Analysis of the most common concept inventories in physics: What are we assessing?

James T. Laverty[1] and Marcos D. Caballero[2]

[1]*Department of Physics, Kansas State University, Manhattan, Kansas 66506, USA*
[2]*Department of Physics and Astronomy and CREATE for STEM Institute, Michigan State University, East Lansing, Michigan 48824, USA and Department of Physics and Center for Computing in Science Education, University of Oslo, N-0316 Oslo, Norway*

(Received 8 November 2017; published 12 April 2018)

Assessing student learning is a cornerstone of educational practice. Standardized assessments have played a significant role in the development of instruction, curricula, and educational spaces in college physics. However, the use of these assessments to evaluate student learning is only productive if they continue to align with our learning goals. Recently, there have been calls to elevate the process of science ("scientific practices") to the same level of importance and emphasis as the concepts of physics ("core ideas" and "crosscutting concepts"). We use the recently developed Three-Dimensional Learning Assessment Protocol to investigate how well the most commonly used standardized assessments in introductory physics (i.e., concept inventories) align with this modern understanding of physics education's learning goals. We find that many of the questions on concept inventories do elicit evidence of student understanding of core ideas, but do not have the potential to elicit evidence of scientific practices or crosscutting concepts. Furthermore, we find that the individual scientific practices and crosscutting concepts that are assessed using these tools are limited to a select few. We discuss the implications that these findings have on designing and testing curricula and instruction both in the past and for the future.

DOI: 10.1103/PhysRevPhysEducRes.14.010123

I. INTRODUCTION

Assessment helps us to understand what students know and are able to do after instruction, it aids us in understanding which aspects of a curriculum are working well for students and which are not, and it provides us with evidence of how well students are meeting our intended learning outcomes [1]. Taken as evidence of learning in physics, different forms of standardized assessment have helped shape many of the major changes that have occurred in physics education over the past 40 years [2–4]. Standardized assessment practices in undergraduate physics education emphasize the use of conceptual pre- and post-tests ("concept inventories")—the outcomes of which have been used to inform changes to curriculum design and instructional practices [5]. A wide variety of studies have been conducted using concept inventories [6–13], and student learning outcomes on such assessments are well documented [14–21].

Physics education researchers, curriculum developers, and instructors have used the outcomes of concept inventories to inform their work. But what are these inventories assessing? What learning goals were used to inform their design? And how well might these concept inventories represent an assessment of the learning outcomes in typical physics courses?

Physics education research (PER) has begun to address a wider variety of learning outcomes over the years [2,22,23]. Courses that were once focused heavily on conceptual understanding now include engagement in scientific practice, development of more sophisticated epistemologies, and achievement of positive attitudinal shifts towards physics. Curriculum design literature argues that aligning assessments and instruction with these goals is critical to helping students achieve these goals (e.g., "backwards design"). In particular, the assessments we use are meant to develop an evidentiary argument for student learning [1,24,25]. Arguably, the common concept inventories in physics are insufficient to address these broader learning outcomes. We are saddled with tools that provide some information, but this information is becoming increasingly incomplete for researchers, curriculum developers, and instructors. It is reasonable to ask: what do our current assessments tell us about student learning? That is, what are we assessing?

In this paper, we address these questions using the framework of three-dimensional learning (3DL), the blending of science process and content in the classroom (more detail can be found in Sec. III) [26]. While this lens

backgrounds a number of important issues (e.g., epistemological development and shifts in identity), it foregrounds engaging students in the process of science (scientific practices) and helping students develop how they organize their knowledge (core ideas and crosscutting concepts). Our analysis makes use of the recently developed Three-Dimensional Learning Assessment Protocol (3D-LAP) [27], a tool that evidences how well assessments provide opportunities to engage students in 3DL. Using the 3D-LAP, we coded the questions appearing on the four most common concept inventories [Force Concept Inventory (FCI), Force and Motion Conceptual Evaluation (FMCE), Brief Electricity and Magnetism Assessment (BEMA), Conceptual Survey of Electricity and Magnetism (CSEM)] to determine the degree to which they can provide evidence of 3DL. This paper provides a brief discussion of standardized assessment in physics (Sec. II), offers an overview of 3DL (Sec. III) with more details in Ref. [26], reviews the 3D-LAP (Sec. IV) and its use to analyze assessment tasks (Sec. V), but defers to Ref. [27] for details, and analyzes the four most common standardized assessments in physics using the 3D-LAP (Sec. VI). We provide concluding remarks in Sec. VII.

II. STANDARDIZED ASSESSMENT IN PHYSICS

Standardized assessment is widely used in physics education to measure learning outcomes in a variety of physics courses [7,14,15,18] including, most recently, upper-division courses [28–33]. It is typical to use these standardized assessments as "summative assessments" for a given course where they are used to gather evidence of what students have learned at the time that they take them, with little intention of using them to help those same students learn physics. That is, we typically assume (even without being explicit about it) that concept inventories attempt to elicit, identify, and track stable cognitive elements. Because of that stability, we neglect any learning that occurs during the assessment itself [34,35]. Some learning may occur when students interact with the measurement tool, but those effects are assumed to be small compared to the learning that has occurred over the time period that people are trying to measure (i.e., one semester) [36].

Concept inventories have typically focused on measuring "conceptual change" or "expertlike thinking." Their development has varied, but often follows a similar procedure [37]. This process usually starts by developing a large number of questions around the target concept—using the current literature on common misconceptions or difficulties around that concept as a guide. These initial questions are usually open ended and are presented to the target audience (students) under test conditions, in think-aloud interviews, or both. The developers then use the students' responses to eliminate or to modify questions that do not meet their standards (e.g., students did not interpret the question as intended or almost everyone got the question right). In addition, developers pay attention to common student responses to the questions. The questions that are deemed appropriate are then converted into multiple-choice questions where the distractor answer options match these most common incorrect responses. For open-ended assessments, it is common for the grading rubric to include the most common incorrect responses [28,29]. The test is readministered to students and modified as necessary until the developers are satisfied with the results. These results might be achieving some sort of stability in student performance, some set of appropriate test statistics, or both. Here, we do not intend to suggest that the development of concept inventories is straightforward or simple; it is not. There is certainly nuance in the design and development of specific inventories. However, the general process described above is quite similar to the development of the commonly used concept inventories in introductory physics.

The Force Concept Inventory is almost certainly the most well known and widely used standardized assessment in introductory physics courses [38]. Both it and the Force and Motion Conceptual Evaluation are designed to evaluate student learning of topics commonly found in the first semester of an introductory physics sequence [39]. Similarly, both the Conceptual Survey of Electricity and Magnetism and the Brief Electricity and Magnetism Assessment were developed to evaluate student learning of topics commonly taught in the second semester of an introductory physics sequence [40,41].

These (and other) concept inventories have provided straightforward, off-the-shelf ways to evaluate instructional practices and curricular materials [5]. Because of this, they have been used routinely to evaluate student learning in interactive environments [19,20,42], to compare student learning in different environments [43,44], and to investigate different learning outcomes for different groups of students within classes [10,42,45]. Using concept inventories in this way aligns with backward design; evidence should be collected to determine if instruction and curricula are helping all students achieve the learning goals we have for them. However, standardized assessments that gather evidence of student learning are only useful if they align with our learning goals. Recently, national reports have highlighted new ways to think about what we want our students to learn, both in K-12 and undergraduate science education. In particular, these reports have emphasized the idea of blending the concepts, on which concept inventories have been focused, and practices of science together into our learning goals [22,46,47].

III. EVOLVING LEARNING GOALS

Recent national calls have emphasized the need for students to engage with science and engineering practices at the same level of emphasis as they engage with science concepts [26,46]. Changes to courses aligned with these

calls broaden the scope of the learning goals in traditional introductory and advanced science courses and, as such, broaden the space for assessment. In physics, discussions of important practices have appeared in the revised advanced placement curriculum [47] and in white papers describing the need for new laboratory and computational experiences for physics students [22,23].

One national report endeavored to synthesize the years of research on student learning in science courses into recommendations for curricula and instruction. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (herein referred to as *Framework*) gives a comprehensive view of merging the concepts and process of science [26]. The underlying idea of the *Framework* is that having students engage in science in the manner that scientists do while using scientific knowledge is a more productive way to build students' understanding of both the process and knowledge of science. By focusing on blending concepts and practices together, we aim to provide our students with a deeper, richer, more enduring learning experience that is likely to benefit both their epistemological and identity development (even though these ideas are backgrounded by 3DL). To be clear, this is not the idea that we must "sacrifice" the content to make room for the process of science; it is that both the concepts and practices *are* the content. That is, while we might sacrifice some *concepts*, the inclusion of practices more than makes up for that sacrifice so that the total amount of *content* is not reduced. While the *Framework* was written for the K-12 education system, it has been argued that these ideas are relevant to higher education [46,48,49].

In this paper, we will use the ideas highlighted in the *Framework* as the basis for our analysis of the concept inventories to investigate how well our current assessments can provide evidence of learning of these broader goals. The *Framework* divides what we want students to learn into three "dimensions" of learning, one that is practice focused and two that are concept focused. A brief description of each of the three dimensions is given here along with an example. We encourage the reader to look at the *Framework* if they are interested in deeper explanations of the dimensions [26].

Scientific practices.—These are the disaggregated components of the process of science. They involve using scientific knowledge to model, predict, and explain phenomena (e.g., developing and using models).

Crosscutting concepts.—These bridge the boundaries between the disciplines of the physical, biological, and geological sciences. These "ways of thinking" are used by each discipline and can be leveraged to help students make connections across the sciences and between their classes (e.g., systems and system models).

Disciplinary core ideas.—These are the foundational concepts that are fundamental to the scientific discipline. In order to qualify as a disciplinary core idea, the concept must (1) be essential to the study of the discipline, (2) be required to explain a wide range of phenomena, and (3) provide a way to generate new ideas and predictions (e.g., energy).

The *Framework* emphasizes that it is vital that all three of these dimensions are blended into instruction, curriculum, and (most importantly for this article) assessments. Herein, we refer to the blending of these ideas as "three-dimensional learning."

In physics, we often use concept inventories to assess the outcomes in our courses (Sec. II), but how well do these inventories represent our shifting goals? In particular,

(1) How well do the four most commonly used concept inventories for introductory physics assess the goals of three-dimensional learning?
(2) For which, if any, of the Scientific Practices, Crosscutting Concepts, and Core Ideas do these concept inventories provide some evidence of student learning?

Note that the concept inventories that we are analyzing were developed well before the idea of three-dimensional learning. We understand that holding them to the standard that they should assess three-dimensional learning is not entirely fair. However, our goal here is not to disparage these assessments. They provide important information regarding conceptual learning in many courses and have helped advance PER in substantial ways. Instead, we aim to survey the current state of standardized assessment in physics education and use this as a step towards discussing the next generation of standardized assessments.

IV. THREE-DIMENSIONAL LEARNING ASSESSMENT PROTOCOL

In 2014, the National Research Council released a document highlighting the importance and challenges of developing assessments for the next generation science standards and (more broadly) three-dimensional learning [50]. To help identify and develop assessments that are capable of eliciting evidence of students engaging with each of the three dimensions, we developed the Three-Dimensional Learning Assessment Protocol [27].

The 3D-LAP was designed with two central purposes:

(1) to help researchers characterize how well assessments align with each of the three dimensions,
(2) to help instructors develop or modify existing assessment tasks so that they have the potential to elicit evidence of students engaging with the three dimensions.

The 3D-LAP uses individual questions or clusters of related questions (referred to herein as a "task") as the unit of analysis. By analyzing only the task itself, the 3D-LAP can be used to determine if the task has the potential to elicit evidence that a student will engage in a scientific practice, crosscutting concept, or core idea [27].

The 3D-LAP was developed as part of a larger project to transform the introductory physics, chemistry, and biology courses at Michigan State University. The development team [made up of the authors and eight additional disciplinary experts, many of whom identify as discipline-based education researchers (DBER)] initially developed a prototype set of criteria for each of the scientific practices, crosscutting concepts, and core ideas based on their descriptions in the *Framework*. Separately, we collected and discussed assessment tasks that exemplified each of the dimensions. We then compared these exemplar tasks with the prototype criteria and used this comparison to revise and refine the criteria [51]. The final criteria took different forms for each dimension: scientific practices each have a list of 2–4 criteria, all of which must be met in order for a task to align with that scientific practice; crosscutting concepts each have a brief description of what is necessary to align with it; and each core idea comes with a list of ideas, at least one of which must be included in a task to qualify as aligning with a core idea.

Both the face and content validity of the 3D-LAP as applied to concept inventories is evidenced by the expertise of the development team. This team included disciplinary experts from physics, chemistry, and biology, some of whom identify as discipline-based education researchers and others that identify as more traditional experts. The development process reinforced the validity of the protocol by continually comparing the theory (*Framework*, research literature, etc.) and the on-the-ground reality (existing assessments). Some of these comparisons included assessment tasks from existing concept inventories in each of the disciplines.

In addition to the validity of the 3D-LAP, it is important to show that multiple coders get the same results when applying it to assessment tasks. In order to establish the reliability of the 3D-LAP when applied to these concept inventories, J. T. L. coded all of the tasks, while M. D. C. coded 25% of the tasks chosen randomly. Cohen's κ is a commonly used measure of interrater reliability for two coders and it does well when looking for levels of agreement in many cases [52]. However, Cohen's κ does yield unexpected and uninformative values when the code appears in almost none (or almost all) of the cases, which is often the case when using the 3D-LAP [53]. It is precisely because of these cases that Gwet's AC_1 was introduced [54]. Gwet's AC_1 is an alternative, more stable measure of agreement, even in cases where the codes appear very (in)frequently.

Our interrater reliability was established using Gwet's AC_1 statistic, obtaining a value of 0.93, 0.79, and 0.91, respectively, for the scientific practices, crosscutting concepts, and core ideas [54]. These values are typically considered good to very good agreement. For these purposes, we only check to see if both coders agreed that the task elicited a dimension or not, without regard to which component of the dimension was coded (i.e., if there is a scientific practice or not, not necessarily which scientific practice). This choice was made because we do not have the sufficient number of tasks needed to investigate the reliability of all 19 components of the 3D-LAP (7 scientific practices, 7 crosscutting concepts, and 5 core ideas).

V. APPLYING THE 3D-LAP

Here, we demonstrate how the 3D-LAP can be applied to assessment tasks in our data set, one that aligns with three-dimensional learning and one that does not. Because concept inventories require significant effort to develop and that effort can be compromised by making the inventories available to the public, we will not reprint any part of them here. Instead, we will describe two questions from the BEMA and refer the interested reader to the original exams for the exact questions [41].

A. Example 1: Alignment with one dimension

Question 19 of the BEMA asks students about the difference in electric potential between any two points in a metal. The answer options all include a declaration of what that potential difference is and a few words that are about either the value of the electric field (answer) or a common incorrect response.

Using the 3D-LAP, we characterize question 19 of the BEMA as providing no evidence that a student has engaged in a scientific practice or crosscutting concept; however, it does elicit the core idea of "interactions are mediated by fields." The most closely associated scientific practice is constructing explanations and engaging in argument from evidence. Column 2 of Table I shows an analysis of the task to determine if it elicits this practice. As shown in Table I, question 19 of the BEMA does ask the student to make a claim about the described situation, but does not present an event, observation, or phenomenon, or ask the student to select evidence or reasoning to support their claim. A student certainly might engage in the practice, but the question as written does not provide any evidence that they are being asked to do so. Similarly, this task does not elicit any of the crosscutting concepts as determined by the 3D-LAP. The most closely associated crosscutting concept is cause and effect: mechanism and explanation. The 3D-LAP criteria for this crosscutting concept is

> *To code an assessment task with cause and effect: mechanism and explanation, the question provides at most two of the following: (1) a cause, (2) an effect, and (3) the mechanism that links the cause and effect, and the student is asked to provide the other(s).*

Question 19 of the BEMA does not ask the student to explain the mechanism that connects the cause to the effect. Unlike with the scientific practices and crosscutting

TABLE I. An analysis of question 19 and question 7 of the BEMA using the 3D-LAP criteria for the scientific practice of constructing explanations and engaging in argument from evidence. An assessment task must meet all of the criteria in order for it to be considered to elicit that dimension.

3D-LAP criteria for aligning with constructing explanations and engaging in argument from evidence	Characterization of BEMA question 19 with 3D-LAP criteria	Characterization of BEMA question 7 with 3D-LAP criteria
1. Question gives an event, observation, or phenomenon.	✗ 1. The question does not present a real-life situation (it takes place in an idealized model).	✓ 1. This question is about a real-world scenario.
2. Question gives or asks student to select a claim based on the given event, observation, or phenomenon.	✓ 2. Question asks student to claim that the potential difference is zero or nonzero.	✓ 2. Question asks student to claim whether or not the rubber sheet is affected by the wall.
3. Question asks student to select scientific principles or evidence in the form of data or observations to support the claim.	✗ 3. Most answer options do not include scientific principles (charge, electric field).	✓ 3. Most answer options include scientific principles (charge, repulsion, polarization).
4. Question asks student to select the reasoning about why the scientific principles or evidence support the claim.	✗ 4. Answer options do not include the reasoning linking the principle and the claim.	✓ 4. Most answer options include reasoning that connects the principles to the claim.

concepts, the task does elicit evidence that a student has engaged with the core idea of interactions are mediated by fields, as the task asks the student specifically about the electric potential (and the correct answer includes the electric field).

B. Example 2: Alignment with three dimensions

Question 7 of the BEMA asks about the interactions between a charged object (wall) and a neutral object (rubber sheet). Each answer option includes a description of what will happen to the rubber sheet and a possible reason why. In contrast to question 19, question 7 of the BEMA does provide evidence that students can engage in a scientific practice, crosscutting concept, and core idea (at least as well as can be done in a multiple-choice question). Column 3 of Table I shows the analysis of this task and gives a brief explanation of why it does align with the criteria for the scientific practice of constructing explanations and engaging in argument from evidence. This task also elicits a crosscutting concept: structure and function. The 3D-LAP criteria for this crosscutting concept is

> *To code an assessment task with structure and function, the question asks the student to predict or explain a function or property based on a structure, or to describe what structure could lead to a given function or property.*

Question 7 asks the student to use the atomic structure of the rubber sheet to predict the behavior of the sheet in response to the charged wall. Like question 19, question 7 also elicits evidence that a student has engaged with the core idea of interactions are mediated by fields.

VI. RESULTS OF CODING CONCEPT INVENTORIES

Looking at the results of coding each question on a concept inventory in aggregate allows us to understand for which aspects of student learning the assessments are eliciting evidence. We have weighted the results of coding with the 3D-LAP using the percentage of points assigned to each question by the inventory authors to address our first research question: How well do the four most commonly used concept inventories for introductory physics assess the goals of three-dimensional learning? We first provide an overview and then discuss results for each concept inventory in turn.

Figure 1 shows that few of the tasks on the concept inventories address all three dimensions. However, most of the tasks do assess at least one of the three dimensions, and few assess no dimensions.

Figure 2 provides a clearer picture of what the current concept inventories are assessing in terms of 3DL. In each concept inventory, the majority of tasks have the potential to elicit evidence of core ideas. Given that these tests were designed to assess conceptual learning, this is what we would expect to find. This also suggests that the 3D-LAP is capable of identifying the kinds of questions that assess important concepts in physics. Crosscutting concepts are assessed significantly less frequently than the core ideas and scientific practices are almost never assessed by these concept inventories. This suggests that concept inventories are assessing students' knowledge about physics concepts (a worthy goal, to be sure), but not necessarily their ability to do physics with those concepts.

a. FCI.—Our coding of the FCI demonstrates that few items have the potential to engage the student with more

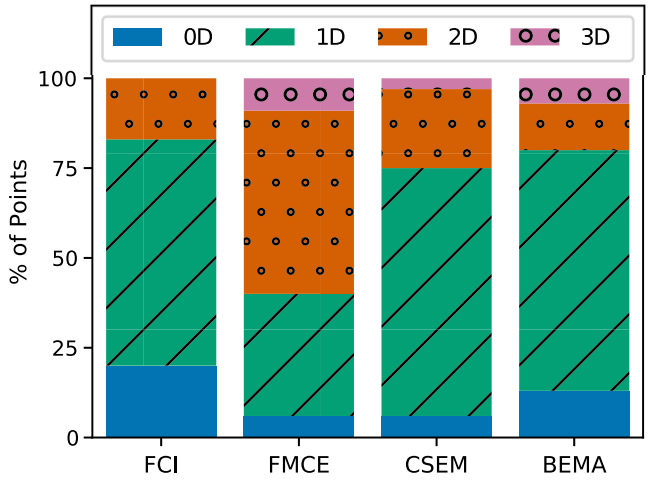

FIG. 1. Percentage of points for each concept inventory that have the potential to elicit evidence of zero, one, two, or all three dimensions.

than one dimension. In fact, no items of the FCI were coded as three-dimensional (Fig. 1). Most of the points that can be awarded to students on the FCI are for one-dimensional questions (63%). There are a small fraction of points awarded for two-dimensional questions (17%) with the rest of the points awarded for answer questions with no dimensions (20%). A close look at Fig. 2 shows why this is the case, 73% of the points can be awarded for questions focused on core ideas. Only a minority of the points are awarded for answering questions that can elicit a crosscutting concept (17%) or scientific practice (7%), so there is very little chance of overlap between the dimensions.

b. FMCE.—The FMCE provides more evidence of 3DL than the FCI (Fig. 1). A small fraction of points on the FMCE (9%) are awarded for answering three-dimensional questions and the majority of points awarded on the FMCE

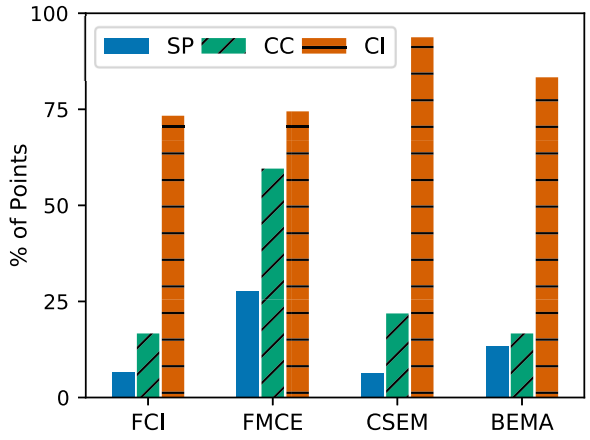

FIG. 2. Percentage of points for each concept inventory assigned to items that have the potential to elicit evidence of a scientific practice (SP), crosscutting concept (CC), or core idea (CI).

are available for answering two-dimensional questions (51%). The FMCE has few points awarded for questions with no dimensions (6%), but a fair percentage for one-dimensional questions (34%). Figure 2 illustrates that the larger percentage of points for two- and three-dimensional questions stem from the greater number of points allotted to assessing Crosscutting Concepts (60%) and Science Practices (28%)—leading to greater overlap with the Core Ideas (74%).

c. CSEM.—For the CSEM (Fig. 1), we again find the greatest number of points available is allotted to one-dimensional questions (69%). Questions with no dimensions (6%) and three-dimensional questions (3%) comprise a minority of the test. Nearly one-quarter of points (22%) are available for answering two-dimensional questions. We find that the majority of points available (94%) appear on questions that contain a core idea (Fig. 2). This result coupled to the low percentage of questions containing a crosscutting concept (22%) or scientific practice (6%) explains the large number of one-dimensional questions on the CSEM.

d. BEMA.—The BEMA is quite similar to the CSEM and FCI (Fig. 1) in that it has a large fraction of points allotted to one-dimensional questions (67%) with few points given for answers to zero-dimensional (13%), two-dimensional (13%), and three-dimensional questions (7%). This result is explained similarly to the CSEM by the observation that the majority of points available on the BEMA are for answering questions with a core idea (83%) while the points available for answering questions aligning with a crosscutting concept (17%) and scientific practice (13%) are low (Fig. 2).

e. Comparing common assessments.—The FCI and FMCE are often used in introductory courses to test students' conceptual understanding of classical mechanics. We have found that these assessments differ in the degree to which they assess three-dimensional learning. In fact, a contingency table analysis of this result shows that the frequency of tasks aligning with zero, one, two, and three dimensions is notably different between the two exams ($\chi^2 = 42.2$, $p \ll 0.05$, $\nu = 3$). We interpret this as suggestive that the FMCE is a better, albeit incomplete, measure of three-dimensional learning in physics when compared to the FCI. We find a similar, but not quite significant, association for the CSEM and BEMA ($\chi^2 = 6.5$, $p = 0.08$, $\nu = 3$). However, here it is less clear which may be the better measure of 3DL, as the CSEM has a higher percentage of points aligning with crosscutting concepts and core ideas, while the BEMA has a higher percentage aligning with scientific practices.

f. Presence of specific components of 3DL.—While the analysis above provides an indication of the presence or absence of the potential to elicit evidence of a student engaging with scientific practices, crosscutting concepts, and core ideas, identifying the specific components that

TABLE II. The scientific practices, crosscutting concepts, and core ideas that are potentially being assessed by each of the four concept inventories.

	Scientific practices	Crosscutting concepts	Core ideas
FCI	Analyzing and interpreting data	Scale, proportion, and quantity	Interactions can cause changes in motion
FMCE	Using mathematics and computational thinking Constructing explanations and engaging in argument from evidence	Scale, proportion, and quantity Stability and change	Interactions can cause changes in motion Energy is conserved
CSEM	Using mathematics and computational thinking	Scale, proportion, and quantity	Interactions can cause changes in motion Interactions are mediated by fields Energy is conserved
BEMA	Using mathematics and computational thinking Constructing explanations and engaging in argument from evidence	Scale, proportion, and quantity Structure and function	Interactions can cause changes in motion Interactions are mediated by fields

appear in each concept inventory requires that we delve more deeply into the coded data. Here, we identify which scientific practices, crosscutting concepts, and core ideas appear on each concept inventory at least once. Through this analysis we aim to answer our second research question: For which of the Scientific Practices, Crosscutting Concepts, and Core Ideas do these concept inventories provide some evidence of student learning?

Table II lists which dimensions appear at least once on each of the concept inventories. Only three (of seven) scientific practices, three (of seven) crosscutting concepts, and three (of five) core ideas are potentially assessed by these four concept inventories. Within the scientific practices, "using mathematics and computational thinking" came up in three of the concept inventories. The crosscutting concept of scale, proportion, and quantity appeared on all four, and the core idea of interactions can cause changes in motion appears on all of them.

VII. CONCLUSION AND DISCUSSION

We used the lens of three-dimensional learning to analyze four of the most common concept inventories used by the physics community to see how well they can assess both students' knowledge of physics concepts and students' abilities to use those concepts to do physics [26,46]. Using the 3D-LAP, we found that almost all of the tasks on these assessments align with at least one of the three dimensions originally defined by the *Framework*, but very few align with all three ($< 10\%$ on each concept inventory). Further analysis suggests that the alignment with dimensions is biased towards traditional conceptual goals, with evidence of eliciting the core ideas being much more common ($>70\%$ on each) than the scientific practices ($<25\%$ on each). Evidence of the crosscutting concepts being elicited was also low, though the FMCE does have notably more tasks aligned with crosscutting concepts than the other three conceptual inventories.

Each concept inventory did align with each of the three dimensions on at least one task. Across the four concept inventories, three of them included at least one task that aligned with the scientific practice of using mathematics and computational thinking. All four contained at least one task that aligned with the crosscutting concept of scale, proportion, and quantity, and the core idea of interactions can cause changes in motion.

While our analysis reveals a number of shortcomings with the most widely used assessments for introductory physics, the work is not without shortcomings. Analyzing these concept inventories using the 3D-LAP means we are looking at whether or not the tasks align with each of the dimensions of 3DL and almost nothing else. We take for granted that it is important for students to be assessed on both the practices and concepts of physics. We do not analyze aspects such as how students interpret the questions, the context in which the assessments are given, or other ways to analyze questions that are known to influence how students respond to them, such as bias and readability [1].

Nevertheless, these results suggest that concept inventories are not productive for gathering evidence of student learning that aligns with three-dimensional learning, particularly with regard to scientific practices and crosscutting concepts. Again, our goal here is not to disparage concept inventories; they were designed to measure students' conceptual understanding and not to align with three-dimensional learning. Our goal was to determine how productive these existing assessments are from the lens of assessing three-dimensional learning, which came along later. This study suggests that there is room for improvement when it comes to aligning standardized assessments in college-level physics with modern learning goals such as engagement in scientific practices. Further, this study suggests that the ability of concept inventories to obtain evidence that students are meeting modern learning goals are tenuous at best.

As discussed in Sec. II, concept inventories have played a vital role in changing the way introductory physics courses are taught and the curricula used for those courses. However, another perspective is that the changes to

curriculum and instruction that have proliferated in physics education would not have succeeded if they did not improve students' scores on concept inventories. It is hard to imagine any of these reforms being successful if the students' gains on the relevant concept inventory were lower in the new environment than in a traditional environment. In the PER community, researchers have developed other methods to investigate student learning as part of their research (e.g., affective measures, interviews, etc.), which might temper this sentiment, but for traditional physics faculty who use these assessments, we may be driving them toward "maximizing" a kind of learning that does not necessarily align with our modern understanding of what we want students to learn [26,46]. It is important to improve standardized assessments in the near future because they can drive curricular and pedagogical change in physics and, thus, have a significant impact on student learning at a large scale.

Given all of this, one question likely jumps to mind: What does a test that assesses scientific practices, crosscutting concepts, and core ideas look like? We do not claim to know the answer to this question, but we will speculate. Such an assessment will likely include all three dimensions in most, but perhaps not all, of its questions. There are certainly things we may want to assess about the dimensions that do not quite rise to the level of being a three-dimensional question. These assessments may also necessarily include tasks that are not multiple choice. For example, it is hard to imagine a task that can assess "constructing explanations" if the task only requires students to *select* an explanation. This has the unfortunate side-effect of making the assessment much more difficult to administer and grade, but there may be ways to turn them back into multiple-choice questions by coupling them together [55].

In the future, we aim to develop standardized assessments that align more fully with three-dimensional learning. There is certainly work being done that such assessments should inform the design of these future assessments [56,57]. Such assessments should be more capable of assessing students' abilities to use the centrally important ideas of physics to model, investigate, analyze, predict, and explain real-world phenomena. Additionally, we intend that such assessments communicate to the larger physics community that our learning goals are shifting to include both concepts and practices.

ACKNOWLEDGMENTS

The authors would like to thank Kansas State University's Department of Physics and the Association of American Universities' STEM Education Initiative for their support. Additionally, we would like to thank the discipline-based education researchers (DBER) community at Michigan State University (especially those involved in the development of the 3D-LAP).

[1] National Research Council, *Knowing What Students Know: The Science and Design of Educational Assessment* (National Academies Press, Washington, DC, 2001).
[2] National Research Council, *Adapting to a Changing World-Challenges and Opportunities in Undergraduate Physics Education* (National Academies Press, Washington, DC, 2013).
[3] L. C. McDermott and E. Redish, Resource Letter: PER-1: Physics Education Research, Am. J. Phys. **67**, 755 (1999).
[4] D. E. Meltzer and R. K. Thornton, Resource Letter ALIP–1: Active-Learning Instruction in Physics, Am. J. Phys. **80**, 478 (2012).
[5] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource Letter RBAI-1: Research-Based Assessment Instruments in Physics and Astronomy, Am. J. Phys. **85**, 245 (2017).
[6] S. J. Pollock and N. D. Finkelstein, Sustaining educational reforms in introductory physics, Phys. Rev. ST Phys. Educ. Res. **4**, 010110 (2008).
[7] R. Beichner, L. Bernold, E. Burniston, P. Dail, R. Felder, J. Gastineau, M. Gjertsen, and J. Risley, Case study of the physics component of an integrated curriculum, Am. J. Phys. **67**, S16 (1999).
[8] E. F. Redish, J. M. Saul, and R. N. Steinberg, On the effectiveness of active-engagement microcomputer-based laboratories, Am. J. Phys. **65**, 45 (1997).
[9] C. M. Sorensen, A. D. Churukian, S. Maleki, and D. A. Zollman, The New Studio format for instruction of introductory physics, Am. J. Phys. **74**, 1077 (2006).
[10] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).
[11] C. Hoellwarth, M. J. Moelter, and R. D. Knight, A direct comparison of conceptual learning and problem solving ability in traditional and studio style classrooms, Am. J. Phys. **73**, 459 (2005).
[12] N. Lasry, E. Mazur, and J. Watkins, Peer instruction: From Harvard to the two-year college, Am. J. Phys. **76**, 1066 (2008).
[13] E. E. Prather, A. L. Rudolph, G. Brissenden, and W. M. Schlingman, A national study assessing the teaching and learning of introductory astronomy. Part I. The effect of interactive instruction, Am. J. Phys. **77**, 320 (2009).
[14] E. Etkina and A. Van Heuvelen, in *Research-Based Reform of University Physics* (2007), Vol. 1.

[15] C. H. Crouch and E. Mazur, Peer Instruction: Ten years of experience and results, Am. J. Phys. **69**, 970 (2001).
[16] C. Crouch, J. Watkins, A. Fagen, and E. Mazur, in *Research-Based Reform of University Physics*, Vol. 1 (Ref. [14]).
[17] R. Beichner, in *Research-Based Reform of University Physics* (Ref. [14]), Vol. 1.
[18] R. Chabay and B. Sherwood, in Research-Based Reform of University Physics (Ref. [14]), Vol. **1**.
[19] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).
[20] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. U.S.A. **111**, 8410 (2014).
[21] J. Von Korff, B. Archibeque, K. A. Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: A 50 k-student study, Am. J. Phys. **84**, 969 (2016).
[22] J. Kozminski, N. Beverly, D. Deardorff, R. Dietz, M. Eblen-Zayas, R. Hobbs, H. Lewandowski, S. Lindaas, A. Reagan, R. Tagg, J. Williams, and B. Zwickl, AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum (2014), https://www.aapt.org/Resources/upload/LabGuidlinesDocument_EBendorsed_nov10.pdf.
[23] E. Behringer, AAPT UCTF Computational Physics Report, 2017, https://www.aapt.org/Resources/upload/AAPT_UCTF_CompPhysReport_final_B.pdf.
[24] G. P. Wiggins and J. McTighe, *Understanding by Design* (Association for Supervision and Curriculum Development, 2005), ISBN 13:978-1416600350, ISBN-10:1416600353, google-Books-ID: N2EfKlyUN4QC.
[25] J. Biggs, Enhancing teaching through constructive alignment, High Educ. **32**, 347 (1996).
[26] National Research Council, *A Framework for K-12 Science Education* (National Academies Press, Washington, DC, 2012).
[27] J. T. Laverty, S. M. Underwood, R. L. Matz, L. A. Posey, J. H. Carmel, M. D. Caballero, C. L. Fata-Hartley, D. Ebert-May, S. E. Jardeleza, and M. M. Cooper, Characterizing college science assessments: The three-dimensional learning assessment protocol, PLoS One **11**, e0162333 (2016).
[28] M. D. Caballero and S. J. Pollock, Assessing student learning in middle-division classical mechanics/math methods, in *Physics Education Research Conference 2013, Portland, Oregon*, PER Conference (2014), pp. 81–84.
[29] S. V. Chasteen, R. E. Pepper, M. D. Caballero, S. J. Pollock, and K. K. Perkins, Colorado Upper-Division Electrostatics diagnostic: A conceptual assessment for the junior level, Phys. Rev. ST Phys. Educ. Res. **8**, 020108 (2012).
[30] S. Wuttiprom, M. D. Sharma, I. D. Johnston, R. Chitaree, and C. Soankwan, Development and use of a conceptual survey in introductory quantum physics, Int. J. Sci. Educ. **31**, 631 (2009).
[31] S. Goldhaber, S. Pollock, M. Dubson, P. Beale, and K. Perkins, Transforming upper-division quantum mechanics: Learning goals and assessment, AIP Conf. Proc. **1179**, 145 (2009).
[32] S. McKagan and C. Wieman, Exploring student understanding of energy through the quantum mechanics conceptual survey, AIP Conf. Proc. **818**, 65 (2006).
[33] E. Cataloglu, Development and validation of an achievement test in introductory quantum mechanics: The quantum mechanics visualization instrument (QMVI), Ph.D. thesis, The Pennsylvania State University, 2001.
[34] E. C. Sayre, Plasticity: Resource justification and development, Ph.D. thesis, The University of Maine, 2007.
[35] E. Sayre and M. Wittmann, Plasticity of intermediate mechanics students' coordinate system choice, Phys. Rev. ST Phys. Educ. Res. **4**, 020105 (2008).
[36] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes* (Harvard University Press, Cambridge, MA, 1980).
[37] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33**, 1289 (2011).
[38] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).
[39] 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 curricula, Am. J. Phys. **66**, 338 (1998).
[40] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).
[41] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).
[42] E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O'Brien, I. Rodriguez, and P. Pamela, Toward equity through participation in Modeling Instruction in introductory university physics, Phys. Rev. ST Phys. Educ. Res. **6**, 010106 (2010).
[43] M. Kohlmyer, M. Caballero, R. Catrambone, R. Chabay, L. Ding, M. Haugan, M. Marr, B. Sherwood, and M. Schatz, Tale of two curricula: The performance of 2000 students in introductory electromagnetism, Phys. Rev. ST Phys. Educ. Res. **5**, 020105 (2009).
[44] M. D. Caballero, E. F. Greco, E. R. Murray, K. R. Bujak, M. Jackson Marr, R. Catrambone, M. A. Kohlmyer, and M. F. Schatz, Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study, Am. J. Phys. **80**, 638 (2012).
[45] A. Madsen, S. B. McKagan, and E. C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).
[46] National Research Council, *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering* (National Academies Press, Washington, DC, 2012).
[47] The College Board, AP Physics 1: Algebra-Based and AP Physics 2: Algebra-Based Curriculum Framework, https://

apcentral.collegeboard.org/pdf/ap-physics-1-course-and-exam-description.pdf (2015).

[48] M. M. Cooper, M. D. Caballero, D. Ebert-May, C. L. Fata-Hartley, S. E. Jardeleza, J. S. Krajcik, J. T. Laverty, R. L. Matz, L. A. Posey, and S. M. Underwood, Challenge faculty to transform STEM learning, Science **350**, 281 (2015).

[49] J. McDonald, Point of view: The next generation science standards: impact on college science teaching, J. Coll. Sci. Teach. **045**, 13 (2015).

[50] National Research Council, *Developing Assessments for the Next Generation Science Standards* (National Academies Press, Washington, DC, 2014).

[51] S. M. Kolb, Grounded theory and the constant comparative method: valid research strategies for educators, J. Emerging Trends Educ. Res. Policy Stud. **3**, 83 (2012).

[52] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. **20**, 37 (1960).

[53] D. V. Cicchetti and A. R. Feinstein, High agreement but low kappa: II. Resolving the paradoxes, Journal of clinical epidemiology **43**, 551 (1990).

[54] K. L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, Brit. J. Math. Stat. Psychol. **61**, 29 (2008).

[55] B. R. Wilcox and S. J. Pollock, Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020124 (2014).

[56] E. Etkina, A. Van Heuvelen, S. White-Brahmia, D. Brookes, M. Gentile, S. Murthy, D. Rosengrant, and A. Warren, Scientific abilities and their assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 020103 (2006).

[57] J. Day and D. Bonn, Development of the concise data processing assessment, Phys. Rev. ST Phys. Educ. Res. **7**, 010114 (2011).