research in any educational domain that has used retrieval practice to improve access to specific knowledge components for self-explaining.

In ACT-R, the probability of retrieval (knowledge access) is based on the activation of the chunk [36], which is the sum of the chunk's base strength and contextual cues. To simplify: the base strength of memories is built by practice and retrieval, while the contextual cues are built through generative learning strategies (i.e., elaboration). Activation from contextual cues—the summed product of attention on the components in the goal times their associative strength with the chunk—is weak in the early stages of skill acquisitions. There are also strong limitations on how much activation we can get from contextual cues [51,52]. On the other hand, the base strength of the chunks—a function of recency and frequency—is easier to build. The most effective method for improving the base strength of memories—improving access to knowledge—is retrieval practice [53–58].

It is also possible to search for principles in the textbook while trying to explain the equations used in worked examples. However, it is inefficient, and some students will probably not even try. On the other hand, the students that retrieve principles during self-explanation get the added memory-strengthening benefits to the generative learning benefits from self-explaining [40,59].

2. Retrieval practice can potentiate all learning strategies

Physics principles and their conditions are also vital knowledge components when using other learning strategies. Therefore, retrieval practice of physics principles and their application conditions may be an intervention that makes all the other learning strategies students use more effective [50]. The increased base strength makes the memories less context dependent [55] and may shift the students' studying from a shallow focus to deeper learning strategies [22,60]. The increased base strength of memories for physics principles may also improve students' working memory capacity for physics [27,28,30]. Moreover, research suggests that it is easier to bind strong memories (high base strength) into more complex knowledge structures [29].

3. Retrieval practice can directly affect problem-solving and conceptual performance

Retrieval practice of principles and their conditions can directly affect post-test performance in two ways. First, it can directly benefit problem solving by improving the planning phase [46], where students model the situation with physics principles (hypothesis 3 in study 2). Additionally, retrieving the principles during problem solving further strengthens the retrieved memories [40] for future problem solving. Second, as mentioned in the previous section, it can reduce the pressure on working memory, which is especially important when the problems are complex and/or when the students have a smaller working memory capacity [28–30].

II. STUDY 1

A. Method

Study 1 was an exploratory analysis of data from two prior and separate experiments, both randomized controlled trials. These experiments were conducted one week apart towards the end of the semester in an introductory mechanics course. The first of these prior experiments investigated retrieval practice of a hierarchical principle structure, reported in Gjerde *et al.* [31]. The second experiment investigated self-explanation of worked examples versus problem solving (with solutions available) or combining the two strategies. The two experiments were also designed so that we could connect and explore the data from the participants who participated in both. Exploring the data from these two experiments, we tried to answer the following two research questions: (1) Does retrieval practice of relevant physics principles (in the first experiment) affect post-test results (in the second experiment)? (2) Are principles, goals, and application conditions important elements to include in self-explanations as measured by their predictiveness for post-test results? We used the combined data from these two experiments to quasiexperimentally investigate question 1. In addition, we used the written self-explanations of a subset of the participants in the second experiment for answering question 2.

1. Participants

The Norwegian Centre for Research Data approved the study. We collected no person-identifying information. The participants provided informed anonymous consent. The self-explanation and outcome data we explored were from the second (self-explanation) experiment, which was conducted during regular, voluntary lecture hours. The number of participants was limited to how many showed up. The participants were 57 undergraduate students from an introductory mechanics course at a large Norwegian university. See study 2 for descriptions of the participants in the subsequent year's cohort, as study 1 was completely anonymous. The students in the course have very similar educational backgrounds where almost everyone has completed two years of physics at the high school level before admission. All participants had equal chances of winning one gift certificate of 2000 NOK (~250 USD) at the end of the experiment.

2. Procedure and materials

Study 1 was an exploratory analysis of relevant data from two prior experiments: A self-explanation experiment where students provided written self-explanations and were tested on conceptual knowledge and problem solving; and a retrieval practice experiment, performed one week prior, where half the students did retrieval practice of important physics principles. Each participant generated a unique participant code in both experiments, which enabled a quasiexperimental test of the effects of retrieval practice on the post-tests in the self-explanation experiment. All the concepts in the course curriculum had been covered in lectures before the two experiments took place.

The retrieval practice experiment.—A week before the self-explanation experiment, a retrieval practice experiment ($N = 80$) was conducted where half of the participants did retrieval practice of important mechanics principles (Newton's laws, conservation of energy, and conservation of linear and angular momentum) and other physics equations (e.g., the definitions of centripetal acceleration and gravitational potential energy) for 70 min, while the other half studied solutions to physics problems for 70 min. The students were randomized into the two conditions. Of the 80 students in the retrieval practice experiment, 24 retrievers and 20 nonretrievers also participated in the self-explanation study one week later. In addition, 13 more students participated in the self-explanation experiment and were added to the group of nonretrievers for the quasiexperimental analysis. Hence, there were 24 retrievers and 33 nonretrievers in the (self-explanation) experiment. The 13 extra students were from the same course, and we expect that the main systematic difference between them and the other participants was that they did not participate in the retrieval practice experiment. We discuss the possibility of other systematic differences in the discussion in Sec. II. C. 1. The data from students who only participated in the retrieval practice experiment were discarded.

The self-explanation experiment.—The overall design of the self-explanation experiment was as follows: 30 min of practice → a 45-min intervening task → a 20-min problem-solving test. The 57 students were randomly assigned to three groups in a one-factorial between-subject design. 21 students were in the problem-solving group (PS), 18 were in the self-explanation group (SE), and 18 were in the combined self-explanation and problem-solving group (SE + PS). During the 30-min practice phase of the self-explanation experiment, the PS group solved problems, the SE group self-explained problems, and the SE + PS group alternated between explaining and solving problems. We did not analyze the differences between these three factors in the current exploratory study because it was not relevant to our research questions. Then followed a 45-min intervening task phase, which consisted of the full mechanics baseline test (MBT) [61]. The final step was a 20-min post-test phase, which consisted of two physics problems. The structure of the experiment was constrained by the need to fit it into a normal double lecture of 90 min.

The practice worksheets contained the same six physics problems for all the participants. Solving the first three problems required using Newton's second law with centripetal acceleration and conservation of mechanical energy. The last three problems required either Newton's second law with centripetal acceleration *or* conservation of mechanical energy. The first two and last three problems were identical to the practice problems used in Badeau *et al.* [7]. The third problem was identical in solution structure to the first two problems but differed in surface details; see Supplemental Material [34] for the practice problems. The problems were presented with their solutions and a box at the bottom of the page for either self-explaining the solution or solving the problem. The solutions had no instructional explanations because instructional explanations can potentially suppress constructive self-explanation activity [62], especially when self-explanations are prompted [63]. The worksheet for the SE group had one problem on each page, with general prompts for self-explanations ("Explain the solution as well as you can. Use the box under each problem. Try to include: 1. Which steps in the solution have the same goal. What is the goal? 2. Which physics principles underlie the solution. 3. How the problem statement indicates that these physics principles can be used.").

3. Measures

To enable exploration of the effects on conceptual knowledge from different types of self-explanations and retrieval practice, MBT items 5, 8, and 10 were selected after the fact as a post-test measure of conceptual understanding. The four authors, whom all have completed more than five years of physics education and have substantial experience with instruction in mechanics, agreed that these three items targeted the relevant concepts for the current study. Since the items were selected after the fact, we also asked three other physics instructors to rate the relevance (from 0: no relevance to 3: very relevant) of these three items together with ten randomly chosen items from the mechanics baseline test. Of the 13 items, items 5, 8, and 10 received the top three relevance ratings, only receiving "2: relevant" and "3: very relevant." Cronbach's alpha for these three items was 0.74. Cronbach's alpha is a measure of the interrelatedness or internal consistency of test items, giving something like a lower bound on the reliability of a measure [64]. Many science education researchers cite values of Cronbach's alpha above 0.70 as acceptable [65]. However, others argue that there is no objective value at which it suddenly becomes acceptable. Measures with lower alpha can be useful [66] and interpretable [67], although a high value is usually preferable. Moreover, Cronbach's alpha is likely to underestimate the reliability of measures that fail to meet the assumption of unidimensionality [67], and measures of conceptual knowledge are probably seldom unidimensional.

The post-test consisted of two problems with the same solution structure as the practice problems. The first problem was identical to the post-test problem used in Badeau *et al.* [7] and had similar surface features to most of the practice problems. The second problem had substantially different surface features but identical solution structure; see Supplemental Material [34] for the post-test problems. The post-test was scored according to the same rubric used by Badeau *et al.* [7] for both problems, with +1 point for each of the nine observable actions in Table I. B. H. (the lecturer) and V. G. scored all the problem-solving post-tests. Interrater agreement was 90% on the individual scoring points. Cohen's unweighted kappa was 0.76, a substantial interrater reliability [68], especially given that there were only two codes and they were not equiprobable [69]. Differences were discussed until agreement. Cronbach's alpha for the problem-solving post-test was 0.96. There was a ceiling effect on the post-test, which tends to attenuate correlational relationships [70].

TABLE I. Scoring rubric for the problem-solving post-test [7].

Observable actions in post-test
+1 Energy conservation recognition
+1 Centripetal acceleration recognition
+1 Correct application of Newton's second law
+1 Identifying that normal force equal zero
+1 Correct initial potential energy
+1 Correct final potential energy
+1 Include final kinetic energy
+1 Substitute correct final velocity
+1 Correct final answer

4. Coding self-explanation categories

The coding of the written self-explanations was organized around the solution step(s) the students were referring to. Each of the explanations was categorized by which of the four SE-elements they included (actions, goals, principles, and conditions of application) and were mutually exclusive. All the self-explanations explained an action or a set of actions, which means that the action element was included by default. Conditions were scored for both conditions of application for principles and situational conditions (referring to physical conditions in the problem or equations as justification) for actions. This resulted in eight categories. See Table II for operationalization and examples of the different self-explanation categories. Half the written self-explanations were coded by the first and last authors. Interrater agreement was 89% with Cohen's unweighted kappa $= 0.85$, an almost perfect interrater agreement [68]. Differences were discussed until agreement. The first author coded the remaining self-explanations.

We did not distinguish between correct and incorrect explanations in the coding scheme for several reasons. How students self-explain is more important than the explanation's present correctness for their long-term learning outcomes [9,71,72]. The students may facilitate later learning by formulating abstract rules (declarative) that are incorrect because they objectify the knowledge [20]. Further, for procedural skills, missing knowledge seems to be the main reason behind students' mistakes, not incorrect knowledge [73]. Finally, incorrect rules will usually not gather enough strength to affect behavior in the long term [74], but there will probably be a positive effect from potentiating the relevant concepts that enter into the incorrect rule [42] for later—correct—rule construction.

TABLE II. Examples of students' self-explanations of worked example solutions.

SE categories	Operationalization	Examples of explanations
Action (A)	These explanations only described actions taken in solution steps without relating it to the three other elements.	"Then, we must divide the answer by the radius."
Action goal (AG)	These explanations assigned goals to actions taken in solution steps, tying together solution steps.	"Use this expression to find the velocity at the bottom of the hill."
Action principle (AP)	These explanations named solution steps by physics principle.	"Use Newton's second law with the forces that act on the block."
Action condition (AC)	These explanations explicated conditions for actions in solution steps.	"The normal force is zero in the limit where the block loses contact with the loop."
Action-goal condition (AGC)	These explanations assigned goals to, and conditions for, actions taken in solution steps.	"We know that the block can't lose contact with the surface. Then the normal force has a limit at $F_N = 0$. Use this to find the minimum velocity."
Action-goal principle (AGP)	These explanations assigned goals to actions taken in solution steps and named them by the physics principles.	"Find the velocity at the top of the hill by using conservation of mechanical energy."
Action-principle condition (APC)	These explanations named solution steps by physics principles and explicated conditions of application for the principle or other conditions for action.	"We assume no friction. We have a difference in height. Mechanical energy is conserved, can use $E_{\mathrm{mek},0} = E_{\mathrm{mek}}$."
Action-goal principle condition (AGPC)	These explanations assigned goals to actions taken in solution steps, named them by the physics principle, and explicated conditions of application for the principle or other conditions for actions.	"This [expression] can be used together with the principle of conservation of mechanical energy to calculate the speed at the bottom. Conservation of mechanical energy can be used because friction is negligible."

A potential drawback is lower correlations with post-test results.

Students in our study sometimes rewrote parts of the problem statement, which could be coded as paraphrases [9]. We excluded these paraphrases from our coding and analysis. Further, they often wrote descriptively of what was done in the solution steps, which we coded as action. These propositions could conceivably be viewed as paraphrases because they did not go beyond the information in the worked example as emphasized in Chi *et al.* [9]. However, we still view these as self-explanations because the students were arguably trying to learn from the examples. They were explicitly told to explain the solutions to problems, which still resulted in action being the largest category. Clearly, the students feel that they are explaining when merely describing actions. Based on the quantification of the qualitative analysis, each student was assigned a total number of self-explanations for each category. The overall averages can be seen in Table III.

TABLE III. The mean number (SD) of different self-explanation categories. Note that $n = 18$.

	Self-explanations
Number of problems self-explained	4.4 (1.7)
Number of total self-explanations	14.9 (7.0)
Number of action only	6.6 (5.3)
Number of action goal	1.9 (1.8)
Number of action principle	1.2 (1.4)
Number of action condition	1.6 (1.9)
Number of action-goal condition	0.3 (0.7)
Number of action-goal principle	1.1 (1.2)
Number of action-principle condition	0.9 (1.3)
Number of action-goal-principle condition	1.3 (2.4)

B. Results study 1

The following sections cover results from our exploratory analysis. As this study is of an exploratory character,

TABLE IV. Correlations between the aggregated self-explanation measures, post-test problem-solving score, and score on MBT items 5, 8, and 10. Note that $n = 18$ for all correlations.

Aggregated SE categories	Individual SE categories	Post-test	MBT (5, 8, 10)
Number of action-only SEs	A	−0.19	−0.12
Number of No-principle SEs	AG, AC, AGC	0.06	−0.02
Number of principle-no-condition SEs	AP, AGP	0.55*	0.56*
Number of principle-and-condition SEs	APC, AGPC	0.52*	0.49*

*$p < .05$.

we do not correct for type I error inflation to avoid overlooking potentially effective (and novel) treatments and relationships because of overly conservative statistics [70]. When considering p values, the effect size estimates and theoretical plausibility should also be considered [70]. For reference, Cohen's d effect sizes of about 0.20, 0.50, and 0.80 are usually treated as small, medium, and large, respectively [75]. Hemphill [76] showed that roughly a third of psychological meta-studies has $r < 0.20$, the middle third has r of 0.20 to 0.30, and the upper third has $r > 0.30$, giving empirical guidelines for small (<0.20), medium (0.20 to 0.30), and large (>0.30) correlations.

1. Retrieval practice

A two-tailed t-test showed a significant effect of doing retrieval practice the week before (retrievers) on the post-test problem-solving score ($M = 14.5, \mathrm{SD} = 4.7$) compared to no retrieval practice the week before (nonretrievers) ($M = 11.3, \mathrm{SD} = 5.5$), $t(53.26) = 2.32$, $p = 0.024$, $d = 0.61$.

We explored the effect of retrieval practice on conceptual understanding through score on MBT-items 5, 8, and 10 [61]. On average, the retrievers scored higher ($M = 2.17$, $\mathrm{SD} = 1.03$) than the nonretrievers ($M = 1.59$, $\mathrm{SD} = 0.98$), on the aggregate measure of these three items, $t(47.21) = 2.06$, $p = 0.045$, $d = 0.56$.

There was no significant difference between the retrievers ($M = 13.5$, $\mathrm{SD} = 5.5$) and the nonretrievers ($M = 13.6$, $\mathrm{SD} = 4.4$) on the full MBT, $t(42.89) = 0.025$, $p = 0.98$.

To check whether the results were due to systematic differences between the 20 nonretrievers that participated in the retrieval experiment and the 13 nonretrievers that did not, we repeated the analysis without the 13 added nonretrievers. Without the 13 added nonretrievers, the effect size was slightly smaller and nearly significant for doing retrieval practice the week before on the post-test problem-solving score ($M = 14.5, \mathrm{SD} = 4.7$) compared to no retrieval practice the week before ($M = 12.0, \mathrm{SD} = 5.4$), $t(38.26) = 1.62$, $p = 0.11$, $d = 0.50$. The effect size on the three conceptual items for doing retrieval practice the week before ($M = 2.17$, $\mathrm{SD} = 1.03$) was still significant—with a lower p value—compared to no retrieval practice the week before ($M = 1.40, \mathrm{SD} = 0.88$), $t(41.99) = 2.63$, $p = 0.01$, $d = 0.79$.

2. The quality and quantity of self-explanations

In the following section, we explore the correlations between the different categories of written self-explanations and the post-test scores. Only the 18 self-explainers (SE group) were included in the analysis of the effect of self-explanation categories. The students in the SE + PS group also solved problems, which may distort associations. They also seemed to put very low effort into explaining the problems and rather concentrated their efforts on solving the problems. Hence, all correlations are for the SE group only.

Because of the low number of subjects and the low number of self-explanations in some categories, we aggregated the different self-explanation categories into those containing principle and condition ("Principle-and-condition SEs": APC and AGPC), those containing principle without condition ("Principle-no-condition SEs": AP, and AGP), and those not containing a principle ("No-principle SEs": AG, AC, and AGC). The action-only explanations (action-only SEs) were kept separate due to their being conceptually different (only descriptive) and their prevalence. See Table IV for the aggregated self-explanation categories, their constituent self-explanation categories, and their correlations with scores.

A multiple linear regression was calculated to predict post-test problem-solving scores based on action-only SEs, the three aggregated SE categories, and whether participants engaged in retrieval practice the week before. A significant regression equation was found [$F(5, 12) = 3.16$, $p = 0.005$], with an $R^2 = 0.57$ (step 2) and adjusted $R^2 = 0.39$. The model was significant, but only principle-and-condition SEs and principle-no-condition SEs were significant as predictors. A multiple linear regression was therefore calculated where only these predictors were included. A significant regression equation was found [$F(2, 15) = 9.06$, $p = 0.003$], with an $R^2 = 0.55$ (step 1) and adjusted $R^2 = 0.49$. Regression coefficients, their standard errors, the standardized coefficients, and the p values can be found in Table V.

A multiple linear regression was calculated to predict conceptual score based on action-only SEs, the three aggregated SE categories, and whether participants engaged in retrieval practice the week before. A significant regression equation was found [$F(5, 12) = 5.67$, $p = 0.007$], with an $R^2 = 0.70$ (step 2) and adjusted $R^2 = 0.58$. Action-only SEs

TABLE V. Hierarchical regression of aggregated SE categories and retrieval practice as predictors of post-test problem-solving score. Note that $n = 18$.

		ΔR^2	B	SE B	β	p
Step 1		0.55				
	Constant		3.83	2.03		0.08
	Principle-no-condition SEs		2.05	0.67	0.53	0.008
	Principle-and-condition SEs		1.08	0.38	0.50	0.01
Step 2		0.02				
	Constant		2.78	3.26		0.41
	Action-only SEs		0.03	0.27	0.02	0.92
	No-principle SEs		0.11	0.46	0.05	0.82
	Principle-no-condition SEs		1.98	0.75	0.51	0.02
	Principle-and-condition SEs		1.05	0.45	0.48	0.03
	Retrieval practice		1.98	2.65	NA	0.47

TABLE VI. Hierarchical regression of aggregated SE categories and retrieval practice as predictors of score on MBT items 5, 8, and 10. Note that $n = 18$.

		ΔR^2	B	$SE\ B$	β	p
Step 1		0.69				
	Constant		−0.12	0.35		0.73
	Principle-no-condition SEs		0.38	0.11	0.50	0.005
	Principle-and-condition SEs		0.17	0.06	0.40	0.02
	Retrieval practice		1.04	0.39	NA	0.02
Step 2		0.01				
	Constant		−0.39	0.52		0.47
	Action-only SEs		0.03	0.04	0.15	0.44
	No-principle SEs		−0.01	0.07	−0.02	0.93
	Principle-no-condition SEs		0.38	0.12	0.51	0.007
	Principle-and-condition SEs		0.19	0.07	0.45	0.02
	Retrieval practice		1.07	0.42	0.42	0.03

and No-principle SEs were not significant predictors. A multiple linear regression was therefore calculated where these measures were excluded. A significant regression equation was found [$F(3, 14) = 10.17,\ p < 0.001$], with an $R^2 = 0.69$ (step 1) and adjusted $R^2 = 0.62$. Regression coefficients, their standard errors, the standardized coefficients, and the p values can be found in Table VI.

C. Discussion—Study 1

1. How did retrieval practice affect performance?

The retrievers scored better than the nonretrievers on both problem solving and conceptual knowledge. However, the design does not enable us to distinguish how retrieval practice affected post-test performances. It could have (i) indirectly affected performance through potentiated learning when studying the worked examples, (ii) indirectly affected performance through potentiating all their studying in the intervening week between the retrieval practice experiment and the self-explanation experiment, or (iii) directly affected performance in one of the ways discussed in Sec. I. C. 3.

This study was a quasiexperimental test of retrieval practice, and the experiments were not designed explicitly for this purpose. Therefore, there are some validity threats that one should keep in mind when interpreting the results. Although it is hard to implement without substantial attrition, a randomized controlled trial would have reduced the potential threats to the statistical conclusion validity and internal validity [70]. The main threat is that the 13 new students may have been fundamentally different from the other participants, and all were added to the nonretriever group. However, there are no apparent reasons to believe that they were systematically different beyond not participating in the retrieval practice experiment. The students were all from the same course, and their backgrounds were relatively homogenous. Moreover, the experiments were conducted during regular lecture hours, and we expect random fluctuations of this size for attendance. The follow-up analysis in Sec. II. B. 1, excluding the 13 additional students, gave much the same results with a slightly smaller effect size on problem solving ($d = 0.50$ vs 0.61) and larger effect size for conceptual test score ($d = 0.79$ vs 0.56). These results add to our confidence that the effects were not due to systematic differences between treatment conditions.

A second potential threat is that the 13 students who did not participate in the retrieval practice experiment lost two hours of studying compared to the other students. However, the analysis where we excluded the 13 students suggests that this effect alone cannot account for the majority of the observed differences.

A third potential threat is that the students were unevenly distributed into groups in the self-explanation experiment and received different treatments. However, the students in the self-explanation experiment—whether they participated in the retrieval practice experiment or not—were randomized into the three conditions in the self-explanation experiment. For the retrievers, 33% were allocated into the PS group, 36% into the SE group, and 30% into the PS + SE group. For nonretrievers, the distribution between the three groups where 42%, 26%, and 35%, respectively. Moreover, the three conditions were not significantly different on the full mechanics baseline test. Hence, the allocation into the (uninvestigated) experimental conditions cannot explain the effects of the retrieval practice.

Overall, it seems plausible that students can benefit from engaging in retrieval practice of essential physics concepts before studying worked examples. Retrieval creates the initial, strong encoding of the relevant concepts and is an easy intervention to implement in the early stages of learning new topics. Self-explanation then becomes a way to expand conceptual understanding in a manner directly relevant to problem solving. If the students cannot retrieve the name of the equation, or even the equation itself, they will be unable to name the action by principle during self-explanation. This problem was exemplified when one of the participants remarked, “I don’t walk around remembering equations six weeks before the exam,” referring to Newton’s second law (the most fundamental law or equation in mechanics). In addition to building the base strength of chunks, retrieval practice builds associative links between the names of physics concepts (e.g., mass and conservation of mechanical energy), their symbols or equations (e.g., m and $K_1 + U_1 = K_2 + U_2$), and the principles’ conditions of application (e.g., only conservative forces acting on the system). Through worked examples and problem solving, the student practices with the many special cases of these principles and conditions. Presumably, they now become more able to connect these special cases hierarchically to the main principle and its abstract condition (e.g., friction and air resistance under the category of nonconservative forces; gravitation and springs as conservative forces). This type of categorization reduces the fan effect [77]. It also reduces the needed number of stored chunks because special cases of equations can be constructed from the principles. Finally, the increased fluency of recalling relevant facts enables the student to use spare capacity on other subprocesses of problem solving [74,78,79].

In the second study, we wanted to do an experimental test of the effects of retrieval practice on post-test performance (hypothesis 1), see whether and how retrieval practice affects the quality of self-explanations (hypothesis 2), and see whether and how it changes their problem solving (hypothesis 3).

2. How did the quality of self-explanations affect performance?

The correlations in Table IV, as well as the regression models in Tables V and VI, indicate that explicating principles is the most crucial self-explanation activity. It is also important to note the high regression weights in Tables V and VI, and the large amount of variance explained (R^2), which indicates that there may be considerable benefits to be gained per self-explanation. Explanations without principles (action-only SEs and no-principle SEs) were neutral or even tending towards negative correlations. However, the qualitative analysis of the self-explanations suggested that physics principles and their conditions are relatively more important for self-explaining the physics modeling part of problems, while actions with goals and situational conditions are relatively more important for self-explaining the procedural part. Unfortunately, the data from study 1 cannot provide a definite answer to this as we had too few participants, and we did not separate explanations of the physics model from those of the mathematical procedures. Therefore, we wanted to explore this further in the second study (hypotheses 4 and 5).

As this was a correlational analysis, we cannot tell whether the students’ self-explanations were already acquired knowledge representations or whether they were constructed (learned) during the experiment, or both. However, the correlations suggest that it is essential to learn—somehow—more complex knowledge representations in terms of the four SE elements. See Sec. IV for a discussion on causality.

III. STUDY 2

A. Method

The first study indicated that retrieval practice may improve problem solving and conceptual scores. We experimentally tested this hypothesis in study 2. We also wanted to see whether and how retrieval practice changes the quality of the students’ self-explanation and problem solving. Further, we wanted to reproduce the analysis of self-explanation categories with more participants and investigate some hypotheses generated from the first study and further reading of the literature, see Secs. III. A. 1 and III. A. 2.

1. Hypotheses for retrieval practice

Hypothesis 1: Retrieval practice of principles and their conditions of application, before self-explaining solutions, improves post-test performance.

Hypothesis 2: Retrieval practice of principles and their conditions of application improves the quality of self-explanations through increased explication of principles and their conditions.

Hypothesis 3: Retrieval practice of principles and their conditions of application increases the probability that students recognize the correct principles and explicate their conditions while solving problems.

2. Hypotheses for self-explanations

Hypothesis 4: Model explanations with principles and conditions are the best predictors of problem-solving performance, but both model explanations with principles (but no conditions) and procedure explanations without principles additionally predict performance.

Hypothesis 5: Only principle-based explanations of the physics model positively predict scores on a conceptual knowledge test.

3. Participants

The study was approved by the Norwegian Centre for Research Data. The participants provided informed consent. The experiment was conducted during regular lecture hours, which meant that the number of participants was limited to how many showed up. The participants were 54 undergraduate students [65% males, modal age 20, median age 21, mean age 21.3 yr ($SD = 2.8$), mostly ethnic Norwegians] from the cohort in the subsequent year of the same introductory mechanics course as in the previous study. The course participants came from a mixture of study programs (approximately 21% nanotechnology, 18% ocean technology, 13% physics, 13% energy, 9% teacher education, 7% petroleum technology, and 9% other). The students in the course have very similar educational backgrounds where almost everyone has completed two years of physics at the high-school level before admission, so there were small differences between study programs. All participants had equal chances of winning one of three gift certificates of 1000 NOK (~110 usd) at the end of the experiment. They were randomly assigned to one of two groups in a one-factorial between-subject design. 29 students were in the treatment group and 25 were in the control group. All the concepts in the experiment had been covered in lectures before the experiment took place.

4. Procedure and materials

Study 2 was a randomized controlled trial on the effects of retrieval practice and a follow-up correlational analysis of self-explanations with more participants than in study 1. To enable randomization and to minimize attrition, the whole experiment was conducted within the same session.

The overall design of the experiment was as follows: A 20-min experimental phase → a 40-min practice phase → a 30-min post-test phase. The first step was a 20-min experimental phase. In this phase, the treatment group did retrieval practice of relevant physics principles and their conditions while the control group wrote self-explanations of solutions to work-energy problems. Then followed a 40-min practice phase where everyone self-explained solutions to six problems. Two problems required the use of conservation of momentum, two problems required the use of conservation of mechanical energy, and two problems required the use of both principles; see Supplemental Material [34] for the practice problems. Both groups performed written work, which was checked for whether participants followed instructions. The final step was a 30-min post-test phase, which consisted of a 15-min problem-solving test and a 15-min conceptual multiple-choice test. The structure of the experiment was constrained by the need to fit it into a normal double lecture of 90 min.

The practice material was similar to those in the first study. The solution structure was changed to separate the physics model and mathematical procedures more clearly. The prompts were also changed to accommodate the changes to the solution structure and to try to ensure that students generated a broad array of self-explanation categories ("Try to include principles, conditions and goals in your explanations: Principles: Which physics principles describe the problem (physics model) and how the equation is set up. Conditions of application for principle: How we know that these physics principles apply in this situation. Actions and goals: Which mathematical actions are taken and the goals of these actions."). It was considered unimportant to get students to finish studying all the problems as the deep structure of the problems was repeated. We included enough problems to be certain that all students would have enough material to study for much longer than the allotted practice time. The post-test in study 2 was too large to be finished by most students, thus ceiling effects were avoided. None of the students finished the post-test or got a full score.

5. Measures

The preselected conceptual multiple-choice test consisted of items 2, 4, 5, 10, 15, 21, and 22 from the Energy and Momentum Conceptual Survey (EMCS) [80]. All of us, who all have completed more than five years of physics education and substantial experience with instruction in mechanics, agreed that these seven items targeted relevant concepts for the current study. Cronbach's alpha for these seven items was 0.62. Because higher alpha values are generally preferable, mainly because low reliabilities lead to underestimation of relationships [66], and because a principal component analysis revealed that two

TABLE VII. Scoring rubric for the problem-solving post-test.

Observable actions in post-test
+1 Recognizing conservation of mechanical energy
+1 Recognizing conservation of momentum
+1 Explicating condition for conservation of mechanical energy
+1 Explicating condition for conservation of momentum
+1 Correct setup of conservation of mechanical energy
+1 Correct setup of conservation of momentum
+1 Correct final answer

TABLE VIII. The aggregated self-explanation categories and their constituents.

Aggregated SE categories	Individual SE categories
No-principle SEs	A, AG, AC, AGC
Principle-no-condition SEs	AP, AGP
Principle-and-condition SEs	APC, AGPC

of the items were unrelated to the rest, we removed items 5 and 21. This improved Cronbach's alpha from 0.62 to 0.69. It was pointed out that deleting these two items in the current study may lead to an overestimation of the true reliability. Although it is a normal practice to remove items to improve alpha [81], it is a controversial practice [82]. A better solution would have been to do a pilot study on the conceptual items and remove these two poorly performing items beforehand.

The problem-solving test consisted of five physics problems, where two problems required the use of conservation of momentum, two problems required the use of conservation of mechanical energy, and one problem required the use of both principles; see Supplemental Material [34] for the post-test problems. We agreed that the practice and post-test problems covered the same concepts, with highly similar solution structures. The post-test was scored according to the rubric in Table VII, with +1 point for each of the observable actions, for a total of 22 possible points. V. G. and V. H. P. scored 10% of the problem-solving post-tests, with an interrater agreement of 97% and Cohen's unweighted kappa = 0.93, an almost perfect interrater agreement [68]. Differences were discussed until agreement. V. G. scored the remaining tests. Cronbach's alpha for the post-test measure was 0.87.

6. Coding self-explanation categories

Because of the low number of explanations in some of the categories, and the insufficient number of participants to differentiate between them in multiple regression models, we aggregated the self-explanation categories as we did in study 1. Since we had more participants in this study, we also coded whether the explanations referred to the physics model or the mathematical procedures, which were differentiated in the solutions the students explained. We included action-only SEs (A) in no-principle SEs since we differentiated between explanations of the model and the procedures. We differentiate between explanations of the physics model from explanations of the mathematical procedures by writing "Model" and "Procedure" in front of the self-explanation categories (e.g., Model: No-principle SEs and Procedure: No-principle SEs). See Table VIII for the aggregated categories in study 2. Hence, there were six different categories of explanations in this study. Ten percent of the written self-explanations were coded by V. G. and V. H. P. Interrater agreement was 91% with Cohen's unweighted kappa = 0.85, an almost perfect interrater agreement [68]. Differences were discussed until agreement. The first author coded the remaining self-explanations. Based on the quantification of the qualitative analysis, each student was assigned a total number of self-explanations for each category. The overall average can be seen in Table IX.

B. Results and discussion for study 2

We do not correct for type I error inflation to avoid overlooking potentially effective (and novel) treatments and relationships because of overly conservative statistics [70]. When considering p values, the effect size estimates and theoretical considerations of plausibility should also be taken into account [70]. The results and discussion are presented together for ease of presentation.

1. Retrieval practice

Hypothesis 1: Retrieval practice of principles and their conditions of application—before self-explaining solutions—improves post-test performance.

TABLE IX. The mean number (SD) of self-explanation categories for the physics model and mathematical procedures for the retrieval group and the control group. Note that $N = 54$.

	Retrieval group ($n = 29$) [model—procedures]		Control group ($n = 25$) [model—procedures]	
No-principle SEs	3.1 (2.2)	6.7 (5.9)	3.6 (2.6)	5.6 (5.2)
Principle-no-condition SEs	1.7 (1.6)	0.1 (0.3)	2.0 (1.3)	0
Principle-and-condition SEs	2.2 (1.9)*	0	1.4 (2.1)*	0

*$p < 0.05$ for difference in simple Poisson regression.

Results: There was no significant effect of doing 20 min of retrieval practice on the post-test problem-solving score ($M = 7.6$, $\mathrm{SD} = 4.7$) compared to self-explaining work-related problems ($M = 5.9$,$\mathrm{SD} = 3.7$), $t(51.54) = 1.48$, $p = 0.14$ (two-tailed). There was also no significant difference between the retrievers ($M = 2.2$, $\mathrm{SD} = 1.5$) and the control group ($M = 2.6$, $\mathrm{SD} = 1.4$) on the conceptual post-test, $t(49.8) = 0.88$, $p = 0.39$, $d = -0.2$.

Nonsignificant replication attempts may still provide additional evidence for an effect [83,84]. Therefore, we performed a continuously cumulating meta-analysis (CCMA) on both results for both studies, using a spreadsheet from Ref. [85]. The CCMA showed a significant effect of doing retrieval practice on problem-solving performance, $d = 0.51$, 95% C.I. [0.13, 0.88], $p = 0.008$. The CCMA revealed no significant effect of doing retrieval practice on conceptual test performance, $d = 0.16$, 95% C.I. $[-0.22, 0.53]$, $p = 0.41$.

Discussion: In study 2, we tested the combination of the short-term indirect effect of improved learning through self-explanation and the direct effect on post-test performance. In study 1, the long-term indirect effect of enhanced learning in the intervening week was also part of the treatment package. The most straightforward interpretation of the results from study 2 is that retrieval practice, compared to self-explanation of problems with related concepts, does not affect post-test performances. However, the continuously cumulating meta-analysis showed that the overall effect of retrieval practice was significant for problem-solving performance—with study 2 providing additional evidence in the form of a decreasing p value—and nonsignificant for conceptual test scores. The difference in patterns of effect sizes in the two studies for problem solving and conceptual test performance suggests some unexplained moderators of the effect, especially for conceptual test performance.

We suspect that time (subsequent study) is an important moderator of the effect of retrieval practice of physics principles and their application conditions. The effect sizes indicate that the direct effect and the short-term indirect effect, through self-explanation, are larger for problem-solving performance than conceptual test performance. We may need a longer time gap between retrieval practice and the post-test—i.e., more intervening study with an increased focus on and access to principles—to affect conceptual understanding. Studying for 40 min after 20 min of retrieval practice of physics principles might not be enough to affect the construction of conceptual knowledge significantly. Importantly, all three pathways (see Sec. I. C) would affect learning if retrieval practice was an integrated part of the students' study habits throughout the semester.

A second potential moderator is the degree of prior elaborative encoding of the principles. There are reasons to believe that students need to encode the principles elaboratively—construct meaningful associative connections within and between principles—before doing retrieval practice to get maximum benefits [50,86]. For example, study 1 was done at the end of the semester when the students had already formed many meaningful associative connections for the relevant principles, while study 2 was done when the relevant concepts were first being introduced in lectures. The lack of prior elaborative encoding may have reduced the effect of retrieval practice.

A third potential moderator is skill in self-explanation. Training students to construct self-explanations might potentiate the effect of retrieval practice because they become more aware of the importance of physics principles and their conditions of application. We believe that, if it is to be used, retrieval practice should be part of an integrated instructional approach where physics principles and their conditions of application are the central concepts [50].

There are also several reasons why this was a conservative test of the effect of retrieval practice. The control group self-explained work-energy problems for 20 min. This might have transferred to improved post-test performance, either through enhanced understanding of energy-related concepts or through practice with self-explanations and potential internalization of prompts. All the relevant physics principles were also available to both groups on a sheet of paper during self-explanation in the practice phase, which reduced the need for being able to retrieve the principles and their conditions. Further, while the practice and post-test problems were essentially isomorphic in study 1, their surface features were substantially different and the solution structures were slightly different in study 2. Hence, one could argue that the problem solving and conceptual post-test in study 1 was near and medium transfer, while the problem solving and conceptual post-test in study 2 were medium and far-transfer, respectively [87]. This difference could have reduced the effect of the intervention, reflected in the lower amount of variance (R^2) explained in study 2, see Tables XI and XIII. Finally, we believe the treatment level was below optimal with the 20 min of retrieval practice.

Hypothesis 2: Retrieval practice of principles and their conditions of application improves the quality of self-explanations through increased explication of principles and their conditions.

Results: The mean number of self-explanation categories by group can be found in Table IX. We performed simple Poisson regression models for the count of the aggregate self-explanation categories. The retrieval group had significantly more Model: Principle-and-condition self-explanations than the control group, ($p = 0.02$), see Table IX. None of the other comparisons were significant.

Discussion: We found that the retrieval group had significantly more Model: Principle-and-condition self-explanations. This is evidence that retrieval practice can improve the quality of self-explanations, the first link in the mediating pathway from retrieval practice through self-explanation to improved post-test performance. This is an important and promising result, as we believe explanations containing both principles and conditions to be the most important for explaining the physics model and the most important type of self-explanation overall for learning to solve physics problems. However, the improved quality of the self-explanations, together with the direct effect on post-test performance, was not sufficient to make the experimental effect on post-test performance significant, see Sec. IV for further discussion.

Hypothesis 3: Retrieval practice of principles and their conditions of application increases the probability that students recognize the correct principles and explicate their conditions while solving problems.

Results: We used binomial regression to test whether retrieval practice affected the proportion of required principles recognized or the proportion of conditions for use of principles explicated during the problem-solving post-test. The model was not significant for recognizing principles, $\chi^2(1) = 0.24$, $p = 0.63$. However, the model was significant for explicating conditions of application, $\chi^2(1) = 11.0$, $p < 0.001$. Retrieval was a significant predictor, with $b = 1.75$, $z = 2.78$, $p = 0.005$, Hosmer-Lemeshow $R^2 = 0.16$, Cox and Snell $R^2 = 0.18$, Nagelkerke $R^2 = 0.26$. The odds ratio was 5.75, with 95% CI [1.9, 24.7].

Discussion: We found no difference in the proportion of correct principles recognized during problem solving. However, the retrieval group explicated a significantly higher proportion of conditions for using principles during the problem-solving post-test (OR = 5.76). This is evidence that retrieval practice can positively affect what students attend to during problem solving. Recognizing principles and conditions of application is part of the potential direct effect on problem solving from retrieval practice. However, the same two physics principles were repeated in all the practice and post-test problems. Therefore, the need to recognize application conditions—for becoming able to solve the problem—was probably reduced relative to regular studying. The repetition of physics principles can also explain why there was no significant difference between the two groups in recognizing the correct principles, as there were only two from which to choose. This may have attenuated the direct effect of retrieval practice on problem solving compared to a more ecologically valid situation with more uncertainty in which principles to apply. We still expect that there was an effect on working memory during problem solving but that this was not sufficient to make the experimental effect significant.

Considering conditions of application for principles during problem solving is presumably very important for becoming able to transfer their declarative knowledge and procedural skills to new problems later, when it is less obvious which principles apply. Therefore, we believe that this is an important result.

2. Self-explanations

Hypothesis 4: Model explanations with principles and conditions are the best predictors of problem-solving performance, but both model explanations with principles (but no conditions) and procedure explanations without principles additionally predict performance.

Results: We calculated Pearson's correlation coefficients for the aggregated self-explanation categories with the post-test problem-solving score, for explanations referring to the physics model and for explanations referring to the mathematical procedures. The correlations can be seen in Table X. Explanations of procedures with principles (Procedure: Principle-and-condition-SEs and Procedure: Principle-no-condition SEs) were nearly nonexistent and, therefore, correlations would be meaningless.

A multiple linear regression was calculated to predict the post-test problem-solving score based on the aggregated SE categories we hypothesized to be most important, Model: Principle-and-condition, Model: Principle-no-condition SEs, and Procedure: No-principle SEs. We did not include retrieval practice vs control as a predictor, as the experimental results were insignificant. A significant regression equation was found [$F(3, 50) = 10.9$, $p < 0.001$], with an $R^2 = 0.40$ and adjusted $R^2 = 0.36$ (step 2). The model was significant, but Model: Principle-no-condition SEs was not a significant predictor. A multiple linear regression was therefore calculated where we excluded this predictor. A significant regression equation was found [$F(2, 51) = 14.4$, $p < 0.001$], with an $R^2 = 0.36$ and adjusted $R^2 = 0.34$ (step 1). Both predictors were significant. Regression coefficients, their standard errors, the standardized coefficients, and the p values can be found in Table XI.

Discussion: The correlations in Table X support our hypothesis that explanations of the physics model containing principles and conditions are the best predictors of problem-solving performance, and that both model explanations with principles (but no conditions) and procedure explanations without principles additionally predict performance. It seems that more complex explanations, containing principles, conditions, and potentially also goals, are suitable for explanations of physics models, while

TABLE X. Correlations between the aggregated SE measures and the post-test problem-solving score. Note that $N = 54$ for all correlations.

Aggregated SE categories	Model SEs	Procedural SEs
No-principle SEs	0.10	0.34^{*}
Principle-no-condition SEs	0.30^{*}	NA
Principle-and-condition SEs	0.50^{***}	NA

$^{a}p < 0.10$
$^{*}p < 0.05$
$^{**}p < 0.01$
$^{***}p < 0.001$.

TABLE XI. Hierarchical regression of aggregated SE categories as predictors of post-test problem-solving score. Note that$N = 54$.

		ΔR^2	B	SE B	β	p
Step 1		0.36				
	Constant		3.25	0.85		<0.001
	Procedural: no-principle SEs		0.26	0.09	0.33	0.004
	Model: principle-and-condition SEs		1.06	0.24	0.49	<0.001
Step 2		0.04				
	Constant		2.58	0.94		0.009
	Procedural: no-principle SEs		0.20	0.09	0.26	0.031
	Model: principle-no-condition SEs		0.59	0.35	0.20	0.099
	Model: principle-and-condition SEs		1.06	0.24	0.49	<0.001

TABLE XII. Correlations between the aggregated SE measures and score on the conceptual test. Note that $N = 54$ for all correlations.

Aggregated SE categories	Model SEs	Procedural SEs
No-principle SEs	−0.18	0.05
Principle-no-condition SEs	0.39**	NA
Principle-and-condition SEs	0.25[a]	NA

[a] $p < 0.10$
* $p < 0.05$
** $p < 0.01$
*** $p < 0.001$.

simpler explanations, without principles, are suitable for mathematical procedures.

In the multiple regression model in Table XI, the only significant predictors of problem-solving performance were Model: Principle-and-condition SEs and Procedural: No-principle SEs. However, the nominal trend was in the hypothesized direction for model explanations with principles and no conditions. One may note the very high correlation and standardized regression coefficient of Model: Principle-and-condition SEs. This may indicate that this is where our efforts should be directed in teaching students how to self-explain.

Procedure-explanations with principles were basically nonexistent. It is possible to construct these types of explanations, but one would have to backtrack to where the equations in the procedures originated. The fact that almost none of the students constructed these types of explanations for procedures is evidence that they at least do not come naturally.

Hypothesis 5: Only principle-based explanations of the physics model positively predict scores on a conceptual knowledge test.

Results: We calculated Pearson's correlation coefficients for the aggregated self-explanation categories with the score on the conceptual test for explanations referring to the physics model and for explanations referring to the mathematical procedures. The correlations can be seen in Table XII. Explanations of procedures with principles (Procedure: Principle-and-condition SEs and Procedure: Principle-no-condition SEs) were nearly nonexistent and, therefore, correlations would be meaningless.

A multiple linear regression was calculated to predict conceptual test scores based on the SE categories we hypothesized to be most important Model: Principle-and-condition, Model: Principle-no-condition SEs, and

TABLE XIII. Hierarchical regression of aggregated SE categories as predictors of conceptual test score. Note that $N = 54$.

		ΔR^2	B	SE B	β	p
Step 1		0.22				
	Constant		1.23	0.36		0.001
	Model: principle-no-condition SEs		0.42	0.13	0.40	0.003
	Model: principle-and-condition SEs		0.20	0.10	0.25	0.048
Step 2		0.01				
	Constant		1.34	0.39	⋯	0.001
	Procedural: no-principle SEs		−0.03	0.04	−0.10	0.47
	Model: principle-no-condition SEs		0.46	0.14	0.43	0.002
	Model: principle-and-condition SEs		0.19	0.10	0.25	0.049

Procedure: No-principle SEs. We did not include retrieval practice vs control as a predictor, as the experimental results were insignificant. A significant regression equation was found [$F(3, 49) = 4.81$, $p = 0.005$], with an $R^2 = 0.23$ and adjusted $R^2 = 0.18$ (step 2). The model was significant, but Procedural: No-principle SEs did not reach significance as a predictor. A multiple linear regression was therefore calculated where we excluded Procedural: No-principle SEs. A significant regression equation was found [$F(2, 50) = 7.02$, $p = 0.002$], with an $R^2 = 0.22$ and adjusted $R^2 = 0.19$ (step 1). Regression coefficients, their standard errors, the standardized coefficients, and the p values can be found in Table XIII.

Discussion: We can see from Table XII that only Model: Principle-no-condition SEs were significantly related to scores on the conceptual test, while Model: Principle-and-condition SEs did not quite reach significance. However, both predictors were significant in the multiple linear regression model in Table XIII. Unlike for problem-solving score, Procedural: No-principle SEs did not significantly predict conceptual test score. These results support our hypothesis that explanations of physics models that contain a reference to physics principles are important for developing conceptual knowledge. Indeed, conceptual multiple-choice tests often remove the need for mathematical procedures while increasing the need for an understanding of the relevant physics principles and when they apply.

IV. GENERAL DISCUSSION

Much educational research has been done on retrieval practice and self-explanations separately. However, there is little to no research on the interplay between retrieval practice and self-explanation. To our knowledge, no one has tried to manipulate levels of relevant prior knowledge for self-explanation using retrieval practice. One of the most critical jobs for physics education researchers is taking general research and theory on learning and finding ways to adapt it to our domain-specific context. Physics is a unique domain because it is highly structured and centered around principles and application conditions. This makes the combination of methods in this paper especially suited for physics.

In this article, we set out to investigate how we can improve students' self-explanations of worked examples in physics by asking two general research questions: (1) What knowledge representations should students seek while self-explaining worked examples to maximize their learning? (2) Can retrieval practice of physics principles and their conditions of application potentiate students' learning from self-explaining worked examples?

Research question 1: We defined four elements of self-explanation (actions, goals, principles, and conditions of application) based on the ACT-R theory and empirical results from the literature. We based our coding categories of written self-explanations on these four elements. In two studies, we explored the correlations between categories of self-explanation and learning outcomes. The results indicate that different combinations of SE elements are useful for different parts of the problem. For example, when explaining the modeling part of solutions—where principles are used to describe the problem—explanations should center around the principles used and their application conditions and descriptions of how the principle is set up in the specific situation. In addition, one might speculate that including the goal of the entire model would help students link the physics model to the mathematical procedures. When explaining the procedural process from the physics model to the final answer, goals and action descriptions should be the main components of the explanations. Different types of explanations can have additive effects [88], but they are only useful to the extent that they support generalization to new and different problems [89,90]. The presented category system clarifies what elements must be included to make self-explanations generalizable. These results also lead us to ask whether it would be wise to more clearly separate the structural elements of worked examples and solutions in physics textbooks, i.e., clearly separating models from procedures. Indeed, Lee *et al.* [91] found that students got an improved final understanding and increased transfer when they studied examples that emphasized the underlying structure.

We have contributed to the literature with a category system that clarifies what high-quality self-explanation means in physics, even distinguishing between the different structural elements in worked examples. We did not investigate whether (high-quality) self-explanations cause better problem-solving or conceptual test scores. However, it is well established that self-explanations cause declarative knowledge representations of procedural problem-solving steps, considering learning theory [37,39] and prior empirical research [12,20]. Indeed, trying to understand an example—self-explanation—is the fundamental way we learn this type of knowledge [37]. Therefore, it seems certain that the knowledge representations described by the categories of self-explanations—whether learned during or before the self-explanation practice phase—are important for post-test performance, also evidenced by the large correlations. Further research is still warranted to increase confidence in the causality of these categories of self-explanations.

Teaching students how to explain might ultimately be more important than directly teaching them specific domain knowledge through lectures, demonstrations, and examples [72]. However, we cannot make any firm suggestions for instructing students in self-explanation. There is a lack of effective instructional methods for teaching self-explanation in the literature [50]. Even though they were prompted, the students in this study tended to generate a low amount of high-quality explanations, something others have also experienced [15]. However, some research indicates that it is possible to affect what types of explanations students construct [88] and that training might improve their self-explanation skills [92]. The category system from this paper may further spark innovative methods and interventions. We suspect that new instructional methods should center around the physics model and principle-and-condition SEs and be implemented through active learning and group discussions. We also suspect that making the relevant knowledge components for self-explanation more accessible must be part of an effective instructional strategy. Retrieval practice is a prime candidate because it is the best learning strategy for improving memory access.

Research question 2: We quasiexperimentally (study 1) and experimentally (study 2) tested the effects retrieval practice of physics principles and their conditions of application had on problem solving and conceptual knowledge. The results from the two studies, together with the continuously cumulating meta-analysis in Sec. III. B. 1, indicate that retrieval practice has a medium-sized positive effect on problem-solving performance while there is no significant effect on conceptual test performance. However, in Secs. I. C, II. C. 1, and III. B. 1, we discussed the three potential pathways of learning from retrieval practice to improved post-test performances, (i) indirectly through improved self-explanation quality, (ii) indirectly through potentiating all learning strategies, or (iii) by directly affecting performance. The pattern of results can provide some insights into these pathways; see the following three paragraphs.

The results from the two studies suggest that the indirect pathway through improved quality of self-explanations leads to an effect that is smaller than 0.4 standard deviations for problem solving and has negligible effects for conceptual knowledge. Notably, the results from the second study (hypothesis 2) established that retrieval practice does improve the quality of self-explanations, even though the intervention only entailed 20 min of retrieval practice and with a conservative control treatment (self-explanation of similar content). One may wonder why the improvements in self-explanation quality did not translate into significantly improved post-test performance. There are some potential interpretations of this finding: First, the improvements in self-explanation quality may have been superficial. Many of the students were unaware of or could not access the relevant physics principles before the 20 min of retrieval practice. We cannot expect these students to suddenly construct deep and insightful self-explanations merely because they can now access the principle. However, it is an essential first step for novice students to connect physics principles and their conditions to their understanding of problem-solving steps. Second, it may mean that preexisting knowledge is behind both self-explanation quality and post-test performance and that the self-explanation practice phase had no effect on the post-tests. However, we find it more likely that both prior knowledge and new knowledge gained through self-explanation affected the student's post-test performances. Indeed, learning theory makes it clear that when you construct high-quality self-explanations of a worked example, you will learn from the experience [36,37,46,93]. Finally, high-quality self-explanation may lead to improved conceptual knowledge, but perhaps not noticeably on such a short time scale. Section III. B. 1 also discusses the two potential moderators of prior elaborative encoding and self-explanation skills.

The results from the two studies suggest that the indirect (long-term) pathway through improved learning from all other learning strategies is important, especially for conceptual test performance. Retrieval practice may be more of a long-term potentiator of all study strategies than a means to specifically and immediately potentiate self-explanation, see Sec. I. C. 2. When the long-term pathway was removed (in study 2), the effects were no longer significant, although the problem-solving post-test performance provided additional evidence in the continuously cumulating meta-analysis (lower overall p value, see Sec. III. B. 1). The finding that retrieval practice affected both the self-explanation quality and the tendency to explicate application conditions during problem solving is indirect evidence that it can improve long-term learning. Lending further support to this hypothesis, Gjerde, Holst, and Kolstø [50] found that students in a mechanics course noticed (long-term) benefits from having engaged in retrieval practice of physics principles and their conditions. Finally, we must consider the possibility that the positive effects in study 1 were due to systematic differences between the groups rather than the effect of retrieval practice, but see the discussion of validity in Sec. II. C. 1.

We experimentally tested one part of the potential direct pathway from retrieval practice to improved post-test problem-solving performance, namely, whether the students were more likely to recognize relevant physics principles and explicate application conditions during problem solving. There was no effect on the proportion of correct principles recognized, perhaps because the same two principles were repeated in all the practice and post-test problems. Retrieval practice may have a larger effect on problem solving during regular studying when it is less obvious which principles to apply. However, the students were much more likely to explicate the application conditions for the principles during problem solving. This has important implications for long-term learning. Considering conditions of application during problem solving is important for becoming able to transfer knowledge to new problems, especially when it is less obvious which principles apply. As discussed in Sec. I. C, we also expect the stronger memories of principles to lead to less pressure on working memory resources and that it becomes easier to embed the principles into more complex knowledge representations. These effects are more relevant for problem solving than conceptual knowledge because of the mathematical complexity. The direct effects on problem solving may have contributed to the effect size of 0.4, but it was not enough to reach statistical significance in study 2.

It is premature to suggest retrieval practice as an intervention with the sole purpose of improving learning from self-explanation. However, retrieval practice of physics principles and their application conditions is intended to potentiate all other learning strategies [31]. Considering (i) the large research literature on how retrieval practice is an effective learning strategy and a promising method for education [54,94,95], (ii) our prior publications on retrieval practice in physics education [31,50], and (iii) that the results in this paper suggest that the potential benefits are quite large and that the potential harm is small or nonexistent, we are comfortable with making a soft suggestion for including retrieval practice of principles and their conditions of application in an overall instructional strategy. A method for doing this in introductory mechanics can be found in Refs. [31,50].

Learning is the result of what the students do and attend to. Moreover, it is a known problem that physics students tend to focus on "formulas" and "getting the right answer." Hence, that we were able to shift the students' focus more towards principles and their conditions of application in their study—during self-explanation and problem solving in this paper—is an exciting result, especially considering how quick and easily implementable the intervention is.

We discuss weaknesses in the two studies in Secs. II. C. 1 and III. B. 1. We also discuss potential moderators of the effect of retrieval practice in Sec. III. B. 1. Further, we cannot say whether our self-explanation results generalize to oral self-explanations. Oral explanations seem to stimulate more conceptual elaboration than written explanations [96]. More research is needed on retrieval practice, self-explanations, and the interplay between the two learning strategies in physics.

V. CONCLUSIONS

In this study, we have begun answering questions about the knowledge representations students should seek from self-explaining worked examples and whether retrieval practice can potentiate self-explanation. For example, it seems that when explaining the physics modeling part of worked examples, students should try to explain how the model is set up, explicate physics principles and their conditions of application, and potentially also how the physics model leads to the goal. Further, when explaining the mathematical procedures in the worked examples, students should try to explain the action, the goal of the action, and the conditions of its application. Our results indicate that retrieval practice improves problem-solving performance, enhances the quality of self-explanations, and increases the tendency to consider application conditions during problem solving. However, we note some potential weaknesses in the paper—primarily related to the design in study 1—that warrant further research before solid recommendations can be given.

For self-explanations, finer-grained research could be done on how the individual SE categories affect different outcomes. Experimental testing of the different self-explanation categories may reveal whether and how they cause learning. There may also be important insights to gain from a similar qualitative analysis of students who explain out loud instead of writing self-explanations. It is also important to explore why some students construct effective explanations while others do not [22] and how educators can affect it. Research should be done to make retrieval practice of physics principles and their conditions maximally effective for potentiating other learning strategies. This may involve researching the hypothesized moderators in Sec. III. B. 1: Intervening time from retrieval practice to test, prior elaboration, and self-explanation skill. The treatment levels of retrieval practice and self-explanation should also be varied. Research should be done to isolate the potential mediating pathways from retrieval practice to post-test performance, discussed in Sec. I. C. Further research should also be done to investigate whether and when retrieval practice can potentiate the learning of conceptual knowledge versus problem-solving skill. Finally, as much current physics education research involves student dialogues, research should be done to investigate how retrieval practice of principles and their conditions can affect the content of student dialogues and other generative learning activities.

ACKNOWLEDGMENTS

The University of Bergen supported this work. We thank Martino Marisaldi, Kjellmar Oksavik, and Kjartan Olafsson for helping us with evaluating the conceptual items in study 1.

[1] C. Angell, Ø. Guttersrud, E. K. Henriksen, and A. Isnes, Physics: Frightful, but fun—Pupils' and teachers' views of physics and physics teaching, Sci. Educ. **88**, 683 (2004).

[2] F. Örnek, W. R Robinson, and M. Haugan, , What makes physics difficult, Int. J. Environ. Sci. Educ. **3**, 30 (2008).

[3] A. Renkl, Learning from worked-examples in mathematics: Students relate procedures to principles, Int. J. Math. Educ. **49**, 571 (2017).

[4] A. Renkl, R. K. Atkinson, U. H. Maier, and R. Staley, From example study to problem solving: Smooth transitions help learning, J. Exp. Educ. **70**, 293 (2002).

[5] R. J. C. M. Salden, K. R. Koedinger, A. Renkl, V. Aleven, and B. M. McLaren, Accounting for beneficial effects of worked examples in tutored problem solving, Educ. Psychol. Rev. **22**, 379 (2010).

[6] X. Zhu and H. A. Simon, Learning mathematics from examples and by doing, Cognit. Instr. **4**, 137 (1987).

[7] R. Badeau, D. R. White, B. Ibrahim, L. Ding, and A. F. Heckler, What works with worked examples: Extending self-explanation and analogical comparison to synthesis problems, Phys. Rev. Phys. Educ. Res. **13**, 020112 (2017).

[8] T. van Gog, N. Rummel, and A. Renkl, Learning how to solve problems by studying examples, *The Cambridge Handbook of Cognition and Education* (Cambridge University Press, Cambridge, England, 2019). p. 183–208.

[9] M. T. H. Chi, M. Bassok, M. W. Lewis, P. Reimann, and R. Glaser, Self-explanations—How students study and use examples in learning to solve problems, Cogn. Sci. **13**, 145 (1989).

[10] M. T. H. Chi and K. A. VanLehn, The content of physics self-explanations, J. Learn. Sci. **1**, 69 (1991).

[11] A. Renkl *et al.*, Learning from worked-out examples: The effects of example variability and elicited self-explanations, Contemp. Educ. Psychol. **23**, 90 (1998).

[12] K. Bisra, Q. Liu, J. C. Nesbit, F. Salimi, and P. H. Winne, Inducing self-explanation: A meta-analysis, Educ. Psychol. Rev. **30**, 703 (2018).

[13] P. Gerjets, K. Scheiter, and R. Catrambone, Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations?, Learn. Instr. **16**, 104 (2006).

[14] B. Rittle-Johnson, A. M. Loehr, and K. Durkin, Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles, Int. J. Math. Educ. **49**, 599 (2017).

[15] A. Renkl, Learning from worked-out examples: A study on individual differences, Cogn. Sci. **21**, 1 (1997).

[16] J. Wittwer and A. Renkl, How effective are instructional explanations in example-based learning? A meta-analytic review, Educ. Psychol. Rev. **22**, 393 (2010).

[17] A. Renkl and R. K. Atkinson, Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective, Educ. Psychol. **38**, 15 (2003).

[18] R. K. Atkinson, A. Renkl, and M. M. Merrill, Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps, J. Educ. Psychol. **95**, 774 (2003).

[19] K. Berthold *et al.*, The double-edged effects of explanation prompts, Comput. Human Behav. **27**, 69 (2011).

[20] M. T. H. Chi *et al.*, Eliciting self-explanations improves understanding, Cogn. Sci. **18**, 439 (1994).

[21] A. Renkl, Toward an instructionally oriented theory of example-based learning, Cogn. Sci. **38**, 1 (2014).

[22] P. A. Alexander, The development of expertise: The journey from acclimation to proficiency, Educ. Res. **32**, 10 (2003).

[23] C. Dignath and M. V. J. Veenman, The role of direct strategy instruction and indirect activation of self-regulated learning-evidence from classroom observation studies, Educ. Psychol. Rev. **33**, 489 (2021).

[24] J. J. Williams and T. Lombrozo, The role of explanation in discovery and generalization: Evidence from category learning, Cogn. Sci. **34**, 776 (2010).

[25] J. J. Williams and T. Lombrozo, Explanation and prior knowledge interact to guide learning, Cogn. Psychol. **66**, 55 (2013).

[26] R. M. F. Wong, M. J. Lawson, and J. Keeves, The effects of self-explanation training on students' problem solving in high-school mathematics, Learning Instr. **12**, 233 (2002).

[27] P. K. Agarwal *et al.*, Benefits from retrieval practice are greater for students with lower working memory capacity, Memory **25**, 764 (2017).

[28] 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, Cogn. Syst. Res. **1**, 99 (2000).

[29] 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).

[30] Z. F. Shen *et al.*, Item strength affects working memory capacity, Memory Cognit. **46**, 204 (2018).

[31] 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).

[32] J. R. Anderson, J. M. Fincham, and S. Douglass, Practice and retention: A unifying analysis, J. Exper. Psychol. Learning Memory Cogn. **25**, 1120 (1999).

[33] J. C. K. Chan, C. A. Meissner, and S. D. Davis, Retrieval potentiates new learning: A theoretical and meta-analytic review, Psychol. Bull. **144**, 1111 (2018).

[34] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.18.010136 for practice and post-test problems.

[35] J. R. Anderson, *How Can the Human Mind Occur in the Physical Universe?*, Oxford Series on Cognitive Models and Architectures (Oxford University Press, Oxford; New York, 2007), p. 290.

[36] 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).

[37] J. R. Anderson and C. Lebiere, *The Atomic Components of Thought* (Lawrence Erlbaum Associates, Mahwah, NJ, 1998), p. 490.

[38] J. R. Anderson and J. M. Fincham, Acquisition of procedural skills from examples, J. Exper. Psychol. Learning Memory Cogn. **20**, 1322 (1994).

[39] J. R. Anderson, J. M. Fincham, and S. Douglass, The role of examples and rules in the acquisition of a cognitive skill, J. Exper. Psychol. Learning Memory Cogn. **23**, 932 (1997).

[40] V. Aleven and K. R. Koedinger, An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor, Cogn. Sci. **26**, 147 (2002).

[41] P. I. Pavlik and J. R. Anderson, Using a model to compute the optimal schedule of practice, J. Exp. Psychol. Appl. **14**, 101 (2008).

[42] J. R. Anderson, *The Architecture of Cognition*, Cognitive Science Series (Harvard University Press, Cambridge, MA, 1983), p. 345.

[43] E. H. Geller, J. Y. Son, and J. W. Stigler, Conceptual explanations and understanding fraction comparisons, Learning Instr. **52**, 122 (2017).

[44] X. Zhu *et al.*, Cue recognition and cue elaboration in learning from examples, Proc. Natl. Acad. Sci. U.S.A. **93**, 1346 (1996).

[45] J. Larkin , J. McDermott, D. P. Simon, and H. A. Simon, Expert and novice performance in solving physics problems, Science **208**, 1335 (1980).

[46] J. R. Anderson and J. M. Fincham, Extending problem-solving procedures through reflection, Cogn. Psychol. **74**, 1 (2014).

[47] M. Wells, D. Hestenes, and G. Swackhamer, A modeling method for high-school physics instruction, Am. J. Phys. **63**, 606 (1995).

[48] K. Berthold and A. Renkl, Instructional aids to support a conceptual understanding of multiple representations, J. Educ. Psychol. **101**, 70 (2009).

[49] 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).

[50] 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).

[51] J. R. Anderson and C. D. Schunn, Implications of the ACT-R learning theory: No magic bullets, in *Advances in Instructional Psychology*, Educational Design and Cognitive Science Vol. 5, edited by R. Glaser (Routledge, London, 2000), p. 1–34.

[52] J. R. Anderson, L. M. Reder, and C. Lebiere, Working memory: Activation limitations on retrieval, Cogn. Psychol. **30**, 221 (1996).

[53] W. J. Hopper and D. E. Huber, Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects, J. Memory Lang. **102**, 1 (2018).

[54] C. A. Rowland, The effect of testing versus restudy on retention: A meta-analytic review of the testing effect, Psychol. Bull. **140**, 1432 (2014).

[55] 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).

[56] P. L. Pirolli and J. R. Anderson, The role of practice in fact retrieval, J. Exper. Psychol. Learning Memory Cogn. **11**, 136 (1985).

[57] N. Kornell, R. A. Bjork, and M. A. Garcia, Why tests appear to prevent forgetting: A distribution-based bifurcation model, J. Memory Lang. **65**, 85 (2011).

[58] 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).

[59] J. Waldeyer *et al.*, Can generative learning tasks be optimized by incorporation of retrieval practice?, J. Appl. Res. Memory Cogn. **9**, 355 (2020).

[60] P. A. Alexander *et al.*, College instruction and concomitant changes in students' knowledge, interest, and strategy use: A study of domain learning, Contemp. Educ. Psychol. **22**, 125 (1997).

[61] D. Hestenes and M. Wells, A mechanics baseline test, Phys. Teach. **30**, 159 (1992).

[62] J. E. Richey and T. J. Nokes-Malach, How much is too much? Learning and motivation effects of adding instructional explanations to worked examples, Learning Instr. **25**, 104 (2013).

[63] S. Schworm and A. Renkl, Computer-supported example-based learning: When instructional explanations reduce self-explanations, Comput. Educ. **46**, 426 (2006).

[64] K. Sijtsma, On the use, the misuse, and the very limited usefulness of Cronbach's alpha, Psychometrika **74**, 107 (2009).

[65] K. S. Taber, The use of cronbach's alpha when developing and reporting research instruments in science education, Res. Sci. Educ. **48**, 1273 (2018).

[66] N. Schmitt, Uses and abuses of coefficient alpha, Psychol. Assess. **8**, 350 (1996).

[67] L. J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika **16**, 297 (1951).

[68] J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, Biometrics **33**, 159 (1977).

[69] R. Bakeman *et al.*, Detecting sequential patterns and determining their reliability with fallible observers, Psychol. Methods **2**, 357 (1997).

[70] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Houghton Mifflin, Boston, 2001), p. 623.

[71] T. Lombrozo, Explanation and abductive inference, *The Oxford Handbook of Thinking and Reasoning* (Oxford University Press, New York, NY, 2012), pp. 260–276.

[72] D. A. Wilkenfeld and T. Lombrozo, Inference to the best explanation (IBE) versus explaining for the best inference (EBI), Sci. Educ. **24**, 1059 (2015).

[73] K. VanLehn, Cognitive skill acquisition, Ann. Rev. Psychol. **47**, 513 (1996).

[74] J. R. Anderson, Acquisition of cognitive skill, Psychol. Rev. **89**, 369 (1982).

[75] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (L. Erlbaum Associates, Hillsdale, 1988), p. 567.

[76] J. F. Hemphill, Interpreting the magnitudes of correlation coefficients, Am. Psychol. **58**, 78 (2003).

[77] J. R. Anderson, The adaptive nature of human categorization, Psychol. Rev. **98**, 409 (1991).

[78] K. E. Woolley, T. D. Huang, and M. Rabinowitz, The effects of knowledge, strategies, and the interaction between the two in verbal analogy problem solving, Contemp. Educ. Psychol. **56**, 91 (2019).

[79] B. D. Mikula and A. F. Heckler, Framework and implementation for improving physics essential skills via computer-based practice: Vector math, Phys. Rev. Phys. Educ. Res. **13**, 010122 (2017).

[80] C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, Am. J. Phys. **71**, 607 (2003).

[81] M. A. Robinson, Using multi-item psychometric scales for research and practice in human resource management, Hum. Resour. Manag. Rev. **57**, 739 (2018).

[82] R. Hoekstra *et al.*, An empirical analysis of alleged misunderstandings of coefficient alpha, Int. J. Social Res. Methodology **22**, 351 (2019).

[83] S. L. Braver, F. J. Thoemmes, and R. Rosenthal, Continuously cumulating meta-analysis and replicability, Perspect. Psychol. Sci. **9**, 333 (2014).

[84] R. Rosenthal, Replication in behavioral research, J. Soc. Behav. Pers. **5**, 1 (1990).

[85] C. G. Sibley, *Utilities for Examining Simple Meta-Analytic Averages [Computer Software]* (University of Auckland, Auckland, NZ, 2008).

[86] 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).

[87] S. M. Barnett and S. J. Ceci, When and where do we apply what we learn?: A taxonomy for far transfer, Psychol. Bull. **128**, 612 (2002).

[88] T. Lombrozo and N. Z. Gwynne, Explanation and inference: Mechanistic and functional explanations guide property generalization, Front. Human Neurosci. **8**, 1 (2014).

[89] T. Lombrozo and S. Carey, Functional explanation and the function of explanation, Cognition **99**, 167 (2006).

[90] T. Lombrozo, Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions, Cogn. Psychol. **61**, 303 (2010).

[91] H. S. Lee, S. Betts, and J. R. Anderson, Embellishing problem-solving examples with deep structure information facilitates transfer, J. Exp. Educ. **85**, 309 (2017).

[92] K. Bielaczyc, P. L. Pirolli, and A. L. Brown, Training in self-explanation and self-regulation strategies—Investigating the effects of knowledge acquisition activities on problem-solving, Cognit. Instr. **13**, 221 (1995).

[93] J. R. Anderson and L. J. Schooler, Reflections of the environment in memory, Psychol. Sci. **2**, 396 (1991).

[94] 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).

[95] H. L. Roediger and M. A. Pyc, Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice, J. Appl. Res. Memory Cognit. **1**, 242 (2012).

[96] L. Jacob, A. Lachner, and K. Scheiter, Learning by explaining orally or in written form? Text complexity matters, Learn. Instr. **68**, 101344 (2020).