Judgments of physics problem difficulty among experts and novices

Witat Fakcharoenphol,[1,2] Jason W. Morphew,[1,3] and José P. Mestre[1,3]
[1]Department of Physics, University of Illinois Urbana-Champaign, Illinois, USA
[2]Faculty of Education and Development Sciences, Kasetsart University, Kamphaeng Saen Campus, Thailand
[3]Department of Educational Psychology, University of Illinois, Urbana-Champaign, Illinois, USA
(Received 1 June 2015; published 23 October 2015)

Students' ability to effectively study for an exam, or to manage their time during an exam, is related to their metacognitive capacity. Prior research has demonstrated the effective use of metacognitive strategies during learning and retrieval is related to content expertise. Students also make judgments of their own learning and of problem difficulty to guide their studying. This study extends prior research by investigating the accuracy of novices' and experts' ability to judge problem difficulty across two experiments; here "accuracy" refers to whether or not their judgments of problem difficulty corresponds with actual exam performance in an introductory mechanics physics course. In the first experiment, physics education research (PER) experts judged the difficulty of introductory physics problems and provided the rationales behind their judgments. Findings indicate that experts use a number of different problem features to make predictions of problem difficulty. While experts are relatively accurate in judging problem difficulty, their content expertise may interfere with their ability to predict student performance on some question types. In the second experiment novices and "near experts" (graduate TAs) judged which question from a problem pair (taken from a real exam) was more difficult. The results indicate that judgments of problem difficulty are more accurate for those with greater content expertise, suggesting that the ability to predict problem difficulty is a trait of expertise which develops with experience.

DOI: 10.1103/PhysRevSTPER.11.020128 PACS numbers: 01.40.-d, 01.55.+b

I. INTRODUCTION

The method that a student utilizes when preparing for an exam, or while taking an exam, is influenced by the students' metacognitive abilities. The ability to effectively learn and recall information relies on students' ability to make judgments about what they know or do not know, and what types of problems tend to be more difficult to solve. Numerous studies have investigated how well people perform tasks and how well they make predictions about their ability to perform these tasks. Much of the research has focused on memory tasks using either matched pair recall tasks or reading comprehension tasks (See Refs. [1] and [2] for a review). Engaging in effective metacognitive strategies during learning and retrieval has been found to be related to one's domain knowledge [3,4]. For example, Griffen, Lee, and Wiley [5] had undergraduates read five different texts about baseball, and then predict how many questions out of five they would get correct on a post test for each passage. They found that expertise was related to the accuracy of their predictions. That is, those more knowledgeable about baseball were more accurate in their metacognitive predictions and made more effective use of domain knowledge in predicting their performance.

In one study looking at predicting problem difficulty in physics, Gire and Rebello [6] asked undergraduate students and instructors to rate the difficulty of kinematics and work energy questions on a scale of 1 to 10. In addition, students were asked to rate the familiarity of the problems and to solve the problems. They found that while students and instructors were similar in many of their predictions, they differed on a subset of questions. Overall they were unable to discern a pattern to explain the difference in difficulty ratings, but did note that students rated context-rich problems as easier than instructors for work-energy problems, but not for kinematics problems. Further, the difficulty predictions made by instructors correlated more strongly to student performance on the questions and to measures of problem complexity than did the student predictions.

Previous studies have investigated physics students' ability to predict and postdict performance. These studies have generally found that students overestimate their own performance on exams, with the overestimates being more pronounced for low-performing students [7]. This is likely due to students both overestimating their own ability [8] and underestimating the difficulty of the problems on the exam. Research employing cued recall tasks has found that students use judgments of learning and problem difficulty to select items for additional study, focusing on items

judged to be more difficult even in the absence of time pressure [9]. This strategy is generally availing as individuals who spend more time studying problems they judge to be more difficult tend to outperform those who study items in other predetermined sequences [10,11]. However, this advantage only holds for those students whose judgments of problem difficulty were aligned with normative measures of difficulty [12]. Together these studies indicate that judgments of problem difficulty affect how students allocate study time even in the absence of actual differences in normative difficulty.

Individuals engage in various metacognitive strategies to estimate the difficulty of tasks. For memory tasks they practice retrieving information in order to make a judgment about whether or not they will remember something in the future. For problem solving tasks they engage in solving practice problems to determine how easy it is for them to retrieve the relevant information and complete the solution process. Perhaps surprisingly, many students make judgments of their performance using factors such as their overall study time, or perception of their skill level, rather than their performance on practice problems [13,14]. Individuals also tend to use their subjective experience when making predictions concerning problem difficulty [15]. However, when an individual is selecting which problems she should solve when studying, she is unlikely to solve all of the problems available to her due to time constraints. In addition, when individuals are under time pressure, such as on an exam, they have to rely on a set of implicit heuristics of what makes problems difficult [16,17]. Further, when individuals know the solution to a problem, then they are no longer able to use subjective experience effectively to make predictions of difficulty. Rather they must rely on analytic alternatives, such as implicit (or explicit in some cases) theories about what makes problems within the domain difficult. However, to develop a good theory concerning problems' difficulty requires the ability to recognize features of a problem that other difficult problems share [15]. In addition, individuals need to develop methods to make predictions when the information provided by these features are in conflict. For example, one feature (e.g., the question requires difficult calculations) may indicate that a problem will be difficult, while another feature (e.g., the calculations are familiar) might indicate that it will be easier.

One strategy that individuals may overuse is domain familiarity when assessing their learning. Shanks and Serra [18] found that participants use familiarity as a source of information when estimating the level of difficulty of recalling a previously studied fact. This information was used in the selection of which facts to restudy, but not in determining the overall time one spends studying. Jacoby and Kelly [15] had students solve anagrams which were either familiar, anagrams which they had previously studied, or new anagrams. These anagrams were either presented alone or paired with a solution. Participants rated previously solved anagrams as easier than newly presented anagrams. In addition, participants who did not have access to the solutions predicted the normative difficulty level more accurately than those who had solutions. This study suggests that individuals tend to use subjective experience to make predictions of difficulty. However, when deprived on the subjective experience through prior knowledge of the solutions, individuals tend to use familiarity to make judgments of problem difficulty.

The use of familiarity as a cue for making predictions of problem difficulty is often a valid strategy, as familiarity and fluency are often correlated with learning. However, the features of a problem that cues familiarity will determine the effectiveness of this strategy. People often use familiarity with terms in the question in their initial judgments of difficulty as opposed to knowledge of the answer in a recall task or knowledge of the procedure in a problem solving task [19]. However, the use of cue familiarity as a predictor is more frequent under conditions where students experience time pressure such as an exam [16].

An interesting yet unanswered question is whether or not introductory physics students are able to accurately judge which problems from typical midterm exams are more difficult. Difficulty is operationalized here as the average score on problems given to large numbers of students to solve, as typically occurs in large-course, multiple-choice exams. Ability to discern the difficulty level of a problem carries important ramifications for both students and instructors. For students taking an exam within a finite time frame, ability to tell which problems are easier can help budget time efficiently—by solving the easy problems first, the student picks "low hanging fruit" thereby accruing points with relatively modest effort. After solving the easy problems the student can then devote the remaining time to tackling the harder problems. In contrast, taking a linear strategy through an exam likely results in students spending a long time solving the hard problems encountered along the way (especially if the student becomes stuck on one or more of them), which could result in not leaving enough time to solve the more manageable problems. Perhaps a major contributing factor for students who perform poorly on exams is one of poor judgment of problem difficulty leading to poor time management. In addition, judgments about problem difficulty may affect the choice of problem solving strategies that students employ [17]. For example, students might use different problem solving strategies on questions they assume are easier, relying on equation hunting or other intuitive strategies for problems they view as easier.

Instructors tacitly use their ability to ascertain problem difficulty in teaching problem solving and in constructing exams. In the former case, instructors tend to scaffold the teaching of problem solving, first picking easier problems to demonstrate techniques and then following with more

difficult problems. In exam construction, a desirable feature of a good exam is to have a range of problem difficulties in order to properly assess students' competence. Misjudgments by instructors resulting in more "easy" problems on an assessment could result in a very high average, thus not providing useful feedback to the instructor or to the students on the latter's level of competence with the material. Conversely, a very difficult exam results in a very low average, which is only useful for selecting out the high achieving students and discouraging the rest.

The two experiments in this study explore the ability of experts and novices to make predictions about the difficulty of physics problems. This study is focused on three main questions. First, what criteria do experts use to make judgments concerning problem difficulty? Second, how accurate are both experts and novices in judging problem difficulty? Third, does the ability to judge problem difficulty improve with experience or content expertise? In the first experiment, expert judgments of problem difficulty from previously administered exam questions was collected, which allowed us to ascertain a numerical assessment of this skill (percent correct judgments). These experts also indicated their rationale for their judgments. Thus, for this set of experts, not only were we able to ascertain the accuracy of their judgments of problem difficulty, but also the extent to which differences among experts' accuracy were due to different judgment strategies or criteria. In short, we could shed light on whether or not experts were homogeneous in terms of using similar judgment strategies or criteria as well as how those strategies or criteria impacted their judgments. In addition, experts' content knowledge may affect their judgment of problem difficulty. This study also sheds light on the accuracy of the tacit and explicit strategies that experts employ when making judgments of problem difficulty.

In the second experiment, judgments of problem difficulty from a subset of items used in experiment 1 were collected from novices and a different set of "near experts" (teaching assistants—TAs—in the introductory course from which novices were recruited). Our data allowed a comparison of the performance of novices and near experts (TAs) on ability to judge problem difficulty. Although we expected to find an advantage for the TAs in discriminating problems based on difficulty, we could not predict how easy or difficult the task of categorizing problems by difficulty would be for both groups, nor could we predict the magnitude of the difference between the groups.

The only study of which we are aware that has investigated differences between predictions of difficulty and student performance in solving physics problems was the Gire and Rebello [6] study referenced above. While they found differences between experts and novices in predicting which problems are more difficult, this study differs in two respects. In the first experiment, experts gave their rationale underlying their prediction of problem difficulty. This allows us to begin to investigate the strategies that experts employ to reason about problem difficulty with introductory physics problems. In the second experiment, the TAs and novices were not given time to solve the problems. Therefore, we could explore how accurate students' judgments were, which likely impacts both their ability to select problems to practice on while studying for an exam and their ability to pace themselves optimally during an exam.

II. EXPERIMENT 1

A. Method

Eight members of the University of Illinois Physics Education Research (PER) Group, including four advanced graduate students, and four faculty members were recruited to participate in the first experiment. All of the members had teaching and research experience related to introductory calculus-based mechanics courses. This group of eight experts was not paid for participation in the study. This group will be referred to as the PER expert group.

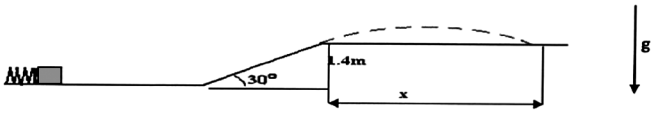
All problems used in the experiments came from previously administered exam problems. The problems were drawn from the first and second midterms administered in a calculus-based introductory mechanics course over previous semesters. Topics covered by these problems included force and motion, energy, and momentum. In addition, problems that combine two of the topics above were also presented to all of the groups. Participants in both experiments always judged problems in pairs by stating which problem in the pair was more difficult for students to solve. All problem pairs were selected out of "*problem sets*" from previously administered exams. Problem sets consisted of groups of 2–4 problems that share a common storyline; that is, a physics situation was described, most often accompanied by a diagram, and several problems were constructed based on the common situation. Thus, all problem pairs referred to the same situation, allowing for judgments based on a common situation; no pair was ever about entirely different topics or physical situations. All problems were multiple-choice, some having 5 choices and some having 3 choices. Pairings consisted of either two 5-choice problems, a 5-choice problem paired with a 3-choice problem, or two 3-choice problems.

Since there was a need to have problem pairs vary significantly in difficulty, we only included problem pairs that differed in difficulty by at least 15 percentage points—that is, the average performance of students solving the pair of problems in a real exam scenario differed by 15% or greater. Seventy nine problem pairs (30 force and motion, 20 energy, 20 momentum, and 9 mixed concept pairs) met the criteria for inclusion and were used in experiment 1. Two sample items used in both experiments are shown in Fig. 1.

The initial pool of 79 items was generated by the first author based only on the criterion that the two problems in

E20) The next two problems refer to the following situation:

A block of mass m=100g is launched horizontally on a frictionless track by compressing a spring of constant k=78N/m. After 1m (from the position of the spring at equilibrium) the track turns upward at a 30° angle as shown in the figure below. The height difference between the lower and upper level is 1.4m.



1. If you compress the spring by a distance of 20cm, what is the horizontal range of the block x shown in the figure?

a. 12cm
b. 33cm
c. 67cm
d. 93cm
e. 123cm

2. Let x be the answer to the previous question. If there is friction between the track and the block, the horizontal range of the block is now:

a. greater than x
b. equal to x
c. smaller than x

FIG. 1. Example problem pair.

the item differed by at least 15% based on exam performance. PER group experts (including the third author) were given the set of 79 items and told to select which problem in the pair would prove more difficult for students who took the exam in which the problems appeared. In addition, PER experts were asked to write the reasoning behind their judgment decision. The items were written on paper, one problem pair per page, with a space for the experts to denote which problem would be harder for students to solve, a place for them to write down a starting and ending time, and room at the bottom of the page for writing the reasoning behind their judgment. The PER experts were given the packet of problem pairs and could perform the task wherever and whenever they desired, and could take as long as they needed to complete all problem pairs. The untimed format was intended to allow the PER experts to reflect on each item as much or as little as they wished (e.g., they could consider how the solution to each problem in a pair would be constructed and ponder what issues might make one problem more difficult than the other); the untimed format also somewhat mimics part of what faculty do when they make up exam questions, namely, think about the level of difficulty of the problems.

B. Results

We begin by addressing the question of what criteria experts use to determine problem difficulty when considering the 79 problem pairs. Two faculty members out the four only provided their predictions of difficulty for the 79 pairs but neither their reasoning nor the timing information. After the PER experts completed the set of 79 problem pairs, one problem was removed from analysis because it was missing key information that would allow the participants to accurately compare the questions.

The rationale given by the six PER experts on the remaining set of 78 questions was analyzed using thematic analysis [20]. In thematic analysis, a subset of the data is analyzed to determine themes that emerge from the data. The same subset of data is then analyzed to determine more refined categories within each theme. The themes and categories are then discussed and refined to create a coding scheme. Finally, the remaining data are coded using the coding scheme. Three main themes emerged concerning the general methods that experts used to predict the reasons underlying the difference in problem difficulty experienced by students. Two raters (the second and third authors—the third author did not provide reasoning so he did not judge his own justifications) categorized the expert rationale with an initial 77% agreement; the raters then met to discuss problems on which there were disagreements in the ratings and to see if some common, agreed-upon rating could be reached. Following the discussion 99% agreement was reached.

1. Types of rationales given by experts

Experts used reasons that focused on the *question context*, the *content type*, and *student characteristics*. Further analysis indicated that within each major theme the rationale experts used to predict the more difficult problem within a problem pair tended to cluster into categories. Example rationales from each category are included in Table I.

Expert comments that focused on the *question context* either addressed the type of question, the distractors used in the response choices, or the wording of the question. Reasons that dealt with the type of question focused on the difference between the general types of questions in the problem pairs (e.g., conceptual versus calculation questions, or problems requiring calculations involving variables versus problems requiring calculations with numbers). Comments that

TABLE I. PER expert rationale category examples.

Theme	Category	Examples
Question context	Question type	"It seems that students would tend to do better on more conceptual problems."
		"No. 1 is pure calculation. No. 2 is conceptual."
	Distractor	"The distractor is more powerful in this one."
		"a, b, and e are good distractors."
	Wording	"No. 1 involves interpreting the expression 'Maximum', which some students are not very good at."
		"Question 2, picture suggests block 2 will go down, and block 1 up. So even [without] calculation you get correct answer."
Content type	More steps	"You need the answer to No. 1 to correctly determine 2."
		"No. 2 requires knowing the answer to No. 1."
	Math	"More calculation required."
		"Math is more difficult to set up properly for No. 2, plus students struggle with ratios. No. 1 is glorified plug and chug."
	Direction	"Students must set up [the equation] with appropriate sign changes in No. 1. In No. 2 the direction of [accel] is apparent."
		"The direction (or existence) of the acceleration is not readily apparent in 2."
	Content	"Stationary systems are easier for students than dynamic systems."
		"No. 1 is Newton's 3rd [Law], very simple."
Student characteristics	Familiarity	"Students are not trained to think about [acceleration] on the side of a vertical track. They are trained to calculate things at the top."
		"This seems to be a point that is not practiced nor emphasized as much."
	Misconceptions	"[Question] 1 prompts several misconceptions."
		"I think students would tend more to the force causes velocity p prim in No. 2."
	Intuition	"Changing masses is more intuitive than changing springs."
		"I think that conceptual question would be easier because it seems more intuitive, i.e., it fits better with everyday experience."
	Carelessness	"Students may easily forget the friction"
		"Students are careless."

addressed distractors focused on the presence or absence of good distractors. Reasons that dealt with the wording of the questions focused on problems or hints intrinsic in the questions (e.g., the answer is obvious from the diagram, or a term that may cause difficulty for some students).

Rationales given by experts that focused on the *content type* either addressed the number of steps involved in solving a problem, the level of math required, the need to consider direction or movement, or the differences in content topics. Comments that dealt with the number of steps focused on the number of steps involved in solving problems. Rationales that described differences in the level of math focused on the difference in mathematical skills need to solve the problems. Reasons that dealt with the need to consider direction or movement focused on differences in understanding of sign and/or direction needed to solve problems. Comments about the difference between specific content topics focused on differences in content assessed by the problems (e.g., dynamic problems are more difficult than static problems).

Expert reasoning that focused on *student characteristics* either addressed the familiarity of the content for students, misconceptions that students may possess, student use of intuition, or carelessness of the students. Reasons that dealt with the familiarity of the content focused on how familiar or unfamiliar particular problems are to the students. Rationales that addressed misconceptions focused on the misconceptions or p prims [21,22] that students hold that may contribute to incorrect answers. Comments discussing intuition focused on intuitions that the students hold that contribute to obviously correct answers. Reasons that dealt with student carelessness focused on students making simple or careless mistakes (e.g., students may forget steps).

Experts tended to use reasons that focused on the content (43.9%) more often than either student characteristics (27.4%) or question context (21.0%). The remaining comments (7.8%) either indicated that the experts were guessing, were unsure of their reasoning, or were unable to be categorized using the coding scheme described above. The frequency with which each category was used can be found in Table II.

2. Successful and unsuccessful types of + rationales used by experts

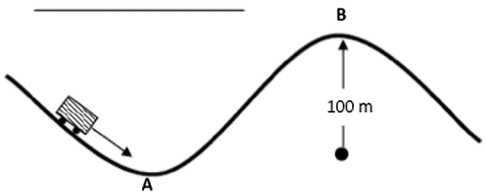
To address the question of which criteria were more successful in judging problems according to problem difficulty chi-square tests were conducted. Experts were more accurate than chance when they used rationales that dealt with distractors, number of steps, level of math, need to

TABLE II. Percentages and chi-square tests of accuracy above chance for PER experts; note $df = 1$.

	N (Percent)	Percent accuracy	Chi square	P
Question context				
Question type	68 (11.0)	60.3	2.882	NS
Distractor	35 (5.7)	85.7	17.857	<0.001
Wording or pictures	27 (4.4)	63.0	1.815	NS
Problem content				
More steps	68 (11.0)	85.3	33.882	<0.001
Math	53 (8.6)	81.1	20.547	<0.001
Direction	23 (3.7)	82.6	9.783	<0.01
Content	128 (20.7)	85.9	66.125	<0.001
Student characteristics				
Familiarity	82 (13.2)	73.2	17.610	<0.001
Misconceptions	13 (2.1)	69.2	1.923	NS
Intuition	46 (7.4)	76.1	12.522	<0.001
Carelessness	29 (4.7)	65.5	2.793	NS
Uncategorizable	41(6.6)	70.7	7.049	<0.01
Guessing or I don't know	7 (1.1)	57.1	0.143	NS

understand direction, content differences, student familiarity, and student intuition. Chi-square results for all rationales are found in Table II. In addition, PER experts tended to report a single reason (68.6%) more often than multiple reasons (31.4%) when giving their rationale for their prediction concerning problem difficulty. Experts who wrote more than one reason to support their prediction concerning problem difficulty were more accurate in identifying the more difficult problem in the pair (81.8% compared to 71.8%). Chi-square tests found that this was a significant difference $[\chi^2(1, N = 461) = 6.71,\ p < .05]$. The odds ratio for this effect is 1.91, indicating that the odds of incorrectly predicting the more difficult question in a problem pair if an individual gave one reason is nearly twice the odds of an individual who gave multiple reasons when making their prediction.

Experts were more accurate than chance when using the term intuition in their rationale, however, the term "intuition" may be used to mean more than one thing. Both Schon [23] and Clement [24] point to multiple meanings of the term intuition within the language. They note that individuals use the term intuition to mean both concrete ideas that are derived from embodied experiences and abstract concepts that logically follow if the individual has the underlying conceptual understanding. In order to determine if experts used intuition in both of these ways, both the question context and expert rationale were analyzed to determine whether the knowledge needed to answer the question was derived from embodied experiences or logically followed from an assumed conceptual understanding. For example, in response to the problem pair in Fig. 1, one expert said that problem number "2 is more conceptual, and the concepts are intuitive." In this case students are likely to have had experiences in which the presence of friction reduces the distance an object travels. This is contrasted by the problem pair presented in Fig. 2, where two experts both indicated that they viewed question number 1 as being "intuitive." In this case, students are unlikely to have a direct experience



N17) A car, which weighs 1000 N, travels over a bumpy road with a constant speed. Gravity acts. The road at point **B** is in the shape of a circle with a radius of 100 meters.

1. Which statement correctly relates the force of the car on the road at points **A** (a valley) and **B** (the top of a hill).

a. The magnitude of the force of the car on the road is larger at point **A** than it is at point **B**.
b. The magnitude of the force of the car on the road is larger at point **B** than it is at point **A**.
c. You need to know the road radius at point **A** to answer this question.

2. With what speed must the car travel such that the force that it exerts on the road at point **B** is 500N?

a. 9.8 m/s
b. 12.4 m/s
c. 16.3 m/s
d. 22.1 m/s
e. 31.3 m/s

FIG. 2. Example problem pair.

with the force that a car exerts on a road, but the answer does logically follow from a conceptual understanding of the problem. Two raters (the first and second authors) categorized expert rationale that used the term intuition according to these two usages with an initial 81% agreement, and following discussion 100% agreement was reached.

When the term intuition appeared in PER experts' rationales, experts used intuition to refer to knowledge associated with everyday experiences about 42% of the time. The remainder of the uses of the term intuition referred to inductive or heuristic reasoning where the answer logically follows if one has a conceptual understanding. Experts were more accurate than chance in their predictions when using intuition in the everyday sense. [$\chi^2(1, N = 20) = 12.80,\ p < 0.01$]. When experts used intuition to indicate that the answer logically follows from a conceptual understanding they were not more accurate than chance in predicting the more difficult question in the problem set [$\chi^2(1, N = 28) = 3.57,\ p > 0.05$].

3. What other factors impact expert performance?

Three additional scores were calculated to aid in the remaining analyses for the six PER experts that provided a rationale for their predictions. First, an accuracy score was obtained by calculating the percentage of correct predictions for each problem pair. Thus the accuracy score represents the percentage of experts who made a correct prediction for each question and had a value of between 0 and 100%. Second, the total number of distinct categories used by the PER experts in making their predictions was obtained for each problem pair. Finally, a problem difficulty difference score was calculated for each pair by subtracting the student performance of the more difficult question from the easier question on the exams. A linear regression was conducted with the accuracy score of each problem pair as the criterion variable, and the problem difficulty difference score and the number of different categories used by the experts to make their predictions as the predictor variables. Both the difference in difficulty and the number of different reasons used predicted accuracy in expert predictions. A summary of the results can be found in Table III.

For problems with a larger difference in student performance, experts were more accurate than chance in identifying the more difficult problem [$t(75) = 2.37,\ p < .05,\ \beta = 0.42$]. The magnitude of the difference in difficulty accounted for 7% of the variance in accuracy. In addition PER experts, as a group, were less accurate as the number of different categories used increased [$t(75) = 2.04,\ p < 0.05,\ \beta = -0.05$]. The more consistent experts were in their rationale (i.e., the fewer the number of total categories used in rationales for a given problem pair), the more accurate the predictions of problem difficulty were as a group. In other words, if the rationales provided by experts for a given problem pair were "all over the map", then the likelihood of a correct prediction of problem difficulty diminished. The number of different categories used by experts accounted for an additional 5% of the variance in accuracy. Overall, these two variables accounted for 12% of the variance in accuracy scores, indicating that while experts were more accurate on problem pairs with larger differences in the magnitude of differences in problem difficulty as experienced by previous students and when their rationales were consistent, other factors contributed to differences in accuracy observed within the PER expert groups.

TABLE III. Predictors of accuracy of judgment scores for PER experts; note $N = 77$.

Criterion variable	Predictor variable	R^2	b weight	F change	sig
Accuracy score	Difficulty difference	0.07	0.42	6.18	0.015
	Number of different reasons	0.05	−0.05	4.56	0.036

While the PER experts performed much better than chance in their overall predictions [77.0%, $\chi^2(1, N = 624) = 180.72,\ p < 0.001$], they obtained an accuracy score of 50% or lower on 16 problem pairs. The questions in these 16 problem pairs were examined to identify common contextual features that might lead to this poor ability to judge problem difficulty. The majority of these problem pairs ($N = 12$) compared questions involving 3 choices to questions involving 5 choices, where the 3-choice question was more difficult. All other things being equal, one would assume that 5-choice questions would make it harder to guess the correct answer, therefore students would tend to perform worse on these types of questions. However, these question pairs also differed on many other dimensions. For example, 3-choice questions typically exhaust all possible answers as distractors (e.g., an object moves to the right, left, or does not move), while for 5-choice questions only a set of possible answer choices are included. Of these problems, 11 required a conceptual understanding to obtain the correct answer for the 3-choice question, while the 5-choice questions were computational. See Fig. 3 for an example of this type of problem pair. In this problem pair the second question was more difficult. Students taking the exam performed almost 20% lower on the second question than the first. Five of the eight experts incorrectly predicted that the second question would be easier for students. The expert rationales indicated that they thought that questions concerning conceptual knowledge were either easier or more intuitive than problems involving computational solutions. For example, one expert indicated that, "students would tend to do better on more conceptual problems". Another expert commented that question "Two [is] intuitive, m goes up so $T <$ mg. [In number] One [students] have to write down two equations and solve."

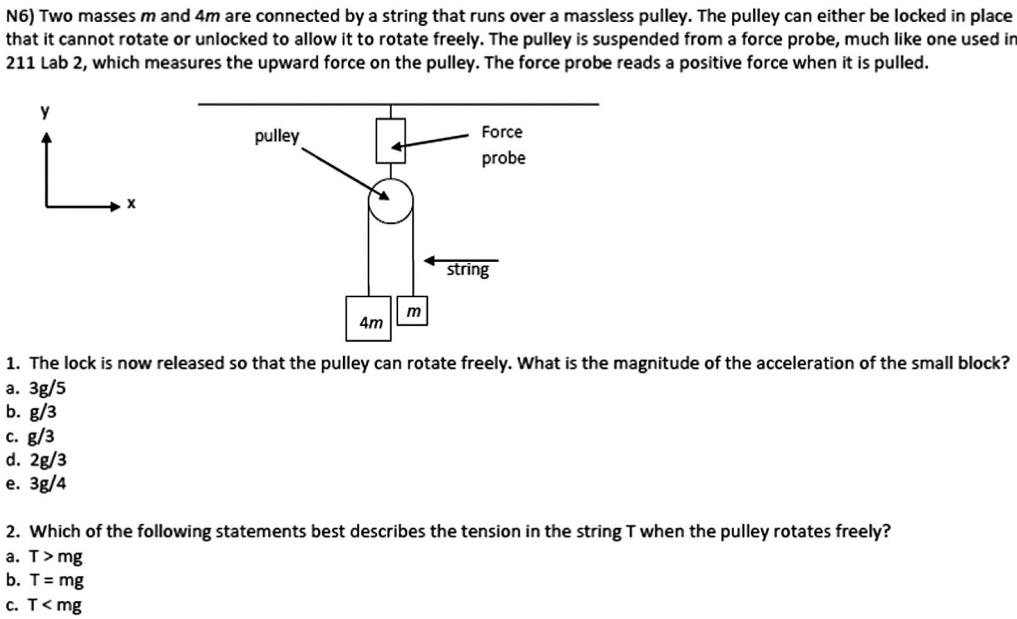


FIG. 3. Example problem pair.

The remaining four problem pairs either compared questions with the same number of choices and involving similar processes to solve the questions ($N = 2$) or compared questions involving 3 choices to questions involving 5 choices, where the 5-choice question was more difficult ($N = 2$). In general, experts were less accurate in their predictions of problem pairs that compared questions with differing numbers of choices, in which the questions with 3 choices were more difficult ($N = 26$, mean $= 60.6\%$, sd $= 23.9\%$) than problems pairs where the questions with 5 choices were more difficult [$N = 27$, mean $= 82.4\%$, sd $= 23.6\%$, $t(51) = 3.35$, $p < 0.01$], or problems pairs where the questions have the same number of choices [$N = 25$, mean $= 85.0\%$, sd $= 15.3\%$, $t(42.776) = 4.36$, $p < 0.001$].

The amount of time that PER experts spent categorizing the problem pairs was calculated by subtracting the end time from the start time as recorded by the experts. This method provided us with a relatively coarse-grained measure of the time experts spent in making their prediction. The majority of predictions were made in 2 min or less (77.2%), while most of the remaining predictions were made between 3 and 5 min (19.4%). Predictions made between 6 and 13 min represented 3.5% of the judgment times. To test our hypothesis that longer times to predict which problem in a pair is more difficult for students suggest the decision is more difficult and hence more prone to error, we compared the average time experts spent judging items where more than half of the experts were correct (62 items), to items where either half or less than half of the experts were correct (16 items). On average, experts spent less time judging the 62 problem pairs in which they had greater than 50% accuracy, than judging the questions in which they had 50% or lower accuracy [$t(76) = -2.52$, $p < 0.05$]. The difference in average time was about 30 sec (1.95 and 2.45 min, respectively) which represents a moderate effect ($d = 0.71$). See Table IV for descriptive statistics.

C. Discussion

The PER expert group data proved to be quite rich. Generally, we were able to categorize PER experts' reasoning into 11 subcategories that could be subsumed into the general categories of *question context*, *content type*, and *student characteristics*. In addition, a 12th category, "other," was used to categorize about 7% of the rationales that did not fit into any of the 11 subcategories. When using seven out of the eleven subcategories, PER experts were statistically better than chance in judging problem difficulty; their best performance came in the general category of *problem content* (with subcategories: number of steps needed to solve a problem, the level of the math involved, whether direction or sign played an important role in the problem, and difficulty differences between different content areas) within which all four subcategories were found to be statistically significant. They were less accurate in judging problem difficulty when using the other

TABLE IV. Descriptive statistics of time spent making predictions for PER experts.

Question accuracy	N	Mean	s.d.	Standard error
50% or lower	16	2.45	0.83	0.20
Greater than 50%	62	1.95	0.67	0.09

two major categories of *question context* and *student characteristics*.

Perhaps most surprising was the PER experts' "at or below-chance" performance on about 20% of the items. Among these 16 items, 12 were items with a 5-choice and a 3-choice multiple-choice problem where the 3-choice problem was more difficult than the 5-choice problem, and in 11 of the 12 the 3-choice item required conceptual understanding of the situation for students to obtain a correct answer. This finding could be due to PER experts having an implicit heuristic that 5-choice problems (which almost always require algebraic or numerical manipulations) are generally more difficult than 3-choice problems (which tend to be more qualitative and conceptual); that is, experts could be overestimating novices' ability to solve conceptual problems, perhaps because those are problems that they (the experts) find easier to solve given their expertise and thus cannot imagine how students would find them harder. In psychology this phenomenon goes by two other names: The illusion of transparency, and the curse of knowledge [25–27]. The former is the tendency of an individual to overestimate how well others understand their own mental state, the latter is the notion that better-informed individuals about a topic can have difficulty understanding how lesser-informed individuals think about the topic.

One surprising finding that is related to the idea of the curse of knowledge is the experts' use of the term intuition. PER experts used the terms intuition or intuitive in more than 7% of their explanations. In this study, experts tended to use the term intuition in two subtly different ways. Experts were more accurate than chance in their predictions only when using the term in the everyday sense. This finding suggests that experts could be overestimating the extent of novices' conceptual understanding, especially when considering problems that can be solved using simple logic following the conceptual understanding. In terms of predicting which problems are more difficult for students to solve, content expertise may deprive one of the subjective experience of solving the problem from the student's point of view, especially for conceptual problems whose solutions are intuitive to the expert. This has implications for physics education. If this means that experienced physics instructors assume (erroneously) that students are developing conceptual understandings through induction from solving calculation-based physics problems, then this assumption needs to be recalibrated. Further, instructional strategies that explicitly emphasize qualitative conceptual understanding in addition to instruction in problem solving should be explored (see Ref. [28] for a review).

Another surprising finding in terms of usefulness of particular subcategories to form rationales to make judgments about problem difficulty among the PER experts was the misconceptions subcategory, which was not statistically significant. Given the attention that misconceptions research has garnered in PER over the last forty years (for a review see Ref. [28]) one might think that PER researchers would be more attuned to making accurate judgments about problem difficulty when invoking the presence of misconceptions. However, it may be that misconceptions are only triggered by certain question contexts. Alternatively, it may be that misconceptions are helpful when dealing with conceptions held by individuals, but not when analyzing the performance of groups of students, especially those enrolled in an introductory calculus-based mechanics course, as these students may not hold the same misconceptions, as a group, as students in other introductory physics courses.

Other findings were less surprising. For example, the longer it took an expert to decide which problem in a pair was more difficult the more likely their judgment would be in error. When rationales combined more than one subcategory to make a judgment about problem difficulty they were more accurate by ten percentage points than when they used a single category. However, the more subcategories used by the experts, as a group, to judge a particular item, the poorer the performance on that item, suggesting that the contrast between some problem pairs made it easier for PER experts to select the more difficult question. Further, the larger the empirical difference in difficulty between the problems in a pair in terms of students' real exam performance, the easier it was for the PER expert group to judge problem difficulty. However, this accounted for only 7% of the variance. An additional 5% of the variance was accounted for by the number of different subcategories used by experts in constructing rationales to judge problem difficulty. This highlights the importance of using multiple reasons to make judgments of problem difficulty. Further research should be done to develop a rich theory of problem difficulty in introductory physics courses with students of differing experience and ability levels.

III. EXPERIMENT 2

Whereas experiment 1 was designed to identify the criteria or rationales used by experts to judge problem difficulty, experiment 2 explored the accuracy of novice and TA judgments of problem difficulty in a timed, examlike situation. Experiment 2 will also allow qualitative comparisons of the accuracy among the three groups, though we note that the experts in experiment 1 were not timed in their judgment task and could possibly solve the two problems in an item before judging their difficulty—something that was not possible for the TAs and novices in experiment 2.

Differences in judgments in problem difficulty may be due to the flaws in self-assessment among novices [12]. Prior research has found that peer assessment is often more accurate than the self-assessment at judging problem difficulty [29–31]. Thus, one would expect novices to be more accurate when predicting which problems are

normatively more difficult. Three groups were used in experiment 2 to investigate differences in judgments and whether judgments are more accurate when made with respect to their peers or themselves. Three main questions were investigated in this experiment. First, how accurate are both experts and novices in judging problem difficulty? Second, does the ability to judge problem difficulty improve with experience or content expertise? Third, is peer assessment more accurate than self-assessment when making judgments of physics problem difficulty?

A. Method

The novice participants in the study consisted of 38 undergraduate students enrolled in an introductory calculus-based mechanics course at the University of Illinois (25 males and 13 females). A class-wide email to students was used to recruit the novice participants, who were paid for their participation. The experts consisted of six TAs (henceforth the TA group) who taught discussion sections in the introductory calculus-based mechanics course. These TA experts were recruited with a similar email solicitation and were paid for their participation in the study.

From the 79 questions used in experiment 1, a subset of the problem pairs were selected for this experiment. Items where six out of the eight PER experts correctly judged which problem in the pair was more difficult according to the exam performance statistics were selected for this experiment. Therefore, all pairs of problems used in the novice-TA study not only differed empirically by at least 15% in difficulty, but were also correctly judged by at least 75% of the PER expert group. The reason for demanding both empirical and judgmental validity in selecting items is intended to provide consistency in item selection and to avoid possible controversies (i.e., a pairing that differed in difficulty by more than 15% in a previous exam administration but that PER experts disagreed on in judging which problem was more difficult). A total of 28 problem pairs (eight force and motion, eight energy, eight momentum, and four mixed concept pairs) were used in experiment 2. The entire set of items used in the novice-TA study is included in the supplemental materials accompanying this article [32].

Participants in the novice-TA study were placed in front of a computer screen, which presented pairs of problems (many with an accompanying diagram—see Fig. 1) delivered through OpenSesame experiment builder. Since differences in judgments in problem difficulty may be due to the flaws in self-assessment [12], and since peer assessment might be more accurate than the self-assessment at judging problem difficulty as defined in this study [29–31], the 38 novices were randomly divided into two groups and given slightly different instructions. One group was asked which problem would be more difficult for them to solve, while the other group was asked which problem would be more difficult for their classmates to solve. The six experts (TAs) were asked which problem is more difficult for their students. All participants were told that they did not have to solve the problem, just indicate which problem is more difficult. Problems and associated figures were displayed for 90 sec, so that participants would have time to read the questions and consider them, but not have sufficient time in which to solve them; the timed task was partly intended to probe students' ability to judge problem difficulty without first solving the problems since those judgments might impact exam-taking habits. After 90 sec, the questions disappeared. Participants then selected which problem was more difficult. After they entered the choice, the next problem pair was displayed. The entire experiment took less than 1 h. Student participants completed experiment 2 shortly after the second midterm to ensure that the relevant physics topics covered in the study had been sufficiently covered in the course.

B. Results

We begin by addressing two questions: (a) Do novices' ability to identify more difficult problems differ based upon the reference group they were prompted to use (themselves or their peers), and (b) do experts (TAs) and novices differ in their ability to identify more difficult problems? A Levine Test of the assumption of homogeneity of variance revealed that an omnibus (overall) ANOVA was appropriate for this data [$F(2,41) = 2.58$, $p = 0.09$]. The omnibus ANOVA was significant indicating that at least one significant difference was found between the groups [$F(2,41) = 7.38$, $p < 0.01$]. In order to determine which groups differed in their ability to predict problem difficulty, two planned *post hoc t* tests were conducted. There was no difference in the ability to determine problem difficulty whether novices used themselves or their peers as the criterion for judging problem difficulty [$t(41) = 0.59$, $p = 0.56$]. Experts (TAs) were more successful than the combined novices groups in their ability to predict the more difficult question from problem pairs [$t(28.15) = 7.87$, $p < 0.001$]. Experts, on average, were 16.4% more accurate in predicting the more difficult question which represents a large effect ($d = 1.68$). Means and standard errors can be found in Table V.

We then addressed the question; does the ability to judge problem difficulty improve with experience or content expertise? To further investigate the differences between experts (TAs) and novices the accuracy of the novices and experts was obtained for each question. The results are shown in Fig. 4. As expected, the experts outperformed the novices on 19 of the 28 questions and did not perform statistically worse on any question. However, the novices did obtain an accuracy score more than 14% higher on two problems, namely, P3 and P15. In addition it appears as if there are two groups of problems in this experiment, 10 in which the novices performed no better than chance, and 18

TABLE V. Accuracy scores for the 28 questions used in experiment 2. Note that the 28 questions selected for the expert novice study were selected based upon high accuracy score among the PER experts.

Group	*N*	Mean	s.d.	Standard error
Novices peer	19	72.0%	9.7%	2.2%
Novices self	19	70.1%	11.2%	2.6%
Experts (TAs)	6	87.5%	3.0%	1.2%
PER experts	8	96.0%	3.0%	1.1%

where they did perform better than chance. No patterns could be found to explain this grouping of questions. Future research should be directed at eliciting the rationale employed by novices in making difficulty judgments on these problems.

C. Discussion

The differential performance between novices and the TA experts indicates that determining problem difficulty is, not surprisingly, a trait of expertise which develops with experience. However, even novices after taking an introductory course were significantly above chance (71%, with chance being 50%) in selecting the more difficult problem in the item pairs. Nevertheless, novice performance in judging problem difficulty suggests this is a difficult task, and may contribute to time-management issues on exams, especially for lower-performing students. That is, the inability to judge problem difficulty during an exam may cause some students to spend inordinate time solving difficult problems in the order they encounter them rather than solving the easy ones first and then returning to the more difficult ones, or to apply inappropriate problem solving strategies. On the other hand, by the time they become TAs, graduate students are quite good at judging problem difficulty (87.5% performance, which is approaching the PER experts' 96% performance on these problem pairs).

On average, novices tended to be equally accurate whether they used themselves or their peers as a reference group. This is not surprising in this context, since students tend to use their subjective experience when making predictions of problem difficulty when engaging in problem solving [15]. In addition, when deprived of the ability to use the subjective experience of solving the problems, the novices had to rely on their implicit understandings of what makes a problem difficult [16,17]. Familiarity may be a strategy that tends to be overused by novices in the absence of the experience of solving the problems [19]. This heuristic may be a good predictor of problem difficulty when deep processing strategies are employed. For example, if a student recalls that problems involving a certain content area tend to be difficult for them to solve, while a different content area tends to be easier, then this heuristic may be productive. However, if the focus is on the terminology in the question or diagrams, then the use of familiarity may be misleading with respect to problem difficulty.

IV. CONCLUSION

Students' ability to effectively study for an exam, or to manage their time during an exam, is related to their metacognitive regulation. Reference [33] presents a three phase view of self-regulatory processes that individuals draw upon during learning depending on whether the processes occur before, during, or after learning or studying. Several studies have found that experts tend to exhibit greater self-regulatory processes and that these processes can be developed through training in metacognitive strategies (see Ref. [33] for a review). Although the design of this study was cross sectional, the results provide evidence

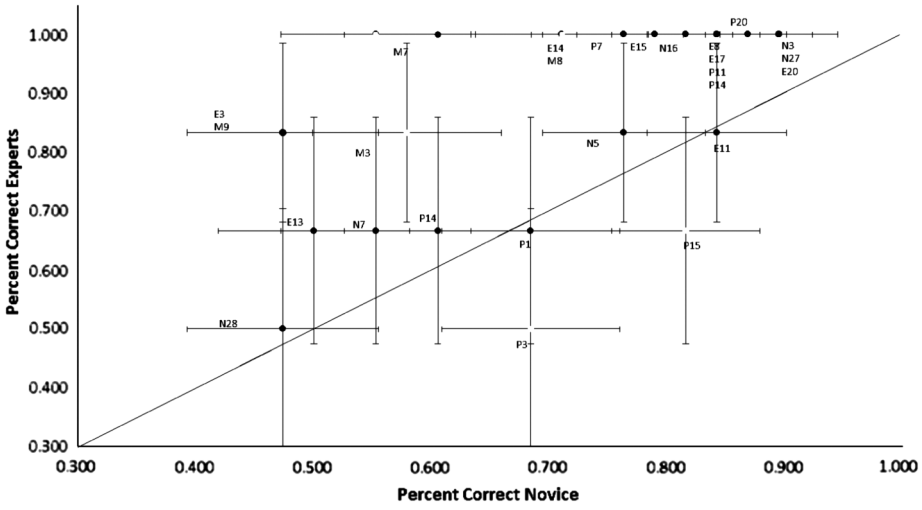


FIG. 4. Percent correct for novices and experts (TAs).

that these processes may be developed through the attainment of content expertise. Alternatively, the development of these skills may be necessary in order to attain content expertise. The two experiments in this study attempted to examine metacognitive regulation among experts and novices by exploring three main questions.

The first question focused on identifying the criteria used by PER experts when judging problem difficulty. We identified 11 different criteria used by PER experts under the three general categories of *question context*, *content type*, and *student characteristics*. PER experts were statistically better than chance when they invoked only 7 of the 11 criteria: attractive distractors under *question context*; number of problem steps, level of math involved, whether the direction or sign of a quantity was pivotal in generating an answer and, and problem contexts (e.g., dynamic vs static), all under *content type*; and the familiarity of the type of problem for the student, and whether students' intuitions aligned with the correct answer, both under the *student characteristics* category. PER experts were no better than chance when invoking the type of question (e.g., conceptual vs computational) and the wording of the question as judgment criteria under the *question context* category, or when invoking misconceptions that students hold that may impact their reasoning on a problem and when they thought students would make careless errors under the *student characteristics* category. Future research should investigate the types of rationales used by novices to make judgments about problem difficulty in order both to explore similarities and difference to the rationales used by experts and to use the information gleaned to help devise pedagogical strategies to improve novices' metacognitive strategies for judging problem difficulty.

The second question explored the accuracy of PER experts, TAs, and beginning physics students in judging problem difficulty, and the third question explored how performance depended on level of expertise. Besides the obvious finding, namely, that PER experts were better at judging problem difficulty than TAs, which in turn were better than novices, findings from this study provided an absolute gauge on just how accurate these groups are, and how progression toward expertise impacted problem difficulty judgment ability. Among the 28 problem pairs that were used in the second experiment with TAs and novices, where the PER experts' judgments were highly accurate (96% correct), the TAs came in at 87% and novices at 70%, indicating that novices were relatively poor at judging problem difficulty, which in turn could impact their budgeting of time in a real exam scenario. Further, since prior research indicates that students use judgments of problem difficulty to allocate their study time [9–11], their relatively poor ability to judge problem difficulty could adversely impact what or how they study to prepare for an exam. Surprisingly, our PER experts were below chance (50% performance) in judging problem difficulty in 20% of the 78 problem pairs they judged. Also somewhat surprising, PER experts appear to have an implicit bias that 3-choice questions (which are predominately conceptual) are easier for students than 5-choice questions (which are predominately computational), a bias that turned out to be inaccurate.

Two methodological issues are worthy of mention. The first concerns the difference in timing used in experiments 1 and 2. In experiment 1 the experts had as much time as they needed to judge which of the two problems in each item was more difficult. Hence they could, if they wished, solve each problem and then render a judgment along with a rationale for the judgment. In experiment 2, the 90 sec allowed to judge each item pair was not sufficient to solve each problem or to deliberate long before making a judgment. That is, in experiment 1 the experts could compare and contrast the solution strategies for the two problems in much more detail than the TAs and novices could in experiment 2. The short time allowed for judgments in experiment 2 may have forced some TAs and students into using superficial attributes (e.g., familiarity) to make judgments. This difference in methodology places limitations in our ability to gauge absolute problem difficulty judgment ability among experts, journeymen (TAs), and novices. Future work that uses the same methodology to explore judgments of problem difficulty among individuals possessing different levels of expertise could provide additional insights.

The second methodological issue concerns using a two-problem judgment task as was done in this study, versus a one-problem task where subjects rate the difficulty on a scale (e.g., 1 to 10), as was done in the Gire and Rebello [6] study. In a two-problem judgment task, the decision is binary (the subjects picks one of the two problems) while in a one-problem task the decision is more subjective, with the subject assigning a numerical value to the perceived difficulty of the problem. Further, the ability to compare or contrast directly the features of the problems in a two-problem task is not present in a one-problem task. This ability to compare or contrast features may have allowed the experts in this study to generate criteria for judgments that may not have been afforded by a one-problem task. The two-problem task was also more akin to what students are faced with when in an exam scenario, since most multiple choice exams in physics have a story line around which two or more problems are presented. Two-problem tasks (and even three-problem tasks where a "model problem" is compared to two "comparison problems") are common in problem categorization tasks where subjects are asked to categorize problems according to similarity of solution [34–36] because they allow the experimenter to manipulate directly the surface features and deep structure of problems. Thus, problem comparison tasks afford the participant the opportunity to make judgments by comparing or contrasting problem attributes or features that would not be as evident in a one-problem task. In Ref. [37] a process model is discussed of case comparison that

describes the processes individuals undertake in completing this type of task. They note that both the type of cases and the experience level of the participant influences the information considered by the participants. The findings of this study support this assertion. In experiment 1, experts were less accurate than chance in judging problem difficulty when the more difficult question was conceptual compared to a calculation question. This suggests that expectations of students' conceptual sophistication may need to be tempered or that pedagogical shifts need to occur in order to facilitate conceptual development. In experiment 2, no patterns in the differences between the problems that affect judgment accuracy were found; however, experts were more accurate than novices in making normative judgment of problem difficulty. Thus, we noted differences in accuracy based on experience level.

As in studies of problem categorization by solution similarity, this study of problem categorization according to problem difficulty tracks with expertise, with better performance exhibited by more knowledgeable individuals. This study did not address how physics ability among novices affects ability to judge problem difficulty. Future research could address how ability to make accurate judgments of problem difficulty differs among both high-performing and low-performing students. One would expect that lower-performing students would be less accurate in their judgments of problem difficulty, but this is an empirical question which needs to be explored. Finally, it might prove fruitful to investigate whether there is a correlation between the ability to judge problems by difficulty and test performance. If so, one could explore interventions to help low-performing students budget their time in exams by training them to distinguish between easy and hard problems; they could initially skip harder problems in an exam, solving the easy problems first, and then spending the remaining time with the difficult problems. In addition, it might prove interesting to explore if helping low-performing students improve their ability to judge problems by difficulty in combination with coaching them in test-taking strategies results in improved exam performance.

[1] K. M. Zabrucky, Knowing what we know and do not know: Educational and real world implications, Procedia Soc. Behav. Sci. **2**, 1266 (2010).

[2] J. L. Feitcher, A. S. Benjamin, and N. Unsworth, The Metacognitive Foundations of Effective Remembering, *Oxford Handbook of Metamemory*, edited by J. Dunlosky and S. K. Tauber [(Oxford University Press) (to be published)].

[3] R. Glaser and M. T. H. Chi, *The Nature of Expertise*, edited by M. T. H. Chi, R. Glaser, and M. Farr (Erlbaum, Hillsdale, NJ, 1988), pp. xv–xxvii.

[4] W. Schneider, Memory development in childhood, *Blackwell Handbook of Childhood Cognitive Development*, edited by U. Goswami (Blackwell, Oxford, 2002), pp. 236–256.

[5] T. D. Griffin, B. D. Jee, and J. Wiley, The effects of domain knowledge on metacomprehension accuracy, Mem. Cogn. **37**, 1001 (2009).

[6] E. Gire and N. S. Rebello, Investigating the perceived difficulty of introductory physics problems, AIP Conf. Proc. **1289**, 149 (2010).

[7] N. S. Rebello, How accurately can students estimate their performance on an exam and how does this relate to their actual performance on the exam?, AIP Conf. Proc. **1413**, 315 (2012).

[8] J. Metcalfe and B. Finn, Evidence that judgments of learning are causally related to study choice, Psychon. Bull. Rev. **15**, 174 (2008).

[9] G. Mazzoni and C. Cornoldi, Strategies in study time allocation: Why is study time sometimes not effective?, J. Exp. Psychol. Gen. **122**, 47 (1993).

[10] A. Koriat, H. Ma'ayan, and R. Nussinson, The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior, J. Exp. Psychol. Gen. **135**, 36 (2006).

[11] J. G. Tullis and A. S. Benjamin, On the effectiveness of self-paced learning, J. Mem. Lang. **64**, 109 (2011).

[12] D. Dunning, C. Heath, and J. M. Suls, Flawed self-assessment: Implications for health, education, and the workplace, Psychol. Sci. Publ. Interest **5**, 69 (2004).

[13] L. Bol and D. J. Hacker, Comparison of the effects of practice tests and traditional review on performance and calibration, J. Exp. Educ. **69**, 133 (2001).

[14] D. Dunning, K. Johnson, J. Ehrlinger, and J. Kruger, Why people fail to recognize their own incompetence, Curr. Dir. Psychol. Sci. **12**, 83 (2003).

[15] C. M. Kelley and L. L. Jacoby, Adult egocentrism: Subjective experience versus analytic bases for judgment, J. Mem. Lang. **35**, 157 (1996).

[16] A. S. Benjamin, Response speeding mediates the contribution of cue familiarity and target retrievability to metamnemonic judgments, Psychon. Bull. Rev. **12**, 874 (2005).

[17] L. M. Reder, Strategy selection in question answering, Cogn. Psychol. **19**, 90 (1987).

[18] L. L. Shanks and M. J. Serra, Domain familiarity as a cue for judgments of learning, Psychon. Bull. Rev. **21**, 445 (2014).

[19] L. M. Reder and F. E. Ritter, What determines initial feeling of knowing? Familiarity with question terms, not with the answer, J. Exp. Psychol. Learn. Mem. Cogn. **18**, 435 (1992).

[20] R. E. Boyatzis, Transforming Qualitative Information: Thematic Analysis and Code Development (SAGE, Thousand Oaks London, and New Delhi, 1998).
[21] A. A. diSessa, Knowledge in pieces, *Constructivism in the Computer Age*, edited by G. Forman and P. Pufall (Lawrence Erlbaum, Hillsdale, NJ, 1988), pp. 49–70.
[22] A. A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).
[23] D. Schon, Intuitive thinking? A metaphor underlying some ideas of educational reform, *D.S.R.E. Working paper WP-8, Massachusetts Institute of Technology*, Cambridge, MA, 1981.
[24] J. J. Clement, in *Implicit and Explicit Knowledge: An Educational Approach*, edited by D. Tirosh (Ablex Publishing Corporation, Norwood, NJ, 1994), pp. 204–244.
[25] C. Wieman, The "Curse of Knowledge," or why intuition about teaching often fails, The Back Page, APS News, November 2007, Volume **16**, Number 10.
[26] The illusion of transparency. Wikipedia entry, retrieved 11–2014. http://en.wikipedia.org/wiki/Illusion_of_transparency.
[27] The curse of knowledge, Wikipedia entry, retrieved 11–2014. http://en.wikipedia.org/wiki/Curse_of_knowledge.
[28] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).
[29] D. A. Risucci, A. J. Torolani, and R. J. Ward, Ratings of surgical residents by self, supervisors and peers, *Surgery Gyne. Obstet.* ***169****, 519 (1989).*
[30] B. M. Bass and F. J. Yammarino, Congruence of self and others' leadership ratings of Naval officers for understanding successful performance, Applied psychology **40**, 437 (1991).
[31] T. K. MacDonald and M. Ross, Assessing the accuracy of predictions about dating relationships: How and why do lovers' predictions differ from those made by observers?, Pers. Soc. Psychol. Bull. **25**, 1417 (1999).
[32] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevSTPER.11.020128 for the problem pairs used in the study.
[33] B. J. Zimmerman, Development and adaptation of expertise: The role of self-regulatory processes and beliefs, *Expertise and Expert Performance*, edited by K. A. Ericsson, N. Charness, R. R. Hoffman, and P. J. Feltovich (Cambridge University Press, Cambridge, England, 2006), pp. 705–722.
[34] P. T. Hardiman, R. Dufresne, and J. P. Mestre, The relation between problem categorization and problem solving among experts and novices, Mem. Cogn. **17**, 627 (1989).
[35] J. L. Docktor, J. P. Mestre, and B. H. Ross, Impact of a short intervention on novices' categorization criteria, Phys. Rev. ST Phys. Educ. Res. **8**, 020102 (2012).
[36] J. L. Docktor, N. E. Strand, J. P. Mestre, and B. H. Ross, Conceptual problem solving in high school physics, Phys. Rev. ST Phys. Educ. Res. **11**, 020106 (2015).
[37] L. Alfieri, T. J. Nokes-Malach, and C. D. Schunn, Learning through case comparisons: A meta-analytic review, Educ. Psychol. **48**, 87 (2013).