Effect of presentation style and problem-solving attempts on metacognition and learning from solution videos

Jason W. Morphew,[1] Gary E. Gladding,[2] and Jose P. Mestre[2]

[1]*School of Engineering Education, College of Engineering, Purdue University, West Lafayette, IN, 47907, USA*
[2]*Department of Physics, College of Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA*

(Received 30 July 2019; published 24 January 2020)

Students must actively engage in problem solving to effectively learn in introductory physics courses. However, students often get stuck and are not able to make progress when solving problems outside of their current ability, particularly when one-on-one tutoring and instructor office hours are a limited resource. One effective technique consists of providing students with worked examples during the problem-solving process. While the benefits of worked examples are well established, less is known about how the format of the worked example affects student learning, or the effect of solution videos on student metacognition. This study presents three experiments investigating how the format of animated worked examples affects student learning and metacognition. The results indicate that students learn equally well from different styles of solution videos that follow multimedia learning principles. In addition, attempting to solve problems before viewing the solution videos facilitates learning for problems just outside a student's current ability, but not for more difficult problems. Further, attempting to solve very difficult problems before viewing animated solution videos can potentially lead to overconfidence, where students believe that they learned more from the solutions than they have actually learned.

DOI: 10.1103/PhysRevPhysEducRes.16.010104

I. INTRODUCTION

Attrition among science, technology, engineering, and mathematics (STEM) undergraduate majors is a significant problem. While over one-quarter of students entering a bachelor's degree program enroll in a STEM major at some point in their career, only half of those students leave having completed a STEM degree [1]. Several factors affect students' decisions to persist, however, the grades students earn in their introductory courses, which are often primarily determined by examinations, are a strong predictor of persistence within STEM majors [2,3]. Given the importance of high-stakes exams typically found in introductory STEM courses, it is important for research to explore possible interventions that might help struggling students better prepare to take exams.

Among the STEM disciplines, many students find introductory calculus-based physics difficult to learn, and not surprisingly, the difficulties manifest themselves most blatantly in exam performance. Typical physics exams require that students apply concepts and procedures to solve problems. At Illinois about 30% of the students enrolled in the introductory mechanics course for scientists and engineers score below a B- on the 3 midterm exams administered in the course, despite 95% of engineering majors at Illinois scoring in the top 5% in ACT math. The great majority of these underperforming students spend considerable time attempting to prepare for the midterm exams, so it is not a lack of time on task that is the reason for poor exam performance. In view of this situation, what can be done to help underperforming students do better on midterm exams? This study explores one promising approach, namely, learning from animated-narrated solution videos (ANSVs) of typical exam problems.

Several strands of research have a bearing on the research described in this study. We begin by briefly reviewing pertinent literature on how students typically prepare for exams and the factors that contribute to poor performance. We continue with a brief review of the literature on learning from worked examples, followed by a discussion of metacognition, and a description of the design of animated solution videos. Finally, we present three studies exploring the extent to which students could learn to solve problems from studying ANSVs that (a) highlighted how concepts and procedures are applied to solve problems, (b) used a worked-examples approach, (c) were designed according to multimedia learning principles, and (d) used problems at the higher end of difficulty from typical midterm exams.

A. Study habits of low and high performing students

Research using survey methods suggest that students study for exams using passive methods, such as rereading and reviewing notes, and tend to focus the majority of the studying one to two days before an exam [4–7]. Furthermore, students do not tend to change their study strategies throughout the semester regardless of the success—or lack of success—of their current study strategies [4]. Although much of the research has not been done in the domain of physics, it is likely that this is also the approach used by most students to study for exams in the STEM disciplines. The passive approach, however, is not well suited for doing well on physics exams. Physics exams require students to display problem-solving proficiency, which in turn requires an understanding of concepts, the conditions for applying them, and the procedures for applying them. Physics homework is designed to develop these skills, however, when students are unable to make progress in solving homework problems, they often resort to finding the solutions online or through their peers. The practice of copying solutions can lead to homework grades not reflecting a mastery of the material. Because some students who struggle on exams still earn most or all of the points on homework assignments, homework scores in large introductory physics classes typically do not correlate with exam scores. The lack of correlation between homework and exam grades suggests that many students would benefit from additional support in developing the conceptual understanding and problem-solving skills needed for success on the exams. The additional support may be particularly useful for students who receive below-average exam grades on the exams, because the ways in which they are successfully completing homework assignments is not sufficient for preparing them for solving problems under the constraints of an exam (timed performance with no resources except an equation sheet and calculator). In particular, providing students with support in the form of solutions that focus on developing conceptual understanding in addition to procedural competence may be particularly useful for lower-performing students.

The literature on expertise strongly indicates that to become good at something (in this case test performance in physics) takes lots of practice at that something. Students often report that they intend to test themselves by solving problems from available practice tests when studying for exams, but do not actually do so, believing that testing is a relatively ineffective study strategy [4]. Students persist in this belief despite extensive empirical evidence that testing is more effective than more passive methods, such as reviewing notes or rereading for long term retention of material [8–12]. This overall situation is particularly problematic for lower-performing students who are more likely to cram and often use passive study strategies such as rereading or reviewing notes rather than engaging in self-testing [5,13,14].

One reason that students might prefer passive methods is that they are not able to make progress when engaging in active problem solving for problems that are outside of their current ability. Presenting students with worked examples is one option that has shown promise for supporting students in solving well-defined computational and procedural problems [15,16], particularly when one-on-one tutoring and instructor office hours are a limited resource.

B. Learning from worked examples

A common technique for teaching students how to solve calculational or procedural problems is to engage them in studying worked examples that provide the entire solution procedure. The worked-example effect describes the robust finding that studying worked examples is more effective at teaching novices and low-performing students how to solve well-structured problem-solving tasks than merely engaging them in problem-solving activities [17,18], or engaging in tutored problem solving [19,20]. Worked examples also better prepare students for learning as compared to working on open-ended problems [21], and are particularly effective when combined with self-explanation or analogical comparison prompts [22,23]. However, worked examples are less effective, and can even be detrimental for higher-performing students [24]. Experts and more knowledgeable novices, such as high-performing introductory students, tend to learn more from solving problems with feedback than from viewing worked examples [25–28].

The worked-example effect is generally interpreted through the lens of cognitive load theory, in which learning is constrained by limitations in working memory capacity. The amount of information elements that are processed in working memory at any time determines the cognitive load experienced by the learner. Current views of cognitive load theory distinguish three sources of cognitive load; extrinsic, intrinsic, and germane [29]. Intrinsic cognitive load refers to the load imposed by the intrinsic nature of the task, or as Sweller [30] defines it, intrinsic load is the number of interacting elements that must be simultaneously processed to understand and learn the material. As one learns, schemas are created which "chunk" the interacting elements reducing intrinsic load.

Cognitive load that is imposed by the design of the learning task rather than from the content itself is either extrinsic or germane load. Extrinsic load is load imposed by the design of the learning task that is unnecessary for learning in that it does not directly relate to schema construction [29]. For example, equation hunting (i.e., searching for equations with the needed variables) is a task characterized by high extraneous load because it imposes a high load on working memory and does not lead to enhanced schema development. Conversely, germane load refers to the load imposed by the design of the learning task that is directly related to schema acquisition and automation [30]. For example, engaging in conceptual

analysis is a task characterized by high germane load because the load imposed on working memory facilitates schema development.

The benefits of worked examples for novices and low-performing students is thought to occur from reductions in extrinsic cognitive load and increases in germane cognitive load [31,32]. As such, worked examples that require students to attend to multiple sources of information, or that provide students with redundant information, are ineffective [32,33]. In addition, worked examples have been found to be associated with increases in learner motivation [34], and increases in the efficiency with which novices process information needed to develop general conceptual understanding [20].

Much of the research has focused on static worked examples, such as those commonly found online, or in textbooks [17,35,36]. However, students have poor recall for conceptual information presented in static solutions even if they attend to the presented conceptual explanations [37]. With advances in technology, worked examples can be dynamically presented through video or animated solutions, through interactive help links, or within cognitive tutoring systems [19,38–41]. The affordances provided by dynamically presented solutions may stem from a reduction in extrinsic cognitive load by directing the learner's attention to the relevant portions of a solution. In fact, dynamically presented solutions have been found to be more effective than static worked examples [42,43]. Dynamically presented solutions, such as those presented in ANSVs, aim to reduce extrinsic cognitive load by guiding the learner's attention, and presenting conceptual information through both visual and aural channels. The reduction of extrinsic cognitive load afforded by dynamically presented solutions may free up cognitive resources to help students better attend to conceptual explanations for the procedural steps contained within a solution. If true, the enhanced germane cognitive load could explain the positive impact that dynamic solutions have compared to static solutions.

In addition to explaining why low-performing students learn from worked solutions, and why dynamic solutions appear to be more effective than static solutions, cognitive load theory can explain why solutions can be less effective for high-performing students. Presenting high-performing students with worked solutions containing already known or no longer needed procedural reminders are processed as redundant information which may increase the extrinsic load for higher-ability learners thus limiting, or even reversing, the benefits from studying worked examples [26,44]. However, methods for presenting worked examples that require the student to engage in actively building their conceptual schemas, such as interleaving problem solving and worked examples, fading worked examples, and presenting incomplete worked examples, may be beneficial for higher-performing individuals [18,45,46].

Previous studies have found that worked examples increase students' self-efficacy for solving similar problems [47], however, few studies have investigated the impact of studying worked examples on the accuracy of students' metacognitive judgments. It may be that providing low-performing students with expert solutions will make explicit the discrepancy between their current level of understanding and the level of understanding expected to solve physics problems in the course. Alternatively, the reduction of cognitive load from viewing the solution videos may also increase the fluency with which individuals process the information given the additional working memory capacity. In addition, because students are viewing the solution videos when reviewing for exams rather than during initial learning, the solutions may present highly familiar procedures. If students overuse the familiarity and fluency cues presented in animated worked examples, they may develop an "illusion of understanding" in which they believe that they have learned the material and will be able to appropriately apply the solution method to a new problem simply because they were able to understand the example [48,49].

C. Metacognition

Metacognition is the act of thinking and reflecting on one's cognitive processes, and is commonly divided into metacognitive knowledge and metacognitive skills, such as monitoring and control. In authentic self-regulated learning contexts, it is generally believed that there is a dynamic and reciprocal relationship between metacognitive monitoring and control processes.

Learners engaged in self-regulated learning monitor their current knowledge state against context dependent and task-specific criteria in order to plan and enact effective study strategies [50,51]. Because of learners' reliance on monitoring their ability in order to make effective metacognitive control decisions, the accuracy of students' metacognitive monitoring (i.e., how closely their estimate of their ability matches their current ability) is paramount in self-regulated learning contexts. As such, it is important for interventions aimed at helping students prepare for course exams to investigate how the interventions impact the accuracy of students' metacognitive monitoring.

Metacognitive monitoring is typically studied by asking learners to make judgments about the current state of their learning at various points in the learning process [52]. Two of these judgments are particularly relevant for studying interventions aimed at helping students prepare for exams; judgments of learning (JOLs) and retrospective confidence judgments (RCJs). JOLs are made after learning the material (e.g., viewing the animated worked examples in this case), but before attempting to solve new problems, while RCJs are made after attempting to solve the problems. Students are likely to make judgments about the state of their learning after viewing worked examples and after

attempting to solve problems. These judgments, along with constraints such as deadlines and perceived task value, likely determine their future studying behavior [50,53].

Early theories concerning the basis for judgments of metacognitive monitoring posited that individuals directly monitored the state of their cognition. This explanation, termed the direct-access hypothesis, predicts the strong relationship that is observed between metacognitive judgments and objective performance. However, this view fails to account for the presence of pervasive metacognitive illusions that suggest that beliefs about memory and metacognitive judgments are independent of objective measures of memory [48,54]. Current theories of metacognition tend to adopt the cue-utilization approach, which asserts that metacognitive judgments are made through inferential processes that utilize beliefs about the connection between learning and cues [e.g., 55].

In the cue-utilization framework, two types of cues (i.e., theory-based and heuristic-based) are predominantly used to make metacognitive judgments. Theory-based cues are related to the characteristics of the task, to the to-be-learned items, or to the learning conditions that one assumes to be related to difficulty of learning. Heuristic-based cues are implicit cues that learners employ as indicators concerning the degree to which items have been learned, such as, the familiarity of the content or the fluency with which they encode the material. Individuals make metacognitive judgments by integrating information from both theory-based cues and heuristic-based cues [55–57]. The extent to which each cue type influences metacognitive judgments is a function of the learning context, motivation of the learner, and the attributions that learners make [58].

While many theory-based cues can be availing, others can be harmful for learning. For example, individuals tend to believe, erroneously, that massed and blocked study is more beneficial than distributed and interleaved study [59], and that additional study sessions will lead to better retention than engaging in testing [60]. The use of availing theory-based cues when making studying decisions likely differ between high- and low-performing students. For example, high-performing students are more likely to use testing as a study strategy [5].

Students also often implicitly use heuristic-based cues, such as familiarity and fluency, when engaged in metacognitive monitoring to make judgments about the progress of their learning. The fluency with which individuals process information is related to metacognitive judgments, with individuals making higher metacognitive judgments in memory tasks for easier to read words and images [61,62], in comprehension tasks for fluently presented lectures or easier to process text [63–66], or in knowledge tasks for questions with familiar terms [67], even when these cues are not diagnostic of learning. The fluency with which an item is retrieved from memory is also related to metacognitive judgments [68,69]. For example, individuals predicting their ability to recall answers from a general knowledge test will often give the highest judgments to the items they answered most quickly, however they recall more items that take longer to answer [70]. The finding that individuals give high metacognitive judgments to easily recalled information is particularly relevant for physics courses where common misconceptions are often fluently retrieved.

D. Design features of animated-narrated solution videos

Worked examples are believed to enhance learning by reducing extrinsic cognitive load and increasing germane cognitive load [31,32]. Dynamically presented information that follows multimedia learning principles also reduce extraneous cognitive load and increase germane cognitive load [71,72], which may explain why dynamic solutions, have been found to be more effective than static solutions [42,43]. Two different "styles" of animated-narrated solution videos (ANSV) were developed for this study, which will be referred to as ANSV self-reflective and ANSV two column. Both styles of ANSVs modeled an expertlike approach where students engage in conceptual analysis of the problem, rather than engage in means-ends approaches, such as equation hunting and followed multimedia learning principles. For example, the ANSVs were relatively brief, ranging between 3 and 7 min in length. In addition, the conceptual and procedural steps were presented using both aural and visual channels. However, care was taken to ensure that the information presented across the two channels was coherent, but not redundant to avoid interference. The solution steps were animated such that the solution steps were presented in small coherent segments, and maintained in the video so that students would have a complete solution to review without the need to rewatch the entire solution, further reducing extraneous load. In both styles, students were free to stop the solution at any point, back up, or replay any segment. In addition, both styles aimed to reinforce a conceptual analysis approach to problem solving by integrating the conceptual and procedural knowledge required to solve the problems and making explicit links between the conceptual knowledge and the mathematical implementations. The ANSV styles differed in two ways besides the superficial differences in the solution layout. First, the styles differed in whether the physical scenarios were animated. Second, the styles differed in the presentation of the algebraic steps used to solve the problem. All of the animated-narrated solutions can be found at the URL address in Ref. [73].

1. Description of ANSV self-reflective

The self-reflective ANSVs begin by presenting students with the problem scenario, then present an expert's solution to a problem in a style depicting an expert thinking aloud about how to construct a solution from "first principles." These solutions are similar to the style used at the

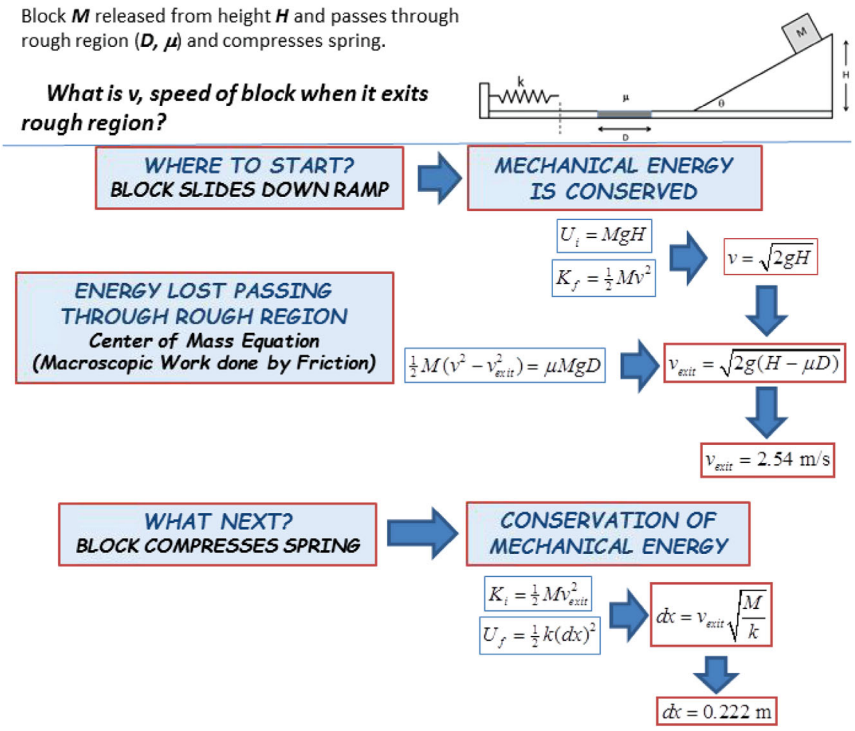

FIG. 1. Example of the self-reflective ANSV style. Solution steps are animated to reduce cognitive load, but remained visible so that students can review the entire solution. The conceptual ideas or procedures are found in the shaded boxes and are discussed and presented before the mathematical formulas.

University of Illinois to generate web-based solutions to problems in the introductory calculus-based course. These solutions alternate presenting the concept or first principles that are applicable to the problem, then the procedure for applying the concept(s) in equation form (see Fig. 1). The self-reflective style aimed to reduce the cognitive load experienced by the students by presenting the solution at a "high level." In other words, the ANSVs tried to avoid presenting students with already mastered information by only presenting the concept and accompanying equation, but not the algebraic manipulations to arrive at a final expression for the quantity being asked for in the problem. That is, once the solution steps were presented, the final symbolic form of the solution was stated and the algebraic steps were assumed to be done "off-line" to generate the final expression. In addition, the physical scenario was not animated, however, narration describing the relevant physics concepts and principles was presented.

2. Description of ANSV two column

The two-column style ANSVs begin by presenting students with an animation of the physical scenario described in the problem accompanied by narration describing the relevant physics concepts and principles (see Fig. 2). The animation of the physical scenario engaged students in mental simulation thereby modeling an additional expert-like approach missing from the self-reflective ANSV style. After orienting students to the problem, the two-column style presents the solution using a two-column style similar to ones used in previous studies [74]. This style explicitly presents the solution in two columns, with the left-column discussing the concepts being applied, and the right column providing the equation that instantiates the concept (see Fig. 3). That is, the solution always presented students with the concept or procedure that was being applied in the left column, and then discussed the mathematical instantiation of the concept in the right column. While the solution steps were being presented, animations were used to direct the students' attention to relevant portions of the solutions, because the use of animations as visual cues has been shown to help direct learners' attention and can facilitate learning [75,76]. Once the mathematical formulas were established, the algebraic steps to obtain the final solutions were explicitly carried out. These steps were animated to reinforce an understanding the mathematical processes carried out, while also reducing students' memory load by presenting the steps sequentially. For example, when finding the cross product of vectors, the vectors appeared and were manipulated on the screen.

E. Research questions

Three experiments were performed to examine how two design aspects affect students' learning and metacognitive monitoring of ANSVs. Experiment 1 examines how the

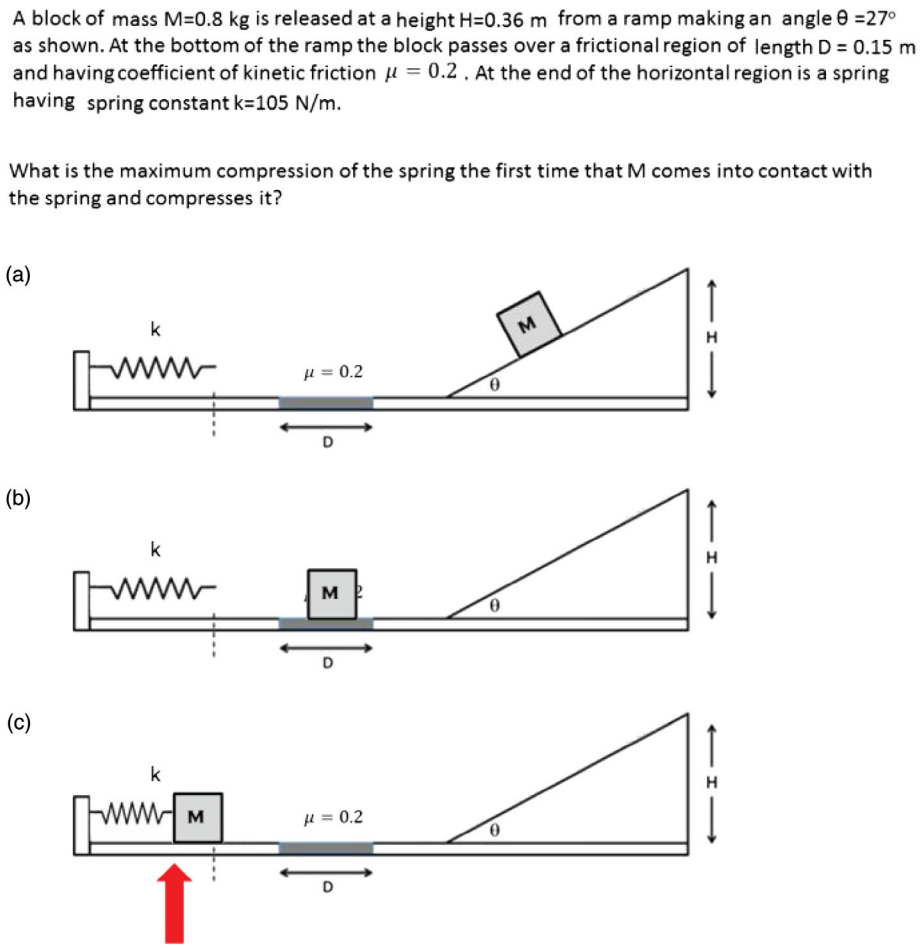

FIG. 2. Example of the animation of the physical scenario. Animations are accompanied by a conceptual analysis of the problem. In this problem, (a) there is an exchange between gravitational potential energy and kinetic energy as the block slides down the ramp. Then, (b) work is done by a nonconservative force (friction) as the block slows down over the region of friction. Finally, (c) there is an exchange between kinetic energy and spring potential energy.

method for presenting the solution (i.e., the solution style) affects students' ability to relearn previous content, as well as the accuracy of their confidence after attempting problems. Because learning gains found in experiment 1 may be due either to viewing the ANSV or to attempting the problem and receiving feedback, experiment 2 investigates whether attempting to solve difficult problems before viewing the solution videos is better for learning and metacognitive accuracy than only viewing the solution videos without first attempting to solve the problems. Experiment 3 replicates the second experiment with less difficult problems and using students from a wider range of abilities.

This study explores three main research questions across three experiments. First, to what extent does the design of the ANSV affect student learning and metacognitive monitoring accuracy? Second, how does attempting to solve problems before viewing the solutions affect learning or metacognitive monitoring accuracy? Third, to what extent does physics ability affect the accuracy of metacognitive monitoring when learning from ANSVs?

II. EXPERIMENT 1

A. Participants

Students enrolled in the second semester (electricity and magnetism) of the Fall 2015 introductory calculus-based physics sequence were recruited for this study. Most students enrolled in the course had completed the first semester in the sequence (introductory mechanics) during the previous semester. The study was conducted in the fall semester meaning that most students had not taken a physics course during the intervening summer semester. Of the 102 students that volunteered, 93 completed the pretest, post-test, and the delayed post-test. The majority of the students who completed all three components of the experiment ($n = 74$) were enrolled in the introductory mechanics course during the previous semester (Spring 2014), and their score on the final exam was available. These students were randomly assigned to conditions using matched-pair random assignment. Of the remaining nineteen students, sixteen students tested out of the mechanics

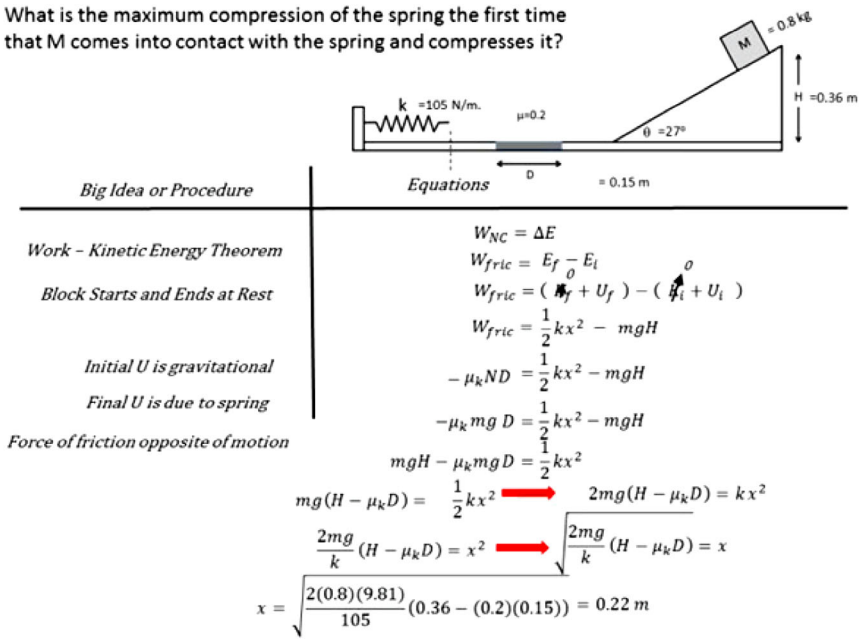

FIG. 3. Example of the two-column ANSV style. Solution steps are animated and remained visible to reduce cognitive load and so students can review the entire solution. The conceptual ideas are found in the right-hand column and are presented and discussed before the relevant mathematical formulas appear in the left-hand column.

course or transferred from another university. The remaining 3 students completed the mechanics course in Summer 2015, or Fall 2014. Measures of prior ability (i.e., course grades from the mechanics course) were not available for these nineteen students, so they were randomly assigned to the conditions.

B. Procedure

The volunteers in this study first completed the pretest on paper. After completing the problems and entering them into the computer, the students received correctness feedback indicating whether their solutions were correct or incorrect, then were asked to view the ANSVs for all five problems. After viewing all of the videos, the students completed the post-test on paper. One week later, the students returned to the lab and completed the delayed post-test on paper. Students did not receive correctness feedback or solutions to any of the problems on the post-test or the delayed post-test.

C. Materials

1. Pretests, post-tests, and delayed post-tests

Participants completed a pretest, a post-test, and a delayed post-test covering center of mass, conservation of energy, conservation of momentum, and work. The pretest consisted of five calculation-based physics problems. The post-test consisted of five calculation-based problems similar to those used on the pretest, and thirteen conceptual physics problems. The delayed post-test consisted of the same five calculation-based problems, and thirteen conceptual physics problems, as on the post-test, however, the order of the problems and the answer choices were changed. All of the questions used in experiment 1 can be found in Ref. [77].

Calculational problems required students to calculate numerical answers. Problems were scored correct if the answers were correct, or if the student made only one minor algebraic or arithmetic error in computing the answer (e.g., making a rounding error on a multistep problem). Conceptual questions were multiple-choice problems that required students to apply their conceptual understanding to determine what change, if any, would occur if the problem's initial conditions were changed. For these questions, students needed to select the correct answer and to explain their answer. Conceptual questions were scored as correct if the correct answer was selected and a correct explanation was provided. All of the questions were scored by two independent graders with an initial interrater agreement of 94%. Following discussion, 100% agreement was reached.

2. Retrospective confidence judgments (RCJs)

After attempting to solve each problem on the pretests and post-tests participants were asked to make RCJs using the following wording: "Circle the number which represents how confident you are that your answer is correct." RCJs were made using five equally spaced percentages (i.e., 0%, 25%, 50%, 75%, 100%). This scale was utilized so that the confidence judgments would be made in the same scale as the measures of performance. A single RCJ

TABLE I. Means, standard errors, and Kruskal-Wallis tests for pretest, post-test, and delayed post-test.

	Two column ($n = 50$)	Self-reflective ($n = 43$)	Kruskal-Wallis	
Exam	$M \pm$ SE	$M \pm$ SE	$\chi^2(1)$	p
Pretest	34.8 ± 3.61	38.1 ± 4.41	0.30	0.58
Post-test calculational	66.0 ± 4.47	64.7 ± 3.93	0.28	0.60
Post-test conceptual	61.2 ± 2.99	58.5 ± 3.09	0.57	0.45
Delayed postcalculational	72.0 ± 3.96	68.4 ± 4.12	0.75	0.39
Delayed postconceptual	64.0 ± 2.80	62.4 ± 3.27	0.09	0.77

score was computed for all participants by calculating the mean of the individual confidence judgments. While students in general provided RCJs after attempting every problem on the pre- and post-tests, about 12% of students failed to make an RCJ on at least one question. Only 0.4% of the individual questions did not have RCJs, and the percentage of missing RCJs ranged from 0% for most questions to 2% for four questions. Because there were no patterns in the missing data, nor did missingness correlate with any dependent variable used in the study, the data were assumed to be missing at random.[1]

3. Metacognitive judgment accuracy

The accuracy of the metacognitive judgments is often measured by examining the calibration of student judgments. Calibration, also known as absolute accuracy, refers to the ability to make judgments that accurately reflect performance [78]. In a learning context, calibration is related to the ability for students to judge when their learning is sufficient to meet the goals they have set for the task. In this study, calibration was measured using bias. Bias was calculated for each participant by subtracting the performance (i.e., the percent of questions answered correctly) from the RCJ [79]. Because bias is a signed measure of calibration, it indicates whether a judgment is higher or lower than performance allowing for the examination of overconfidence and underconfidence. In this study, bias was calculated so that positive bias indicates overconfidence and negative bias indicates underconfidence.

D. Methods

To examine differences between conditions, t tests and Kruskal-Wallis tests were conducted on the pretest scores, post-test scores, and RCJ bias scores. Normality was assessed using visual inspection of the histograms and q-q plots, as well as conducting Shapiro-Wilk tests of normality. When normality could be assumed, independent samples t tests were conducted. When normality could not be assumed, nonparametric Kruskal-Wallis tests were conducted.

To examine whether students were able to relearn from viewing the ANSVs, two difference scores were calculated by subtracting the percentage on the pretest from the percentage of similar calculation problems solved correctly on the post-test and the delayed post-test, respectively. Because the difference scores were normally distributed and a homogeneity of variance can be assumed, two independent samples t tests were conducted.

To examine the change in metacognitive bias on the calculation questions between the pretest and the post-test, a 2×2 (test bias $\times$ style) mixed ANOVA with test bias as the repeated measure and video style as the between-subjects variable was conducted.

E. Results

There was no difference on the pretest for students who completed the introductory mechanics course during the previous semester and those who did not, $\chi^2(1) = 0.53$, $p = 0.46$, therefore these students were not analyzed separately. Descriptive statistics for pretest, post-test, and delayed post-test scores are found in Table I. A difference between the groups was not detected on the pretest. The Kruskal-Wallis tests failed to detect a difference between the styles for solving similar calculation problems or conceptual transfer problems on either the post-test or the delayed post-test.

The difference scores and t-test results are presented in Table II and displayed in Fig. 4. The t tests failed to detect a difference between the styles in learning for either the post-test or the delayed post-test.

Bias scores for the RCJs are given in Table III and presented in Figs. 5 and 6. No differences between the groups were detected from the t tests. The 2×2 (test bias $\times$ style) mixed ANOVA showed a significant main effect of test, $F(1, 91) = 12.09$, $p < 0.001$, indicating that both groups on average were less overconfident after viewing the ANSVs. However, neither the main effect of style, $F(1, 91) = 0.04$, $p = 0.85$, nor the interaction were significant, $F(1, 91) = 0.29$, $p = 0.59d$, indicating that no difference in bias was detected between the two styles, and

[1]Because the data can be assumed to be missing at random, and only 88% of the participants made RCJs for every question, multiple imputation procedures are more appropriate to address the missing RCJ data. This analysis was also conducted, however, because these results led to identical conclusions, the simpler analysis is presented here. For interested readers, the analysis using multiple imputation can be found in Ref. [77].

TABLE II. Means, standard errors, and Kruskal-Wallis tests for pretest, post-test, and delayed post-test.

	Two column ($n = 50$)	Self-reflective ($n = 43$)		
Exam	$M \pm$ SE	$M \pm$ SE	$t(91)$	p
Post-pre	31.2 ± 3.7	26.5 ± 3.4	0.93	0.36
Delayed post-pre	37.2 ± 3.4	30.2 ± 4.4	1.27	0.21

that the decrease in overconfidence was similar for the two styles.

III. EXPERIMENT 2

In experiment 1, students scored between 30 and 40 percentage points higher on the post- and delayed post-tests after viewing the ANSVs. In addition, students were more calibrated in their confidence for questions discussed in the ANSVs. However, the students were engaged in active problem solving, and received correctness feedback, in addition to viewing ANSVs. As such, Experiment 1 does not allow for the isolation of the effect of viewing the ANSVs from the effect of receiving feedback after attempting to solve problems. It is possible that much of the observed relearning was due in large part to engaging in testing with feedback, which has been shown to be effective for learning [9,10]. In addition, the observed improvement in metacognitive calibration may be due to the underconfidence with practice effect, where individuals tend to be overconfident on the first trial of a task, but see their overconfidence decline, even to the point of becoming underconfident as soon as the second trial [80–82].

Experiments 2 and 3 address these issues by randomly assigning students either to attempt the problem before viewing the ANSVs or to view the ANSVs without attempting to solve the problems. If the observed learning gains are largely due to the problem solving with correctness feedback then the students who attempt the problems before viewing the solutions would be expected to demonstrate larger learning gains. In addition, the underconfidence with practice effect predicts that students who attempt the problems before viewing the solutions would also have lower bias scores than students who only watch the videos. Because experiment 1 failed to detect any differences in learning between the two styles, only the two-column ANSV video style was used to investigate the remaining research questions in experiments 2 and 3. In experiment 2 we investigate the effects of attempting to solve problems before viewing the ANSVs. We expect that attempting to solve the problem allows students to identify the particular areas where they need assistance. The feedback received from the problem attempt may allow students to focus on specific components of the videos, reducing their cognitive load. Alternatively, attempting the problem before viewing the ANSVs may lead to increased fluency with the surface features, leading to overconfidence.

A. Participants

An email was sent to the 373 students enrolled in the Fall 2015 introductory mechanics course who had scored at or below the 45th percentile on the first two midterm exams (i.e., their average exam score was below 76%) of an introductory physics course were recruited to participate in a study to help them prepare for the third exam in the course. Seventy students volunteered and were randomly assigned either to attempt to solve calculation-based physics problems before viewing the solution videos ("attempt first" condition) or to only watch the solution videos without attempting the problems ("view only" condition). The experiment was completed over two sessions to minimize student fatigue. In each session, students completed half of the pretest problems, viewed the corresponding ANSVs, then completed the corresponding post-test questions. The two sessions were separated by

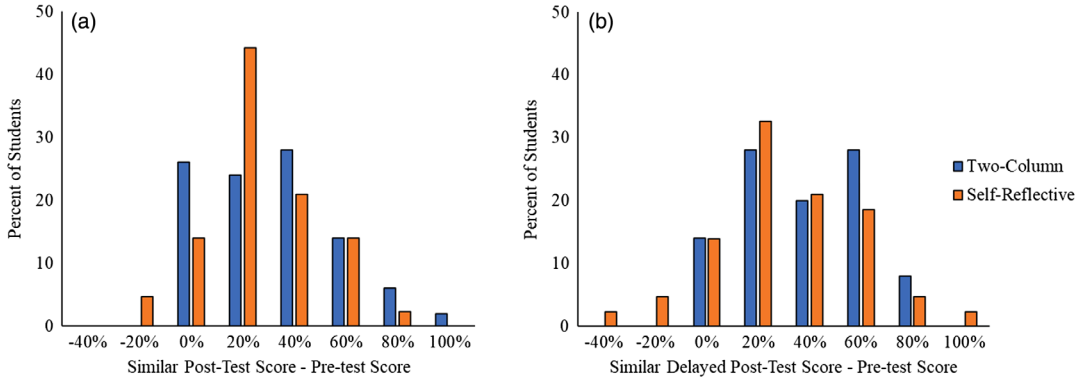

FIG. 4. Histogram of difference scores for (a) post-test and (b) delayed post-test.

TABLE III. Mean RCJs bias by video style.

	Two column ($n = 50$)	Self-reflective ($n = 43$)		
Measure	$M \pm$ SE	$M \pm$ SE	$t(91)$	P
Pretest	16.3 ± 3.88	13.9 ± 3.76	−0.44	0.66
Post-test isomorphic	4.3 ± 3.04	5.1 ± 3.52	0.18	0.86
Post-test transfer	16.6 ± 2.55	17.7 ± 2.56	0.30	0.76
Delayed post-test isomorphic	-4.9 ± 2.93	3.7 ± 3.58	1.89	0.06
Delayed post-test transfer	14.7 ± 2.73	19.2 ± 2.94	1.11	0.27

one to two days. Of the 70 students who volunteered, 60 participants completed both sessions. The two conditions differed in attendance resulting in 26 participants in the attempt first condition, and 34 participants in the view only condition. The 10 students who did not complete both sessions were excluded from the data analysis.

B. Procedure

The problems and ANSVs covered topics that would appear on the third course exam (i.e., rotational motion, angular kinematics, and angular momentum). Participants in the attempt first condition completed a pretest where they solved difficult calculation-based physics problems and made RCJs for each problem. After attempting all of the problems, the participants viewed ANSVs for the calculation problems that they had just attempted. Participants were then asked to make a JOL by indicating how many problems similar to the problems in the ANSVs they would now be able to correctly solve. Finally, the participants completed a post-test consisting of "similar" calculation-based physics problems and conceptual-based "transfer" problems.

Participants in the view only condition answered nine survey questions about their typical study habits, then viewed the same ANSVs as the attempt first condition after answering survey questions about their typical study habits. Participants were then asked to make a JOL by indicating how many problems similar to the problems in the ANSVs they would now be able to correctly solve. Finally, the participants completed a post-test consisting of calculation-based physics problems similar to the problems covered in the ANSVs and conceptual-based transfer problems.

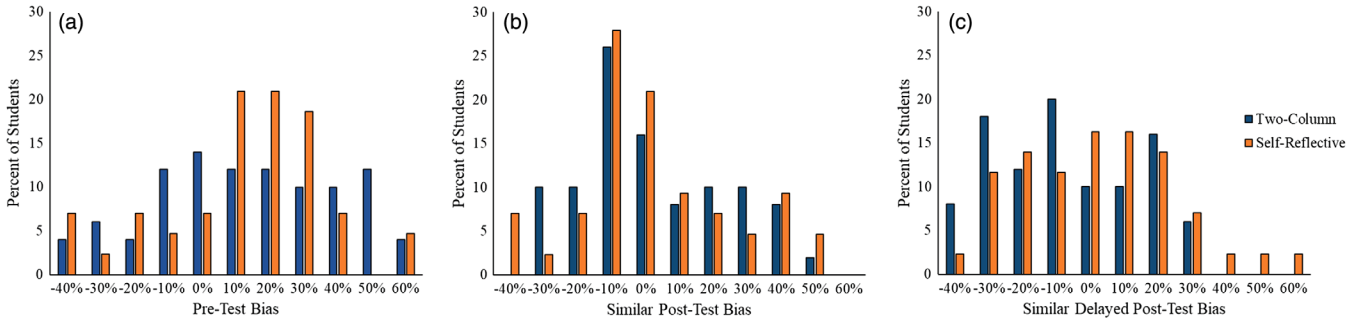

FIG. 5. Histogram of RCJ bias for (a) pretest, (b) isomorphic post-test, and (c) isomorphic delayed post-test.

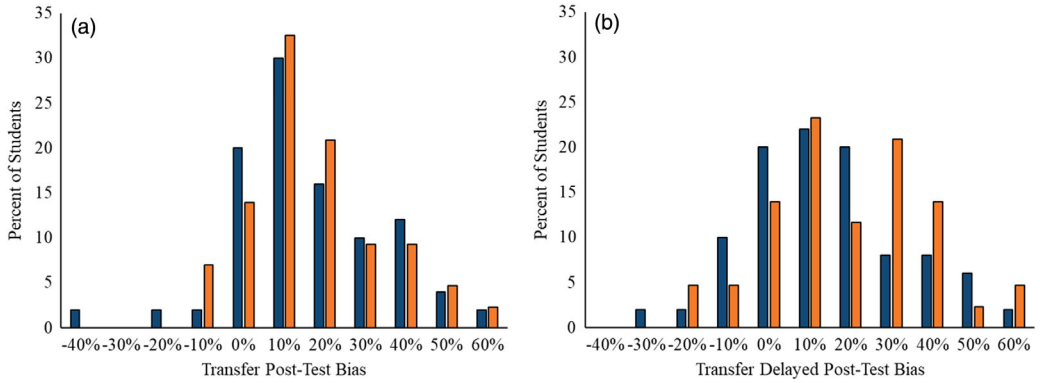

FIG. 6. Histogram of RCJ bias for transfer questions for the (a) post-test and (b) delayed post-test.

C. Materials

1. Pretests and post-tests

The pretest consisted of nine calculation-based physics problems. The post-test consisted of nine calculation-based physics problems similar to the problems covered in the ANSVs, and eight transfer problems. All of the questions used in experiment 2 can be found in Ref. [77]. All of the calculation problems required students to calculate numerical answers. Problems were scored correct if the answers were correct, or if the student made only one minor algebraic or arithmetic error in computing the solution (e.g., rounding errors on multistep problems). The transfer problems were multiple choice conceptual problems that required the participants to apply conceptual reasoning that was discussed in the ANSVs. For these questions, students needed to select the correct answer and to explain their answer. Conceptual questions were scored as correct if the correct answer was selected and the explanation was correct. All of the questions were scored by two independent graders with an initial interrater agreement of 95%. Following discussion, 100% agreement was reached.

2. Retrospective confidence judgments

Students made RCJs after attempting every problem on the pretests and post-tests as in experiment 1. A single RCJ score was computed by taking the mean of the individual confidence judgments. In general students provided RCJs after attempting every problem on the pretests and post-tests. Only 2.8% of the individual questions did not have RCJs, and the percentage of missing RCJs ranged from 0% for most questions to 10% for a single calculational question on the post-test. Because there were no patterns in the missing data, nor did missingness correlate with any dependent variable used in the study, the data was assumed to be missing at random.[2]

3. Judgments of learning

After viewing all of the ANSVs, participants were asked to indicate how many problems similar to the problems in the ANSVs they would now be able to correctly solve.

4. Metacognitive judgment accuracy

Metacognitive accuracy was measured using bias as described in experiment 1 both for JOLs and for mean RCJs.

[2]Because the data can be assumed to be missing at random, and only 68% of the participants made RCJs for every question, multiple imputation procedures are more appropriate to address the missing RCJ data. This analysis was also conducted, however, because these results led to identical conclusions, the simpler analysis is presented here. For interested readers, the analysis using multiple imputation can be found in Ref. [77].

D. Methods

To examine differences between conditions, t tests and Kruskal-Wallis tests were conducted on the pretest and post-test scores. Normality was assessed using visual inspection of the histograms and q-q plots as well as conducting Shapiro-Wilk tests of normality. When normality could be assumed, independent samples t tests were conducted. When normality could not be assumed, nonparametric Kruskal-Wallis tests were conducted. Effect sizes were calculated using Cohen's d.

To examine learning from ANSVs, within-subjects tests for repeated measures were used to compare scores on the pretest with scores on the post-tests. Normality of the difference scores was assessed using visual inspection of the histograms and q-q plots as well as conducting Shapiro-Wilk tests of normality. When normality could be assumed, dependent-measures t tests were conducted. When normality could not be assumed, nonparametric Wilcoxon signed-rank tests were conducted. Effect sizes were calculated using Cohen's d for correlated measurements d_z, which is interpreted similarly to Cohen's d in between-subjects designs [83].

To examine whether attempting the problem before viewing the solution affects the relationship between student ability and the accuracy of metacognitive judgments, three one-way analyses of covariance (ANCOVAs) were conducted on the metacognitive bias scores. Homogeneity of regression (i.e., that the covariate has the same effect within each group) was assessed as in Ref. [84], and could be assumed except where noted. Effect sizes for the ANCOVAs were calculated using partial eta squared η_p^2, which represents the proportion of the total variance uniquely accounted for by each variable [83]. Partial eta squared values of 0.01, 0.06, and 0.14 are considered small, medium, and large, respectively [85]. Effect sizes for the simple regressions were calculated using the square correlation coefficient R^2, which represents the proportion of the total variance accounted for by the independent variable.

E. Results

1. Learning from ANSVs

The mean performance on the pretest and post-test for both conditions is given in Table IV. To investigate whether students learned from viewing the ANSVs a nonparametric Wilcoxon signed-rank test was conducted. The results indicate that students who completed the pretest scored about 24 percentage points higher on the post-test for calculation problems ($S = 165.50$, $p < 0.001$, $d_z = 1.17$).

2. Effect of attempting problems before viewing ANSVs

Although the individuals were randomly assigned to conditions, potential differences in prior physics ability was investigated using the exam average from the first two

TABLE IV. Mean course exam, performance, RCJs, and JOLs on the pre- and postassessments by condition. Note that attempt first ($N = 26$), view only ($N = 34$).

	Attempt first	View only
Measure	$M \pm$ SE	$M \pm$ SE
Exam 1/2 AVG	62.2 ± 2.43	61.9 ± 2.04
Exam 3	67.0 ± 3.28	67.5 ± 3.17
Pretest		
Performance	30.8 ± 4.94	... ...
Confidence	44.4 ± 4.84	... ...
Post-test performance		
Similar calculational	55.0 ± 5.77	60.9 ± 3.69
Conceptual transfer	39.4 ± 4.08	40.1 ± 3.57
Post-test RCJs		
Similar calculational	52.4 ± 5.24	53.1 ± 2.59
Conceptual transfer	66.0 ± 5.24	65.7 ± 1.99
JOLs	78.4 ± 3.08	65.4 ± 2.66

course exams. A nonparametric Kruskal-Wallis test failed to detect a difference between the conditions in prior physics ability, $\chi^2(1) = 0.03$, $p = 0.86$, indicating that the conditions were of approximately similar ability. Mean scores on the pre- and post-assessments for both groups are given in Table IV. No differences were found between the conditions for either the similar calculation problems, $t(58) = 0.89$, $p = 0.37$, or the conceptual transfer problems, $t(58) = 0.12$, $p = 0.91$. This suggests that the observed learning gains were largely due to learning from the ANSVs rather than engaging in problem solving with feedback. In addition, given the low scores on the pretest and the difficulty of the material, this result may suggest that attempting problems that are outside of students' zones of proximal development [86] may not be facilitative. In other words, when students are not able make much progress in solving problems, students may not receive much benefit from attempting to solve the problem before viewing the video.

3. Effect of problem attempts and ability on metacognitive judgments

JOLs.—On average, participants in the attempt first condition were overconfident by about 23 percentage points compared to a four-and-a-half percentage point overconfidence in the view only condition (Table V). The distribution of the bias scores for the JOLs was relatively normal and the assumption of homogeneity of regression held. Therefore, a one-way ANCOVA with JOL bias score as the response variable, condition as the between-subjects variable, and the average exam grade as the covariate was conducted. The results indicate that participants in the attempt first condition were more overconfident in making JOLs than participants in the view only condition, $F(1, 57) = 11.70$, $p = 0.001$, $\eta_p^2 = 0.15$, even though they were largely unsuccessful at solving the initial problems. In addition, participants with lower exam averages were more overconfident in making JOLs than participants with higher exam averages, $F(1, 57) = 9.89$, $p = 0.003$, $\eta_p^2 = 0.13$, consistent with the Kruger-Dunning effect.

TABLE V. Mean RCJ, and JOL bias by condition. Note that attempt first ($N = 26$), view only ($N = 34$).

	Attempt first	View only
Measure	$M \pm$ SE	$M \pm$ SE
JOLs	23.4 ± 5.14	4.6 ± 3.43
Similar RCJs	-2.6 ± 4.73	-7.8 ± 3.24
Transfer RCJs	26.6 ± 4.53	25.7 ± 3.70

RCJs.—For the similar calculation problems, participants in the attempt first condition were underconfident by about 2 percentage points on average, compared to an eight-percentage point underconfidence in the view only condition (Table V). The distribution of the bias scores for the RCJs was relatively normal and the assumption of homogeneity of regression held. Therefore, a one-way ANCOVA with RCJ bias score as the response variable, condition as the between-subjects variable, and the average exam grade as the covariate was conducted. The results failed to detect a difference between the conditions in the accuracy of the RCJs for similar problems, $F(1, 57) = 1.05$, $p = 0.31$, $\eta_p^2 = 0.02$. However, participants with lower exam averages were more overconfident in making JOLs than participants with higher exam averages, $F(1, 57) = 7.50$, $p < 0.01$, $\eta_p^2 = 0.11$, consistent with the Kruger-Dunning effect. In other words, the overconfidence of the attempt first group displayed after viewing the ANSVs was largely not present after attempting to solve the new problems.

For the conceptual transfer problems, participants were overconfident in both conditions by more than 25 percentage points. The distribution of the bias scores for the RCJs was relatively normal, however, the assumption of homogeneity of regression did not hold. This indicates that the effect of physics ability on the RCJs differed between conditions. Therefore, to examine the effect of attempting to solve the problems before viewing the ANSVs, an independent samples t tests was conducted on the bias scores. The difference between conditions was not significant, $t(58) = 0.16$, $p = 0.87$. To examine the effect of ability on the accuracy of RCJs, a simple linear regression was conducted for each condition with RCJs bias score as the dependent variable, and the average exam grade as the independent variable. For participants in the attempt first condition, the accuracy of the transfer problem RCJs was not related to physics ability, $\beta = 0.26$, $F(1, 24) = 0.47$, $p = 0.50$, $R^2 = 0.02$. Conversely, higher ability participants were more accurate in the accuracy of the transfer problem RCJs than lower ability participants in the view

only condition, $\beta = -0.68$, $F(1, 32) = 5.28$, $p = 0.03$, $R^2 = 0.14$. This suggests that attempting to solve problems before viewing ANSVs may have different metacognitive effects for solving conceptual transfer problems for students at different ability levels. However, because this effect was unexpected, the results should be considered preliminary and future studies should attempt to replicate this finding.

IV. EXPERIMENT 3

It was expected that the experience of attempting to solve the problem before viewing the solution would allow students to identify the components of the problem that they did not understand, thus allowing them to focus on learning the content applicable to their needs. Contrary to expectations, students who attempted to solve the problems before viewing the solutions did not score higher on the post-test than students who only viewed the solutions. However, the material covered by the ANSVs in experiment 2 is typically some of the most difficult material in the course for students to master. Given the low scores on the pretest and the difficulty of the material, the lack of a difference on the post-test between the conditions may suggest that having students attempt problems that they are initially unable to make progress in solving is not beneficial for learning. Experiment 3 explores this possibility by using participants with a wider ability range, and covering the material used in experiment 1, which is typically easier for students to learn.

A. Participants

An email was sent to all students enrolled in the Fall 2018 introductory mechanics course. Seventy-four students volunteered and were randomly assigned either to attempt to solve calculational physics problems before viewing the solution videos or to only watch the solution videos without attempting the problems. Of these students, 49 students completed both sessions—21 in the attempt-problem-first condition, and 28 in the view the solution only condition. Only the data from the participants who completed both sessions were included in the data analysis.

B. Procedure

The procedures for each condition were identical to the procedures described in experiment 2. Student performance was assessed using a pre-test and a post-test consisting of six calculational and five conceptual physics problems covering center of mass, conservation of energy, conservation of momentum, and work. All questions can be found in Ref. [77]. Scores on the first course exam were used as a measure of physics ability in this experiment.

In addition, students viewed solutions for both calculational and conceptual problems as in experiment 1. The procedures for scoring the calculational and conceptual problems were the same as in experiments 1 and 2. All of the questions were scored by two independent graders with an initial interrater agreement of 96%. Following discussion, 100% agreement was reached.

One student did not make a JOL after viewing the videos during one session. This student was not included in the JOL analyses, but was included in all other analyses. While students in general provided RCJs after attempting every problem on the pre-tests and post-tests, about 25% of students failed to make an RCJ on at least one question. Only 2.8% of the individual questions did not have RCJs, and the percentage of missing RCJs ranged from 0% for most questions to 6% for four questions on the post-test. Because there were no patterns in the missing data, nor did missingness correlate with any dependent variable used in the study, the data were assumed to be missing at random. A single RCJ score was computed for all participants by calculating the mean of the individual confidence judgments.[3]

C. Results

Descriptive statistics for grades on the first course exam, performance on the pre-tests and post-tests, JOLs, and RCJs are given in Table VI. The distribution of scores on the first course exam was normally distributed, therefore independent-samples t tests were performed to investigate differences between the conditions. No difference was found between the conditions for physics ability, $t(47) = 0.51$, $p = 0.61$, $d = 0.15$. To investigate the effect of attempting problems before viewing the ANSVs for different problem types, the calculational and conceptual questions were investigated separately. Distributions for both calculational and conceptual problems were not normally distributed, therefore Kruskal-Wallis tests were performed to investigate differences between the conditions. The conditions did not differ for conceptual problems, $\chi^2(1) = 0.68$, $p = 0.41$, $d = 0.31$. However, participants in the attempt-first group scored marginally higher on the post-test calculational problems, $\chi^2(1) = 3.80$, $p = 0.051$, $d = 0.49$.

Descriptive statistics for the JOL and RCJ bias scores for both conditions are given in Table VII. The distribution of the JOL and RCJ bias scores were normally distributed, and homogeneity of regression could be assumed. Therefore, to investigate how attempting to solve problems before viewing ANSVs affected the accuracy of JOLs and RCJs, two ANCOVAs were conducted with the bias scores as the

[3]Because the data can be assumed to be missing at random, and only 75% of the participants made RCJs for every question, multiple imputation procedures are more appropriate to address the missing RCJ data. This analysis was also conducted, however, because these results led to identical conclusions, the simpler analysis is presented here. For interested readers, the analysis using multiple imputation can be found in Ref. [77].

TABLE VI. Mean performance and RCJs on the pre and postassessments by condition. Note that attempt first ($N = 21$), view only ($N = 28$).

	Attempt first	View only
Measure	$M \pm$ SE	$M \pm$ SE
First course exam	72.2 ± 3.51	69.9 ± 2.84
Pretest calculational problems		
Performance	58.7 ± 6.15	··· ···
RCJ	60.1 ± 4.14	··· ···
Pretest conceptual problems		
Performance	67.6 ± 6.09	··· ···
RCJ	68.3 ± 3.57	··· ···
JOLs	86.5 ± 4.09	81.7 ± 2.66
Post-test calculational problems		
Performance	90.5 ± 2.95	83.9 ± 2.49
RCJ	79.8 ± 3.22	76.7 ± 3.00
Post-test conceptual problems		
Performance	75.2 ± 4.34	70.0 ± 3.78
RCJ	83.3 ± 2.31	79.7 ± 3.33

between-subjects variable and the score from the course exam as the covariate.

Two outliers were identified, one in the attempt-first condition whose JOL bias was more than 30 percentage points below the next lowest bias, and one in the view only condition whose JOL bias was more than 20 percentage points above the next highest bias. These individuals were removed from the analysis of JOL accuracy,[4] but were not removed from the RCJ analyses. The ANCOVAs indicated that there was not a significant difference in JOL bias between the conditions, $F(1,44) = 0.76$, $p = 0.39$, $\eta_p^2 = 0.02$, nor was JOL accuracy related to physics ability, $F(1,44) = 0.41$, $p = 0.52$, $\eta_p^2 = 0.01$. Participants in both conditions were underconfident by about ten percentage points when making RCJs for calculational problems, and overconfident by almost ten percentage points when making RCJs for conceptual problems. The conditions did not differ in RCJ bias for calculational problems, $F(1,46) = 0.25$, $p = 0.62$, $\eta_p^2 = 0.01$, or for conceptual problems, $F(1,46) = 0.03$, $p = 0.86$, $\eta_p^2 < 0.01$. In addition, RCJ accuracy for calculational problems was related to physics ability, as participants with higher course exam scores were less overconfident in making RCJs than participants with lower exam scores, $F(1,46) = 4.12$, $p = 0.04$, $\eta_p^2 = 0.08$. However, RCJ accuracy for conceptual problems was not related to physics ability, $F(1,46) = 1.04$, $p = 0.31$, $\eta_p^2 = 0.02$.

[4]Including these individuals in the ANCOVA results in a heterogeneity of regression (i.e., the relationship between the dependent variable and the covariate differing between the conditions), which violates the assumptions for ANCOVA. Therefore, the individuals were removed from the ANCOVA analysis to avoid violating this assumption. However, the results of the ANCOVA are similar when these two individuals are included in the analysis.

TABLE VII. Mean RCJ bias and JOL bias on the post-test by condition. Note that attempt first ($N = 21$), view only ($N = 28$).

	Attempt first	View only
Measure	$M \pm$ SE	$M \pm$ SE
JOL bias	3.0 ± 4.2	4.1 ± 3.1
Overall RCJ bias	-2.5 ± 3.6	0.6 ± 3.4
Calculational RCJ bias	-10.7 ± 4.8	-7.2 ± 3.2
Conceptual RCJ bias	8.1 ± 4.2	9.7 ± 4.6

V. GENERAL DISCUSSION

Prior work has demonstrated that students like to learn how to solve calculational problems from studying worked examples during initial learning [17,33,87], and when reviewing for exams [41], especially when paired with problem solving [88]. This study extends the prior work by demonstrating that students learned from ANSVs for both calculational and conceptual physics problems. This study also demonstrates, in experiment 1, that students enrolled in a calculus-based introductory physics course relearn similarly from ANSVs that present a procedural approach to problem-solving while also modeling expertlike mental simulation and conceptual analysis, and from ANSVs that presents a high-level solution that derives equations from first principles. In addition, students were more accurate in monitoring their understanding as measured by lower RCJ bias. The similar student outcomes in experiment 1 suggests that students are likely to benefit from ANSVs that adhere to multimedia learning principles. The similarity between the attempt first and the view only conditions in experiments 2 and 3 suggest that these improvements were likely due to the information provided in the video solutions and not simply from attempting the problems and receiving correctness feedback.

Students who attempted to solve problems before viewing the ANSVs were overconfident in their JOLs for difficult material (experiment 2), but not for easier material (experiment 3). This is a potentially important finding for students within self-regulated learning environments, because material that receives high JOLs tends to be dropped from future studying [89]. It may be that attempting to solve problems before viewing ANSVs reduces the extrinsic cognitive load without increasing germane cognitive load, leading to greater feelings of fluency when viewing the ANSV than students who did not know the problem before viewing [58,90]. When working with difficult content, this increased fluency is not diagnostic of learning, which could lead students to become overconfident and result in an "illusion of understanding." This result parallels the findings from Refs. [63–65] that found

students are more confident after viewing more fluent lectures even though the enhanced fluency often does not lead to greater learning. When working with easier content the enhanced fluency is more diagnostic of learning, leading students to make more accurate judgments. Alternatively, students in the attempt-first condition in experiment 3 were not able to demonstrate higher overconfidence when making JOLs because the scores on the post-test were near ceiling. Future work should investigate the cause of the overconfidence by varying the difficulty of the problems shown in the ANSVs.

Finally, across all three experiments students were more overconfident in making RCJs for conceptual problems compared to calculational problems, even though they scored higher on the calculational problems on the post-tests. One possible reason for this pattern is that individuals often hold the belief that conceptual problems are easier to solve than calculational problems in physics [91]. In addition, students often hold robust misconceptions for conceptual questions for which the intuitive answer is not correct [92]. These findings suggest that it is important to incorporate worked examples for both calculational and conceptual problems to help students prepare for exams that require students to use both computational and conceptual solutions.

While students appear to benefit from worked examples in this study, the relatively small sample sizes suggest that the findings should be replicated before definitive conclusion are made. In addition, although prior research has found that students benefit from animated learning materials compared to static learning materials [42,43], the conditions and contexts in which animated solutions leads to better learning than static solutions is unclear and more research is needed [93]. This study also did not investigate how these students performed on the actual course exams, as it is unreasonable to expect to significantly help students prepare for exams using only a handful of questions across one or two sessions. To investigate the larger question of helping low-performing students prepare for exams, future research should investigate the use of ANSVs in an integrated course management environment. In addition, prior research has found that the extent to which students are able to learn from worked examples depends on how well they provide correct self-explanations of the solutions and the underlying conceptual understanding to themselves [94]. However, learners often do not spontaneously engage in productive self-explanation of the steps given in procedural solutions [95]. One reason for the lack of spontaneous self-explanation could be that novices and low-performing students do not have the baseline knowledge required for elaborate self-explanation. These underprepared students may benefit from solutions which make explicit the rationale, theories, and implicit thought processes that an expert utilizes when producing a solution. Future work should investigate the effect of methods using techniques that involve more active learning that can increase the germane cognitive load, such as providing students with solutions that require students to elaborate or self-explain the material [96,97], that engage in analogical comparison [22], or that use adaptive fading of worked examples [20,45].

Finally, prior work has demonstrated that providing students with instruction and training in making accurate self-assessment leads to better performance and greater monitoring accuracy [98,99], but providing students with only feedback on their accuracy can lead to greater overconfidence and less effective metacognitive control [100]. Given these findings, future work should investigate the effect that metacognitive training and feedback has on the ways in which students interact with solution videos.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation under Grant No. DRL 1252389. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[1] X. Chen, STEM Attrition: College Students' Paths into and out of STEM Fields, Statistical Analysis Report No. NCES 2014–001 (National Center for Education Statistics, College Park, MD, 2013).

[2] S. T. Fiske, J. G. Cromley, T. Perez, and A. Kaplan, Undergraduate STEM achievement and retention: Cognitive, motivational, and institutional factors and solutions, Policy Insights from Behavior. Brain Sci. **3**, 4 (2016).

[3] B. King, Changing college majors: Does it happen more in STEM and do grades matter?, J. Coll. Sci. Teach. **44**, 44 (2015).

[4] R. N. Blasiman, J. Dunlosky, and K. A. Rawson, The what, how much, and when of study strategies: Comparing intended versus actual study behavior, Memory **25**, 784 (2017).

[5] M. K. Hartwig and J. Dunlosky, Study strategies of college students: Are self-testing and scheduling related to achievement?, Psychon. Bull. Rev. **19**, 126 (2012).

[6] J. D. Karpicke, A. C. Butler, and H. L. Roediger III, Metacognitive strategies in student learning: Do students practice retrieval when they study on their own?, Memory **17**, 471 (2009).

[7] N. Kornell and R. A. Bjork, The promise and perils of self-regulated study, Psychon. Bull. Rev. **14**, 219 (2007).
[8] A. C. Butler and H. L. Roediger III, Testing improves retention in a simulated classroom, European J. Cogn. Psychol. **19**, 514 (2007).
[9] J. L. Little, E. L. Bjork, R. A. Bjork, and G. Angello, Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting, Psychol. Sci. **23**, 1337 (2012).
[10] K. B. McDermott, P. K. Agarwal, L. D'Antonio, H. L. Roediger III., and M. A. McDaniel, Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes, J. Experimental Psychol., Applied **20**, 3 (2014).
[11] L. E. Richland, L. S. Kao, and N. Kornell, Can unsuccessful tests enhance learning? in *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (Cognitive Science Society, Austin, TX, 2008), p. 2338–2343.
[12] H. L. Roediger III and A. C. Butler, The critical role of retrieval practice in long-term retention, Trends Cognit. Sci. **15**, 20 (2011).
[13] J. Geller, A. R. Toftness, P. I. Armstrong, S. K. Carpenter, C. L. Manz, C. R. Coffman, and M. H. Lamm, Study strategies and beliefs about learning as a function of academic achievement and achievement goals, Memory **26**, 683 (2018).
[14] R. A. Gurung, J. Weidert, and A. Jeske, Focusing on how students study, J. Scholarship Teach. Learn. **10**, 28 (2010).
[15] S. Kalyuga, P. Chandler, J. Tuovinen, and J. Sweller, When problem solving is superior to studying worked examples, J. Educ. Psychol. **93**, 579 (2001).
[16] F. Nievelstein, T. van Gog, G. van Dijck, and H. P. A. Boshuizen, The worked example and expertise reversal effect in less structured tasks: Learning to reason about legal cases, Contemp. Educ. Psychol. **38**, 118 (2013).
[17] J. Sweller and G. A. Cooper, The use of worked examples as a substitute for problem solving in learning algebra, Cognit. Instr. **2**, 59 (1985).
[18] J. L. Booth, K. M. McGinn, L. K. Young, and C. Barbieri, Simple practice doesn't always make perfect: Evidence from the worked example effect, Policy Insights from Behavior. Brain Sci. **2**, 24 (2015).
[19] R. Schwonke, A. Renkl, C. Krieg, J. Wittwer, V. Aleven, and R. Salden, The worked-example effect: Not an artefact of lousy control conditions, Computers Human Behav. **25**, 258 (2009).
[20] J. C. M. R. 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).
[21] I. Glogger-Frey, C. Fleischer, L. Gruny, J. Kappich, and A. Renkl, Inventing a solution and studying a worked solution for learning from direct instruction, Learning Instr. **39**, 72 (2015).
[22] 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).
[23] M. T. H. Chi, Constructing self-explanations and scaffolded explanations in tutoring, Appl. Cogn. Psychol. **10**, S33 (1996).
[24] S. Bokosmaty, J. Sweller, and S. Kalyuga, Learning geometry problem solving by studying worked examples: Effects of learner guidance and expertise, Am. Educ. Res. J. **52**, 307 (2015).
[25] G. Cooper, S. Tindall-Ford, P. Chandler, and J. Sweller, Learning by imagining, J. Exp. Psychol. Appl. **7**, 68 (2001).
[26] S. Kalyuga and A. Renkl, Expertise reversal effect and its instructional implications: Introduction to the special issue, Instruc. Sci. **38**, 209 (2010).
[27] J. Leppink, N. J. Broers, T. Imbos, C. P. M. van der Vleuten, and M. P. F. Berger, Self-explanation in the domain of statistics: An expertise reversal effect, Higher Educ. **63**, 771 (2012).
[28] D. S. McNamara, E. Kintsch, N. B. Songer, and W. Kintsch, Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text Cognit. Instr. **14**, 1 (1996).
[29] J. Sweller, P. Ayres, and S. Kalyuga, *Cognitive Load Theory* (Springer, New York, NY, 2011).
[30] J. Sweller, Cognitive load theory: Recent theoretical advances, in *Cognitive Load Theory*, edited by J. L. Plass, R. Moreno, and R. Brunken (Cambridge University Press, Cambridge, New York, 2010), pp. 29–47.
[31] J. Sweller, Cognitive load during problem solving: Effects on learning, Cogn. Sci. **12**, 257 (1988).
[32] M. Ward and J. Sweller, Structuring effective worked examples, Cognit. Instr. **7**, 1 (1990).
[33] R. Tarmizi and J. Sweller, Guidance during mathematical problem solving, J. Educ. Psychol. **80**, 424 (1988).
[34] F. Paas, J. E. Tuovinen, J. J. G. van Merrienboer, and A. A. Darabi, A motivational perspective on the relation between mental effort and performance: Optimizing learner involvement in instruction, Educ. Technology Res. Develop. **53**, 25 (2005).
[35] S. Kalyuga, P. Ayres, P. Chandler, and J. Sweller, The expertise reversal effect, Educ. Psychol. **38**, 23 (2003).
[36] T. van Gog, L. Kester, and F. Paas, Effects of worked examples, example-problems, and problem-example pairs on novices' learning, Contemp. Educ. Psychol. **36**, 212 (2011).
[37] A. D. Smith, J. P. Mestre, and B. H. Ross, Eye-gaze patterns as students study worked-out examples in mechanics, Phys. Rev. ST Phys. Educ. Res. **6**, 020118 (2010).
[38] I. Belski and R. Belski, Impact of dynamic (videotaped) worked examples on knowledge transfer, in *Proceedings of the 24th Annual Conference of the Australasian Association for Engineering Education-AAEE2013*, edited by C. Lemckert, G. Jenkins, and S. Lang-Lemckert (Griffith University, Nathan, Australia, 2013), p. 3A2.
[39] I. Belski and R. Belski, Student-generated dynamic worked examples as videos to enhance learning in STEM, in *Student-generated Digital Media in Science Education: Learning, Explaining and Communicating Content*, edited by G. Hoban, W. Nielsen, and A. Shepherd (Routledge, New York, NY, 2016), p. 122–135.

[40] G. Gladding, B. Gutmann, N. Schroeder, and T. Stelzer, Clinical study of student learning using mastery style versus immediate feedback online activities, Phys. Rev. ST Phys. Educ. Res. **11**, 010114 (2015).

[41] J. P. Mestre, J. W. Morphew, and G. E. Gladding, Learning from different styles of narrated-animated solutions among low-performing students, in *Proceedings of the 2015 Physics Education Research Conference, College Park, MD*, edited by A. D. Churukian, D. L. Jones, and L. Ding (AIP, New York, 2015), pp. 223–226.

[42] M. M. Lusk and R. K. Atkinson, Animated pedagogical agents: Does their degree of embodiment impact learning from static or animated worked examples?, Appl. Cogn. Psychol. **21**, 747 (2007).

[43] T. N. Hoffler and D. Leutner, Instructional animation versus static pictures: A meta-analysis, Learn. Instr. **17**, 722 (2007).

[44] S. Kalyuga and A. Renkl, Expertise reversal effect and its instructional implications: Introduction to the special issue, Instr. Sci. **38**, 209 (2010).

[45] 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).

[46] J. ter Vrugte, T. de Jong, S. Vandercruysse, P. Wouters, H. van Oostendorp, and J. Elen, Computer game-based mathematics education: Embedded faded worked examples facilitate knowledge acquisition, Learn. Instr. **50**, 44 (2017).

[47] K. J. Crippen and B. L. Earl, The impact of web-based worked examples and self-explanation on performance, problem solving, and self-efficacy, Comput. Educ. **49**, 809 (2007).

[48] R. A. Bjork, Assessing our own competence: Heuristics and illusions, in *Attention and Performance XVII: Cognitive Regulation of Performance: Interaction of Theory and Application*, edited by D. Gopher and A. Koriat (MIT Press, Cambridge, MA, 1999), p. 435–459.

[49] C. M. Mills and F. C. Keil, Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth, J. Exp. Child Psychol. **87**, 1 (2004).

[50] R. Ariel, J. Dunlosky, and H. Bailey, Agenda-based regulation of study-time allocation: When agendas override item-based monitoring, J. Exp. Psychol. Gen. **138**, 432 (2009).

[51] J. Metcalfe and N. Kornell, A region of proximal learning model of study time allocation, J. Memory Lang. **52**, 463 (2005).

[52] J. Dunlosky and K. W. Thiede, Metamemory, in *The Oxford Handbook of Psychology*, edited by D. Reisberg (Oxford University Press, Oxford, UK, 2013), pp. 283–298.

[53] A. D. Castel, A. S. Benjamin, F. I. Craik, and M. J. Watkins, The effects of aging on selectivity and control in short-term recall, Memory Cognit. **30**, 1078 (2002).

[54] N. Kornell, M. G. Rhodes, A. D. Castel, and S. K. Tauber, The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments, Psychol. Sci. **22**, 787 (2011).

[55] R. Ariel and J. Dunlosky, The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting, Memory Cognit. **39**, 171 (2011)..

[56] D. L. Dinsmore and M. M. Parkinson, What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration, Learn. Instr. **24**, 4 (2013).

[57] R. Jersakova, R. J. Allen, J. Booth, C. Souchay, and A. R. O'Connor, Understanding metacognitive confidence: Insights from judgment-of-learning justifications, J. Memory Lang. **97**, 187 (2017).

[58] A. Koriat, R. Nussinson, and R. Ackerman, Judgments of learning depend on how learners interpret study effort, J. Exper. Psychol., Learning, Memory, Cogn. **40**, 1624 (2014).

[59] R. A. Schmidt and R. A. Bjork, New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training, Psychol. Sci. **3**, 207 (1992).

[60] N. Kornell and L. K. Son, Learners' choices and beliefs about self-testing, Memory **17**, 493 (2009).

[61] M. G. Rhodes and A. D. Castel, Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions, J. Exper. Psychol., General **137**, 615 (2008).

[62] C. L. Yue, A. D. Castel, and R. A. Bjork, When disfluency is—and is not—a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory, Memory Cognit. **41**, 229 (2013).

[63] S. K. Carpenter, M. M. Wilford, N. Kornell, and K. M. Mullaney, Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning, Psychon. Bull. Rev. **20**, 1350 (2013).

[64] M. J. Serra and D. A. Magreehan, Instructor fluency correlates with students' ratings of their learning and the instructor in an actual course, Creative Educ. **7**, 1154 (2016).

[65] A. R. Toftness, S. K. Carpenter, J. Geller, S. Lauber, M. Johnson, and P. I. Armstrong, Instructor fluency leads to higher confidence in learning, but not better learning, Metacognition Learn. **13**, 1 (2018).

[66] F. M. Zaromb, J. D. Karpicke, and H. L. Roediger III, Comprehension as a basis for metacognitive judgments: Effects of effort after meaning on recall and metacognition, J. Exper. Psychol., Learning, Memory, Cogn. **36**, 552 (2010).

[67] L. M. Reder and F. E. Ritter, What determines initial feelings of Knowing? Familiarity with question terms, not with the answer, J. Exper. Psychol., Learning, Memory, Cogn. **18**, 435 (1992).

[68] R. Ackerman and H. Zalmanov, The persistence of the fluency-confidence association in problem solving, Psychon. Bull. Rev. **19**, 1187 (2012).

[69] A. Koriat and H. Ma'ayan, The effects of encoding fluency and retrieval fluency on judgments of learning, J. Memory Lang. **52**, 478 (2005).

[70] A. S. Benjamin, R. A. Bjork, and B. L. Schwartz, The mismeasure of memory: When retrieval fluency is misleading as a metacognitive index, J. Exper. Psychol. General **127**, 55 (1998).

[71] R. E. Mayer, Multimedia learning, Psychol. Learn. Motivation **41**, 85 (2002).
[72] R. E. Mayer and R. Moreno, Nine ways to reduce cognitive load in multimedia learning, Educ. Psychol. **38**, 43 (2003).
[73] https://bit.ly/2Yr3wFG.
[74] J. L. Docktor, N. E. Strand, J. P. Mestre, and B. H. Ross, Conceptual problem solving in high school physics, Phys. Rev. Phys. Educ. Res. **11**, 020106 (2015).
[75] B. B. de Koning, H. K. Tabbers, R. M. J. P. Rikers, and F. Paas, Attention cueing as a means to enhance learning from an animation, Appl. Cogn. Psychol. **21**, 731 (2007).
[76] B. B. de Koning, H. K. Tabbers, R. M. J. P. Rikers, and F. Paas, Attention guidance in learning from complex animation: Seeing is understanding?, Learn. Instr. **20**, 111 (2010).
[77] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.16.010104 for all of the questions used in the study as well as the analyses using multiple imputation methods to deal with missing data.
[78] M. G. Rhodes, Judgements of learning: Methods, data, and theory, in *The Oxford Handbook of Metamemory*, edited by J. Dunlosky and S. K. Tauber (Oxford University Press, New York, 2015), pp. 65–80.
[79] G. Shraw, Measuring metacognitive judgments, in *Handbook of Metacognition in Education*, edited by D. J. Hacker, J. Dunlosky, and A. C. Graesser (Routledge, New York, NY, 2009), pp. 415–429.
[80] B. Finn and J. Metcalfe, The role of memory for past test in the underconfidence with practice effect, J. Exper. Psychol., Learning, Memory, Cogn. **33**, 238 (2007).
[81] A. Koriat, L. Sheffer, and H. Ma'ayan, Comparing objective and subjective learning curves: Judgment of learning exhibit increased underconfidence-with-practice, J. Exper. Psychol., General **131**, 147 (2002).
[82] M. J. Serra and J. Dunlosky, Does retrieval fluency contribute to the underconfidence-with-practice effect?, J. Exper. Psychol., Learning, Memory, Cogn. **31**, 1258 (2005).
[83] D. Lakens, Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-test and ANOVAs, Frontiers Psychol. **4**, 1 (2013).
[84] G. Keppel and T. D. Wickens, *Design and Analysis: A Researcher's Handbook*, 4th ed. (Pearson Education, Upper Saddle River, NJ, 2004).
[85] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Erlbaum, Hillsdale, NJ, 1988).
[86] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes* (Harvard University Press, Cambridge, MA, 1980).
[87] M. M. Recker and P. Pirolli, Modeling individual differences in students' learning strategies, J. Learn. Sci. **4**, 1 (1995).
[88] J. G. Trafton and B. J. Reiser, Studying examples and solving problems: Contributions to skill acquisition, in *Proceedings of the 15th conference of the Cognitive Science Society*, edited by M. Polson (Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1993), pp. 1017–1022.
[89] J. Metcalfe and B. Finn, Evidence that judgments of learning are causally related to study choice, Psychon. Bull. Rev. **15**, 174 (2008).
[90] A. Koriat, Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning, J. Exper. Psychol., General **126**, 349 (1997).
[91] W. Fakcharoenphol, J. W. Morphew, and J. P. Mestre, Judgments of physics problem difficulty among experts and novices, Phys. Rev. ST Phys. Educ. Res. **11**, 020128 (2015).
[92] J. L. Docktor and J. P Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).
[93] R. E. Mayer, M. Hegerty, S. Mayer, and J. Campbell, When static media promote active learning: Annotated Illustrations versus narrated animations in multimedia instruction, J. Exper. Psychol., Applied **11**, 256 (2005).
[94] 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).
[95] P. Gerjets, K. Scheiter, and R. Carambone, Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations?, Learn. Instr. **16**, 104 (2006).
[96] A. Renkl, R. Stark, H. Gruber, and H. Mandl, Learning from worked-out examples: The effects of example variability and elicited self-explanations, Contemp. Educ. Psychol. **23**, 90 (1998).
[97] J. J. G. van Merrienboer, J. G. Schuurman, M. B. M. De Croock, and F. Paas, Redirecting learners' attention during training: Effects on cognitive load, transfer test performance and training, Learn. Instr. **12**, 11 (2002).
[98] R. Azevedo and J. G. Cromley, Does training on self-regulated learning facilitate students' learning with hypermedia?, J. Educat. Psychol. **96**, 523 (2004).
[99] D. Kostons, T. van Gog, and F. Paas, Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning, Learn. Instr. **22**, 121 (2012).
[100] S. F. Raaijmakers, M. Baars, F. Paas, J. J. G. van Merrienboer, and T. van Gog, Effects of self-assessment feedback on self-assessment and task-selection accuracy, Metacognit. Learn. **14**, 21 (2019).