Do students know what they know? Exploring the accuracy of students' self-assessments

Beth A. Lindsey¹ and Megan L. Nagel²

¹Physics, Penn State Greater Allegheny, McKeesport, Pennsylvania 15132, USA

²Chemistry, Penn State Greater Allegheny, McKeesport, Pennsylvania 15132, USA

(Received 16 December 2014; published 17 July 2015)

We have conducted an investigation into how well students in introductory science classes (both physics and chemistry) are able to predict which questions they will or will not be able to answer correctly on an upcoming assessment. An examination of the data at the level of students' overall scores reveals results consistent with the Dunning-Kruger effect, in which low-performing students tend to overestimate their abilities, while high-performing students estimate their abilities more accurately. Similar results have been widely reported in the science education literature. Breaking results out by students' responses to individual questions, however, reveals that students of all ability levels have difficulty distinguishing questions which they are able to answer correctly from those that they are not able to answer correctly. These results have implications for the future study and reporting of students' metacognitive abilities.

DOI: 10.1103/PhysRevSTPER.11.020103

PACS numbers: 01.40.Fk

I. INTRODUCTION

Historically, a great deal of work in the field of physics education research (PER) has focused on the improvement of student conceptual understanding [1,2]. Another focus has been on students' epistemologies and their beliefs about the nature of physics and of learning in physics [3,4]. More recent work examines students' self-efficacy [5], and their reasoning skills more broadly [6]. Many curricula and educational interventions [7–11] have been devised and shown repeatedly to be effective at improving students' conceptual understanding [12–16] or their beliefs about learning physics [17,18]. In order to truly demonstrate understanding, however, students must be able to not only answer questions correctly but also to recognize their correctness. In other words, students should be able to reflect on, and be metacognitive about, their learning. Metacognition is loosely defined as “thinking about one's own thinking,” but some authors note that “studies (of metacognition) do not support a coherent understanding of this concept” [19]. Here, we examine one aspect of metacognition: students' ability to evaluate their own understanding. In particular, we have investigated their metacognitive calibration, which describes how accurately someone assesses their own knowledge: “A well calibrated individual correctly assesses his state of knowledge, knowing when he knows, and knowing when he does not know. In contrast, the self-assessments of knowledge of the poorly calibrated person are uncorrelated with actual states of knowledge” [20]. To demonstrate understanding, students

are expected to recognize questions that they are capable of answering correctly, and to distinguish them from those that they are not capable of answering correctly. The ability to recognize what one does and does not know may be a prerequisite skill for allowing students to reflect on and regulate their own learning [21]. This ability of self-reflection may in turn lead to an additional increase in understanding [22].

Previous studies in chemistry courses [23] have shown that students who do not perform well on course assessments typically overestimate their abilities. Those students who perform near the top of the class tend to assess their abilities more accurately and even appear a bit underconfident. This is an instance of a well-studied phenomenon in the educational psychology literature known as the Dunning-Kruger effect [24]. Dunning and Kruger suggest that the students who do not perform well lack not only the content knowledge, but also the metacognitive skills that would allow them to recognize their lack of content knowledge. The Dunning-Kruger effect is a result of this lack of metacognitive skill, which leads to the much greater disparity between self-assessed skill and actual skill for students who perform poorly than for students who perform well.

This paper presents an investigation into student ability to distinguish items that they can answer from those that they cannot answer correctly. We explore factors that may influence this ability, including attributes of the students themselves (as measured by their exam performance) and attributes of the questions (for instance, the topic being assessed). While the main focus of the study is an introductory calculus-based mechanics class, data from a chemistry course (General Chemistry I) are presented to serve as a comparison and counterpoint.

The primary tool used in this investigation is a Knowledge Survey [25]. A Knowledge Survey (KS)

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

consists of a large set of questions (≥ 100), administered before students complete a course assessment (such as the final exam). Students do not attempt to answer these questions. Instead, they indicate for each question which of the following levels best matches their confidence in their ability to answer correctly:

- (A) I could answer this question correctly.
- (B) I could answer 50% of this question, or know where to quickly get the information I need.
- (C) I am not confident that I could answer this correctly.

A subset of the questions on the KS are closely matched (albeit not identical) to questions on the course assessment. Students' responses to these questions on the KS can then be compared against their actual performance on the course assessment. The KS can thus give an indication of whether students can accurately identify questions that they are able to answer correctly. For students, reviewing the KS can provide them a quick and easy map of where they feel they must invest some time (for choice B) or substantial time (for choice C) in studying. (In practice, we found that very few students selected choice C on the KS, as will be described below. In our analyses, we have thus grouped choices B and C together.)

Although the work presented here provides evidence for the Dunning-Kruger effect as it is traditionally reported, a more detailed analysis of the data gives other indications. When examined on a question-by-question basis, students in the top third of the class are actually not any more successful at distinguishing between items they are able to answer correctly and those that they are not able to answer correctly than are students in the bottom third, suggesting that the Dunning-Kruger effect may not tell the whole story.

II. BACKGROUND

Many studies have looked at how students assess their own knowledge after completing an assessment or exam. Dunning and Kruger reported that students tend to overestimate their own scores on assessments, and that such overestimation is typically worst among subjects who perform the worst on these assessments. They link this to poor metacognitive skills among those subjects [24,26]. Such an effect has also been reported extensively in the sciences. For example, Rebello [27] reported that when asked to estimate their scores at the conclusion of an exam, students in a physics course tended to overestimate their own score and to underestimate the scores of their classmates. He also found that the overestimation was worse among the students who scored in the bottom third of the class. He suggested that encouraging low-performing students to reflect on why they might have overestimated their performance might enable those students to improve their performance on future assessments. The Dunning-Kruger effect has also been reported in General Chemistry [28] and Organic Chemistry [29] courses, when students were asked to assess their own performance *after*

completing an assessment. Furthermore, when students are asked to make predictions about their performance *before* completing an assessment, a similar phenomenon arises. Bell and Volckmann [23], for instance, reported an observation of the Dunning-Kruger effect when students were asked to complete a Knowledge Survey prior to taking an associated assessment.

The status of the Dunning-Kruger effect as a real effect, however, has been called into question. Krueger and Mueller [30] suggest that the observed result may be due to a combination of statistical regression and "better than average" effects (most people believe that they are better than the average person). Other investigators have suggested that its prevalence may be related to the difficulty (or students' perception thereof) of the test items at hand [31]. Despite these concerns, reports of the Dunning-Kruger effect continue to be prevalent in the science education literature. These conflicting conclusions indicate that further analysis is warranted.

Each of the studies described above effectively looks only at students' aggregate score—i.e., how well they are able to predict their overall score on an assessment, or how well a series of judgments on a KS reflects students' overall score on the final exam [23]. Other research, primarily in the fields of psychology and educational psychology, examines how well students' confidence on individual items matches their scores on those items. Rosenthal *et al.* found that when students in a psychology class were asked to rate their confidence in each item they responded to on a final exam, the average of their confidence judgments correlated with their overall exam score. These students were also asked to predict their overall exam score before taking the exam, and "postdict" their exam score after completing the exam. The pre- and postdicted overall exam scores did not correlate with the students' actual exam scores. This suggests that the students were more accurate in making a series of microlevel judgments (evaluating the correctness of individual answers to test items) than they were in making accurate macro-level judgments of their overall score [32]. Sinkavich, in a study of students in educational psychology courses, found a similar result, and also observed that students as a whole were more likely to get correct an item to which they had given a higher confidence rating than one to which they had given a lower confidence rating [33]. Sinkavich also observed that this effect was much more pronounced for "good" students (those scoring in the top 33% of the class) than for "poor" students (those scoring in the bottom 33% of the class). For poor students, he reported little variation in the percent of correct responses given across different confidence ratings. These results collectively suggest that analysis of student responses at a microscale might be used to examine other perspectives on the Dunning-Kruger effect.

Within physics education, studies have examined whether student confidence can be used as an indicator

of the presence of strong alternative conceptions. Potgieter *et al.* [34] compared student confidence on items that many students get wrong due to the presence of strong alternative conceptions relative to those that students get wrong for other reasons (e.g., lack of problem-solving skills). They found that student confidence tends to be more inflated on items for which strong alternative conceptions exist; the inflation of confidence on those items for which students simply lacked a needed skill was much lower.

Many studies have looked at various measures of student calibration, including student ability to predict their overall scores on assessments, as well as student judgments of their confidence on individual test items. No prior study, however, has examined at both a macro- and a microlevel how well students are able to assess their knowledge (i.e., to predict which specific items they will be able to answer and which they will not be able to answer) *before* attempting to answer the items. The ability to make such determinations of knowledge is necessary if students are to be able to direct their own studying—if they are to be able to engage in self-regulated learning.

III. METHODS

A. Student population

This study was conducted at Penn State Greater Allegheny, a small campus of The Pennsylvania State University located in Western Pennsylvania. This paper will focus on research conducted in two courses: Physics 211 (introductory calculus-based mechanics) and Chemistry 110 (General Chemistry I). Both of these courses are offered every semester, and primarily serve STEM majors in their first two years of study. In Physics 211, approximately 25% of the students are female and 75% are male. Data associated with the General Chemistry course were only collected from students concurrently enrolled in the General Chemistry laboratory (Chemistry 111), as the researchers had greater access to this course and the lab setting allowed ample time for administration of the KS. For many students in Chemistry 110, the lab course is taken concurrently to fulfill either major or general education requirements. The subset of students from which data were collected represents roughly 70% of the total students enrolled in the general chemistry lecture. The General Chemistry course and lab are typically 30% female. Students in both the physics and the chemistry courses typically attend under the “2 + 2” program: They complete the first two years of their major at the campus, and then move to another larger campus to complete the final two years. Throughout this study, Physics 211 was taught by an author of this paper (B. A. L.). The instructor made full use of the results of Physics Education Research (PER): the course was taught in an interactive manner including interactive materials developed in-house as well as Peer Instruction questions [16], small-group problem

solving, and tutorial worksheets such as *Tutorials in Introductory Physics* [7]. Data from Chemistry 110 come from sections taught by a faculty member who was not otherwise involved in the research. It was taught in a traditional lecture format with little emphasis on interactive engagement.

B. Development of the Knowledge Survey

The primary tools used in this study were the Knowledge Surveys (KS) themselves. We developed a KS for each of the courses involved. Each KS consisted of a large number (≥ 100) of questions, which were meant to mirror the breadth and depth of the final exam of the course in which the KS was deployed.

For the physics KS, questions were drawn from exams given in prior semesters, from the PER literature, and from well-known Concept Inventories such as the Force Concept Inventory (FCI) [35] and the Force and Motion Conceptual Evaluation (FMCE) [36]. The “questions” thus consisted of short problems to be solved, conceptual questions requiring complex reasoning, and conceptual questions that appealed to students’ common sense. Students were already familiar with each of these question types—all three had been common throughout the semester on in-class activities, homework assignments, and course exams. Sixteen questions were carefully matched to specific questions from post-course assessments: Eight were drawn from the course final exam, and another eight were drawn from the FCI. The same final exam was used in all sections of each course from which data are drawn. In each case, the “matched” questions on the KS contained a version of the question stem from the assessment, but did not include the multiple-choice answer options that would be present when students actually completed the assessment. (Although the final exam included both free-response and multiple-choice elements, only the multiple-choice questions were matched to KS questions, for ease of unambiguously identifying a response as “correct” or “incorrect.”) For an example of a pair of KS questions and the corresponding matched exam questions, see Fig. 1.

For the chemistry KS, questions were primarily drawn from exams given in prior semesters, and covered a wide range of topics typical of a first semester of chemistry. The researchers had access to the actual final exam given in the course, which was composed entirely of multiple-choice questions. Twenty-five of the KS questions were nearly equivalent to questions appearing on the final exam. These 25 questions were used to create the matched subset.

C. Administration of the KS

The KS was administered in the first half of the final week of the semester. Students were given as much time as they needed, but were told that it should not take them more than half an hour to complete the KS. Completing the KS was a course requirement for all students and a small

Knowledge Survey Questions	Final Exam Questions
<p>Indicate A [Could answer now] if you feel confident that, given enough time and access to your cheat sheet, you can answer the question sufficiently well for graded test purposes.</p> <p>Indicate B [Could quickly find the answer] if you could answer now at least 50% of the question or if you know precisely where you could quickly get the information (within 20 min.) and then could complete the answer for graded test purposes.</p> <p>Indicate C [Could not answer] if you are not confident that you could adequately answer the question for graded test purposes at this time.</p>	
<p>1. A string is used to whirl a stone of mass 0.5 kg around in a horizontal circle. The stone is moving with a speed of 10 m/s, and the radius of the circle is 0.5 m. What is the tension in the string?</p>	<p>1. A string is attached to a 0.040-kg air hockey puck resting on an air hockey table, which is level and has negligible friction. The string is used to whirl the air hockey puck in a horizontal circle with a radius of 0.25 m. The puck moves with a constant speed of 5 m/s. What is the tension in the string?</p> <p>A. 0 N B. 0.39 N C. 3.61 N D. 4 N E. 4.39 N</p>
<p>2. A string is used to whirl a stone of mass 0.5 kg around in a vertical circle. The radius of the circle is 0.5 m. The stone is moving with a speed of 10 m/s at the BOTTOM of its path. What is the tension in the string when the stone is at its LOWEST point?</p>	<p>2. The 0.040-kg air hockey puck is now whirled in a vertical circle with a radius of 0.25 m. What is the tension in the string when the puck is at the bottom of the circle if the speed of the puck at that point is 5 m/s?</p> <p>A. 0.39 N B. 0.59 N C. 3.61 N D. 4 N E. 4.39 N</p>

FIG. 1. A pair of questions from the physics KS, and the corresponding questions from the physics final exam.

amount of course credit was awarded for doing so. The students were given the opportunity to sign an informed consent form; only those students who gave consent are included in our sample.

In Physics 211, the KS was administered online. Students had access to the KS online after they had completed it. Thus they could, in principle, refer back to questions on the survey as they were preparing for the final exam. In Chemistry 110, the KS was administered on paper during a lab session. The researcher collected the KS and students did not have subsequent access to it.

D. Scoring the KS

In previous studies, scores on the KS have been calculated by mapping every “A” response to a large point value (e.g., 100 points), every “C” response to a small point value (e.g., 0 points), and every “B” response to an intermediate point value (e.g., 50 points) [23,37,38]. The total KS score could then be calculated by averaging the scores of all the items. In practice, however, we found that very few students selected choice “C” (which would indicate that they recognize that they are unable to answer a particular question). Our analysis focuses on only those items that were carefully matched between the KS and the end-of-course assessments. From this question pool, “C” accounted for only about 10% of choices overall, and between 25% (chemistry) and 58% (physics) of students never selected choice C for any of the matched questions.

It was decided that students might not be adequately distinguishing between choices “B” and “C”, and thus these two responses were combined in this analysis. Every “A” response was coded as 100 points, and every “B” or “C” response was coded as 0 points. This would clearly distinguish between items that students are absolutely certain they are able to answer (for which they have chosen A) and those items for which they experience any doubt in their ability to answer. As in other studies, the total score on the KS was calculated by averaging the scores for the individual items. Since our focus was on the subset of KS items that were matched to end-of-course assessment items, a score is calculated for just that subset. This score is referred to as the “KS-matched score.” The actual assessment items were graded as correct or incorrect, and a student’s score on the matched assessment items was calculated as the percent of the matched items to which they had responded correctly. These two scores (the KS-matched score and the actual score on the matched assessment items) are both reported as percentages that can easily be compared to one another.

IV. RESULTS

A. Results from the physics course

1. Matched KS and assessment items: Overall score

There was an overall correlation between students’ KS-matched scores and the scores that they actually

FIG. 2. Scatterplot of students' KS-matched scores versus their actual scores on the matched assessment items for Physics 211.

achieved on the matched assessment items, $r(58) = 0.511$, $p < 0.01$ (see Fig. 2). This suggests that overall students who are doing well in the course recognize that they are doing well, and indicate that they believe they are able to answer correctly many of the items on which they will be assessed. These findings are consistent with those reported in other work [23]. As can be seen from Fig. 2, however, the data are very noisy. At the level of the individual student, there is a large variation in how accurately they assess their abilities. To compare our results with those reported by Dunning and Kruger as well as many others, we grouped our students by ability levels. We used students' overall performance on the course final exam as the best available metric of their actual "physics ability" at the end of Physics 211. Students were grouped into thirds ("bottom third," "middle third," and "top third") based on their exam score. When plotted as in Fig. 3, a pattern similar to the one reported by Dunning and Kruger (1999), and replicated in

FIG. 3. Average KS-matched scores and actual scores on the matched assessment items broken out by student groupings based on the overall score on the final exam for Physics 211.

many other studies [23,28,29] emerges: students in the top and middle third of the class appear to be well calibrated, while students in the bottom third tend to overestimate their abilities. These results appear to provide support for the Dunning-Kruger effect, suggesting that students who perform poorly on course assessments not only lack content knowledge, they also lack metacognitive awareness of themselves. In the next section, however, we present an alternate analysis of our data, which suggests a different interpretation of the results.

2. Comparing predictions and assessments for individual items

In addition to calculating an overall score on the KS and comparing this to the student's exam score, for each of the matched questions, we examined how a student's selection on the KS (A vs B or C) related to their actual performance on that item. The number of items to which a student had responded A and had subsequently given a correct response to was divided by the total number of items to which that student had responded A to determine the proportion of A items that were correct. These items could be considered "true positives"—students believed (according to their KS responses) that they could answer these questions correctly, and they went on to do so. The proportion of B and C items that were correct was similarly determined by dividing the number of items to which a student had responded B or C and had subsequently answered correctly by the total number of items to which they had responded B or C. These items could be considered "false negatives"—students' KS responses indicated that they did not believe they could answer these questions correctly, and yet they did so on the assessment. Results of these calculations are shown, broken out by student ability level, in Fig. 4. Note

FIG. 4. Average fraction of KS questions to which students had responded A or B or C which they then went on to answer correctly, broken out by student groupings based on overall final exam score for Physics 211. Error bars represent the 95% confidence interval of the mean.

that while the percentage of “A” questions and “B or C” questions to which students responded correctly appears similar within each group, these percentages are being calculated out of different values. For instance, students at the top of the class selected “A” much more frequently than they selected “B” or “C”, an average of 11.4 times and 3.4 times, respectively, out of the 16 matched questions. (The averages do not sum to 16 because not all students gave a response to every KS question.) For the students at the bottom and middle of the class, the values were much more equal, with students choosing “A” or “B or C” on average 7.9 times and 7.1 times, respectively, for the students in the middle of the class, and 7.1 times and 7.9 times, respectively, for students at the bottom of the class. Nine students were not included in this analysis as they had either never selected B or C or never selected A for any of the matched KS questions.

We expected that students would answer a larger percentage of items correctly for which they had selected A than they would items for which they had selected B or C. This was not the case. A mixed-methods analysis of variance did not reveal a significant main effect for the students’ selection of A vs B or C, $F(1, 48) = 2.096$, $p = 0.154$. In other words, in the physics class, students were just as likely to get right an item for which they had selected B or C as they were one for which they had selected A. This suggests that students may not be distinguishing correctly between what they know and what they do not know when making selections on the KS. There was also no interaction effect between students’ ability level and the proportion of A items that were correct vs the proportion of B and C items that were correct, $F(2, 48) = 1.443$, $p = 0.246$, suggesting that the lack of an ability to distinguish what they know from what they don’t know holds across students of all ability levels. Students in the top third of the class, for example, get right about 70% of the questions for which they have chosen A, but they also get right about 70% of the questions for which they have chosen B or C—suggesting that they are not correctly identifying all of the questions that they are able to answer correctly. For the bottom third of the class, students responded correctly on about 43% of the questions to which they had responded A, but were correct on only about 25% of the questions for which they had selected B or C. The disparity between the fraction of A items to which students in the bottom third of the class responded correctly and the fraction of B or C items to which they responded correctly appears larger than for the students in the top third of the class, although this difference did not rise to the level of statistical significance. At the very least, students who appear (based on a comparison between their overall KS-matched score and their actual score on the matched questions) to be more well calibrated are not actually distinguishing more clearly between which items they are able to answer correctly and which items they are

not able to answer correctly than are students who appear to be less well calibrated. This suggests that the method of using students’ predicted exam score and comparing to their actual exam score may obfuscate key aspects of students’ metacognitive abilities.

In addition to comparing the proportion of A items that were correct and the proportion of B or C items that were correct for each student, it was also possible to compare these proportions for each question, and determine whether there is a particular question type or topic that led to more accurate student judgments. Fisher’s exact test was calculated for each of the matched questions to determine whether the proportion of correct responses for that question differed by the judgment of ability made by students on the KS for that question. For the physics question set, only one of the sixteen matched questions yielded at significant result (Bonferroni corrected $p = 0.011$). This question is one of those shown in Fig. 1. It involved an object of known mass attached to a string being rotated in a vertical circle of known radius; students were given the velocity of the object at the bottom of its path, and asked to calculate the tension in the string at that point (exact values of the radius, the mass, and the velocity differed between the KS and the exam, but the questions were otherwise identical). Fifty-nine students responded to this item on both the KS and the final exam; of those, 17 responded to this item with an A on the KS, while the remaining 42 responded with a B or C. 15 of the 17 who responded A (88%) answered this question correctly on the final exam, while only 17 of 42 (40%) who responded B or C answered the question correctly on the final exam. On both the KS and the final exam, this question was immediately preceded by a similar question (also shown in Fig. 1) in which the object was being rotated in a *horizontal* circle. On that question, Fisher’s exact test did not reveal a significant difference between the proportion of correct responses provided at each KS judgment level (Bonferroni corrected $p = 1$). Sixteen out of 20 students (80%) who responded A answered the question correctly, while 27 out of 38 (71%) who responded B or C answered correctly.

B. Results from the chemistry course

As in Physics 211, there was an overall correlation between students’ KS-matched scores and the scores that they actually achieved on the matched assessment items for Chemistry 110, $r(58) = 0.465$, $p < 0.01$. For this course, however, when grouped into thirds based on their overall final exam scores, students in the top third of the class were underconfident while students in the middle and bottom thirds were somewhat overconfident (see Fig. 5). Again, this is consistent with the Dunning-Kruger effect and similar to results seen in equivalent courses when students postdict their exam scores [28].

In contrast to the Physics 211 students, Chemistry 110 students were, as a group, able to distinguish between the

FIG. 5. Average KS-matched scores and actual scores on the matched assessment items broken out by student groupings based on the overall score on the final exam for Chemistry 110.

items they could answer correctly and items they could not answer correctly. Results for the fraction of A items to which students responded correctly and B or C items to which the students responded correctly, broken out by student ability level, are shown in Fig. 6. A mixed-methods analysis of variance revealed a significant main effect for the students' selection of A vs B or C, $F(1, 57) = 17.503$, $p < 0.001$. However, the interaction effect between a student's selection of A vs B or C and their ability level was not significant, $F(2, 7) = 2.275$, $p = 0.112$. This suggests that once again students who scored well on the exam overall were not more successful than those who scored poorly at distinguishing between the items that they would get right and those that they would get wrong.

FIG. 6. Average fraction of KS questions to which students had responded A or B or C which they then went on to answer correctly, broken out by student groupings based on overall final exam score for Chemistry 110. Error bars represent the 95% confidence interval of the mean.

As in Physics 211, Fisher's exact test was calculated for each of the matched questions in Chemistry 110 to determine whether the proportion of correct responses for that question differed by the judgment of ability made by students on the KS. Two of the twenty-five matched questions yielded a significant result (Bonferroni-corrected $p = 0.05$ and 0.025 , respectively). Both of these questions were based on the topic of intermolecular forces, or the electrostatic attractions experienced by neighboring molecules as a result of uneven electron distribution.

V. DISCUSSION

The KS delivered in Physics 211 and the one delivered in Chemistry 110 were not expected to be analogous to one another. The courses in which they were deployed were taught in significantly different ways. Different authors developed the Knowledge Surveys, and the surveys emphasized different aspects of course content. In particular, there was a much higher emphasis on conceptual understanding in the physics KS and exam than on the chemistry KS or exam. Despite these differences, there are several features of the data that are strikingly similar. In both physics and chemistry courses, when average scores for the matched subset of the KS questions and the exam questions were compared, students who do poorly on the course exam appear to be overconfident, while those who score in the middle of the exam distribution appear to be more well calibrated. In both cases, however, when individual judgments were compared, there was no difference between the ability groups in terms of how well they were able to distinguish between those individual items they were able to answer correctly, and those that they were not able to answer correctly.

As noted above, we expected that students would respond correctly for a higher percentage of items to which they had responded A than for those to which they had responded B or C. In other words, a well-calibrated student would respond correctly to most of the items for which they had selected A (leading to a high "fraction of A correct") and would respond incorrectly to most of the items for which they had selected B or C (leading to a low "fraction of B-C correct"). For a student who is less metacognitively aware, these percentages would be expected to be more similar or possibly even reversed. If students at the top of the class are truly more metacognitively aware than students at the bottom of the class (as the Dunning-Kruger effect would suggest), then the disparity between "fraction of A correct" and "fraction of B or C correct" would be stronger for students at the top of the class than at the bottom of the class. In neither the physics course nor the chemistry course was this observed to be correct—in fact, the disparity between "fraction of A correct" and "fraction of B or C correct" appears, if anything, to be larger for students at the bottom of the class than for students in the middle or top of the class in both courses.

In Physics 211, students had access to the KS after they completed it. Thus it could be argued that perhaps the students in the top and middle thirds of the class were successful at studying from the KS before completing the final exam. That might result in their doing as well on the items for which they chose B or C as they did on those items for which they chose A. In Chemistry 110, however, students did not have access to the KS as a study aid after they had completed it. In that course, the disparity between the fraction of items to which students had responded “A” that they went on to answer correctly and the fraction of items to which students had responded “B” or “C” that they went on to answer correctly is relatively constant across the groups of students, regardless of the ability levels of the students. This suggests that the students in the bottom third of the class are not less successful from distinguishing the items they can answer from those they cannot answer, even in a course where students do not have the opportunity to study from the KS.

The role that the questions themselves play in contributing to students’ judgments of what they do or do not know remains an open question. Potgieter *et al.* found that students tend to be overconfident in their responses to items for which strong alternative conceptions exist [34]. In this study, we expected that such an effect might result in students being overconfident on the FCI items, for which there are frequently strong alternative conceptions [35]—i.e., it was expected that students would select A on the KS for the FCI items more frequently than they would for the final exam items, even if they were not actually more likely to get the FCI items right. Our observations support this expectation. Students in Physics 211 were much more likely to select A for the FCI questions than they were for questions drawn from the final exam. For the eight FCI items, the mean percentage of students responding A on the KS was 76%, while for the eight final exam items it was 43%. An independent-samples Mann-Whitney U test of the percentage of students who responded A for each of these questions revealed that there was a significant difference between these results, $p < 0.01$. On average, students were more likely to predict that they could answer an FCI question correctly than they were to predict that they could answer a final exam question correctly. The actual percent of students responding correctly to each of these questions, however, was 60% correct for the FCI questions and 50% correct for the final exam questions. This was not a significant difference (independent samples Mann Whitney U test, $p = 0.442$). This is consistent with the idea that students as a whole may be overconfident on items for which there exist strong alternative conceptions.

Another aspect of the questions that may influence student ability to determine whether or not they will answer a question correctly is the extent to which the question can be answered using plug-and-chug techniques with a simple memorized formula. As described above, the only question

on the physics KS for which students as a group were significantly able to discriminate whether or not they would be able to answer correctly involved forces and motion in the context of a vertical circle. Students were not able to discriminate at a significant level as to whether they would correctly answer a question involving motion in a horizontal circle. It is unclear why one of these questions would lead to better discrimination of student ability than the other, but one hypothesis is that most students were uncertain that they could answer either question correctly (more than 65% of students chose B or C on the KS for each of these questions), but were able to plug given values into a single formula ($F_c = mv^2/R$) and get a correct answer for the tension in the case of the horizontal circular motion (leading them to a correct answer on the exam). In other words, students may have been able to respond correctly to the horizontal circle question without understanding *why* their response was correct. Poor calibration for this question might therefore be indicative of the rote use of a formula without relevant conceptual understanding. This strategy would not work for the vertical circle question—only students who were able to construct correct free-body diagrams for the object, and set up and solve a relationship representing Newton’s second law for the object in question would be able to respond to the second question correctly. Most students who had these skills may have recognized that fact, leading them to select A on the KS.

The results from the chemistry course suggest that students might be able to recognize certain topics as being more challenging, and adjust their KS responses appropriately. As described above, both questions for which Fisher’s exact test yielded a significant result (indicating that students as a group were able to distinguish whether or not they could respond correctly) centered on the topic of intermolecular forces. This topic, unlike many others in general chemistry, requires the synthesis of a great many pieces of prior course knowledge. At the very least, students must have developed a mental model of matter as being composed of distinct, but interacting particles [39]. From there, students must know how to calculate the total number of valence electrons in a molecule, generate a Lewis structure, determine the molecular shape based on this structure, make predictions about the molecular polarity based on the shape, and then, finally, make a determination about the intermolecular attraction experienced by each of the molecules presented in the question. As in the physics class, there were other questions for which an equivalent or greater number of students responded B or C on the KS, but more of these students then went on to give a correct response on the exam. A correct response on the exam might thus not represent true understanding—if students did not realize that they would be able to respond correctly, they may have arrived at the correct answer by happening upon the correct formula without understanding

why it applies. For the intermolecular forces questions, students who have the necessary skills might recognize their ability to be successful, while students who are uncertain how to proceed would be unlikely to happen upon a correct response by rote, or by simply plugging into a memorized formula.

In any discussion about student confidence, a question of the effects of gender or membership in a group that is traditionally underrepresented in the sciences tends to arise. Unfortunately, the small numbers of women and underrepresented minorities enrolled in either Physics 211 or Chemistry 110 (only approximately 25%–30% of the students in the course are female or come from an underrepresented group) make it difficult to draw any conclusions about the effects that this status may have. Given the small numbers involved in our samples, no statistically significant effects were observed between results from males and females. We caution, however, that this does not necessarily imply that males and females were behaving in similar ways in making their selections on the KS, merely that our sample was too small to detect any differences.

VI. CONCLUSIONS

On one level, results from this study provide support for the Dunning-Kruger effect, even when students are asked to predict their performance before completing an assessment. (With a few exceptions [23], most prior reports of the Dunning-Kruger effect occurred when students were asked to estimate their score after completing an assessment.) Students in the bottom third of the class tend to overestimate their abilities, while students in the top third of the class or middle of the class appear to be more well-calibrated. An analysis of the data at the level of individual questions, however, reveals that in another sense students in the top third of the class are not necessarily more metacognitively aware than students in the bottom third of the class. They are not any more successful at distinguishing between the items they will answer correctly and those that they will answer incorrectly than are bottom students. It is possible that top students recognize that they are performing well in the class overall and give KS responses consistent with this observation, without actually distinguishing items that they are able to answer correctly from items that they are not able to answer correctly.

The results presented here support the idea that the ability to answer questions correctly is insufficient to demonstrate true understanding of a concept or idea. The ability to recognize that their understanding is correct is also a necessary part of true understanding. Many students are able to answer questions correctly (for instance, in the case of horizontal circular motion) even if

they do not indicate a belief in their ability to answer these questions correctly. In general, students are not successful at identifying all questions that they will be able to answer correctly, nor do they recognize all of those questions that they are unable to answer correctly. Students in the bottom third of the class in particular tend to overestimate their abilities. Students in the top third of the class went on to answer a larger percentage of the items for which they had selected A on the knowledge survey correctly than did students in the bottom third of the class, but students in all groups failed to identify some of the items that they were able to answer correctly, and failed to answer correctly some of those items that they believed they could. If students become more realistic in their self-assessments, and recognize the questions that they are not able to answer correctly, they might be better able to push themselves to develop appropriate understanding. This might also allow them to focus their studying and class preparation time to more precisely fit the topics they do not understand. In all classes, students of all ability levels, not just the poor performers, could benefit from instruction targeted at increasing their metacognitive awareness.

The results presented here suggest several implications for future research. In particular, the specific attributes of questions for which students are more successful at discriminating whether they will be answer correctly or not bear further examination. Research involving a larger pool of students might provide insights into the effects of gender and underrepresented status that were not readily apparent in this study. The results presented here indicate that simply examining the overall ability of students to predict their score on an assessment may not provide sufficient insights into the metacognitive skillfulness of a person or group. An examination of data at the level of individual judgments on individual topics provides a different view of students' metacognitive abilities, and should not be neglected in research studies. Finally, the extent to which students can be trained to be more metacognitively aware could be investigated. A future research project involves investigating whether students who are asked to practice making metacognitive judgments on a daily basis (using a personal response system such as iClicker) increase their skill at making accurate knowledge judgments over time.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Edward Bittner for allowing us access to his chemistry class for the purpose of taking data. Thanks are also due to the Office of Academic Affairs at Penn State Greater Allegheny, for providing financial support for this project.

- [1] L. C. McDermott and E. F. Redish, Resource letter: PER-1: Physics education research, *Am. J. Phys.* **67**, 755 (1999).
- [2] National Research Council, *Adapting to a Changing World—Challenges and Opportunities in Undergraduate Physics Education* (National Academies Press, Washington, DC, 2013).
- [3] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010101 (2006).
- [4] A. Elby, Helping physics students learn how to learn, *Am. J. Phys.* **69**, S54 (2001).
- [5] V. Sawtelle, E. Brewe, and L. H. Kramer, Exploring the relationship between self-efficacy and retention in introductory physics, *J. Res. Sci. Teach.* **49**, 1096 (2012).
- [6] M. Kryjevskaja, M. R. Stetzer, and N. Grosz, Answer first: Applying the heuristic-analytic theory of reasoning to examine student intuitive thinking in the context of physics, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020109 (2014).
- [7] L. C. McDermott and P. S. Shaffer and the Physics Education Group at the University of Washington, *Tutorials in Introductory Physics* (Prentice-Hall, Upper Saddle River, NJ, 2002).
- [8] L. C. McDermott and the Physics Education Group at the University of Washington, *Physics by Inquiry* (John Wiley & Sons, New York, 1996).
- [9] R. W. Chabay and B. A. Sherwood, *Matter and Interactions*, 3rd ed. (John Wiley & Sons, Hoboken, NJ, 2010).
- [10] E. Mazur, *Peer Instruction: A User's Manual* (Addison-Wesley, Reading, MA, 1996).
- [11] D. R. Sokoloff and R. K. Thornton, *Interactive Lecture Demonstrations* (John Wiley & Sons, Hoboken, NJ, 2004).
- [12] B. A. Lindsey and A. Y. Liu, Combining two reform curricula: An example from a course with well-prepared students, *Am. J. Phys.* **81**, 545 (2013).
- [13] P. W. Laws, Calculus-based physics without lectures, *Phys. Today* **44**, 24 (1991).
- [14] N. D. Finkelstein and S. J. Pollock, Replicating and understanding successful innovations: Implementing tutorials in introductory physics, *Phys. Rev. Spec. Top. - Phys. Educ. Res.* **1**, 010101 (2005).
- [15] M. Kohlmyer and M. Caballero, Tale of Two Curricula: The performance of 2000 students in introductory electromagnetism, *Phys. Rev. ST Phys. Educ. Res.* **5**, 020105 (2009).
- [16] C. H. Crouch and E. Mazur, Peer Instruction: Ten years of experience and results, *Am. J. Phys.* **69**, 970 (2001).
- [17] B. A. Lindsey, L. Hsu, H. Sadaghiani, J. W. Taylor, and K. Cummings, Positive attitudinal shifts with the Physics by Inquiry curriculum across multiple implementations, *Phys. Rev. ST Phys. Educ. Res.* **8**, 010102 (2012).
- [18] E. Brewe, A. Traxler, J. de la Garza, and L. H. Kramer, Extending positive CLASS results across multiple instructors and multiple classes of Modeling Instruction, *Phys. Rev. ST Phys. Educ. Res.* **9**, 020116 (2013).
- [19] A. Zohar and Y. J. Dori, Introduction, in *Metacognition in Science Education*, edited by A. Zohar and Y. J. Dori (Springer Netherlands, Dordrecht, 2012), pp. 1–19.
- [20] A. M. Glenberg and W. Epstein, Calibration of comprehension, *J. Exp. Psychol. Learn. Mem. Cogn.* **11**, 702 (1985).
- [21] N. Stone, Exploring the relationship between calibration and self-regulated learning, *Educ. Psychol. Rev.* **12**, 437 (2000).
- [22] D. B. May and E. Etkina, College physics students' epistemological self-reflection and its relationship to conceptual learning, *Am. J. Phys.* **70**, 1249 (2002).
- [23] P. Bell and D. Volckmann, Knowledge surveys in general chemistry: Confidence, overconfidence, and performance, *J. Chem. Educ.* **88**, 1469 (2011).
- [24] J. Kruger and D. Dunning, Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments, *J. Pers. Soc. Psychol.* **77**, 1121 (1999).
- [25] E. Nuhfer and D. Knipp, The knowledge survey: A tool for all reasons, *To Improv. Acad.* **21**, 59 (2003).
- [26] J. Ehrlinger, K. Johnson, M. Banner, D. Dunning, and J. Kruger, Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent, *Organ. Behav. Hum. Decis. Process.* **105**, 98 (2008).
- [27] 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).
- [28] S. Pazicni and C. F. Bauer, Characterizing illusions of competence in introductory chemistry students, *Chem. Educ. Res. Pract.* **15**, 24 (2014).
- [29] A. G. Karatjas, Comparing college students' self-assessment of knowledge in organic chemistry to their actual performance, *J. Chem. Educ.* **90**, 1096 (2013).
- [30] J. Krueger and R. a. Mueller, Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance, *J. Pers. Soc. Psychol.* **82**, 180 (2002).
- [31] K. A. Burson, R. P. Larrick, and J. Klayman, Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons, *J. Pers. Soc. Psychol.* **90**, 60 (2006).
- [32] G. T. Rosenthal, B. Soper, R. R. McKnight, A. W. Price, M. Boudreaux, and K. C. Rachal, Do students know if they answered particular questions correctly on a psychology exam?, *J. Instr. Psychol.* **37**, 57 (2010).
- [33] F. Sinkavich, Performance and metamemory: Do students know what they don't know?, *J. Instr. Psychol.* **22**, 77 (1995).
- [34] M. Potgieter, E. Malatje, E. Gaigher, and E. Venter, Confidence versus performance as an indicator of the presence of alternative conceptions and inadequate problem-solving skills in mechanics, *Int. J. Sci. Educ.* **32**, 1407 (2010).
- [35] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, *Phys. Teach.* **30**, 141 (1992).
- [36] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The force and motion

- conceptual evaluation and the evaluation of active learning laboratory and lecture, *Am. J. Phys.* **66**, 338 (1998).
- [37] B. A. Lindsey and M. L. Nagel, Assessing students' meta-cognitive calibration with knowledge surveys, *AIP Conf. Proc.* **1513**, 258 (2013).
- [38] N. Bowers, M. Brandon, and C. D. Hill, The use of a knowledge survey as an indicator of student learning in an introductory biology course, *Cell Biol. Educ.* **4**, 311 (2005).
- [39] H.-J. Schmidt, B. Kaufmann, and D. F. Treagust, Students' understanding of boiling points and intermolecular forces, *Chem. Educ. Res. Pract.* **10**, 265 (2009).