Resource Letter RBAI-2: Research-based assessment instruments: Beyond physics topics

Adrian Madsen, Sarah B. McKagan, Eleanor C. Sayre, and Cassandra A. Paul

Citation: American Journal of Physics **87**, 350 (2019); doi: 10.1119/1.5094139

View online: https://doi.org/10.1119/1.5094139

View Table of Contents: https://aapt.scitation.org/toc/ajp/87/5

Published by the American Association of Physics Teachers

RESOURCE LETTER

Resource Letter RBAI-2: Research-based assessment instruments: Beyond physics topics

Adrian Madsen and Sarah B. McKagan
American Association of Physics Teachers, College Park, Maryland 20740

Eleanor C. Sayre
Department of Physics, Kansas State University, Manhattan, Kansas 66506

Cassandra A. Paul
San Jose State University, San Jose, California 95192

(Received 18 October 2018; accepted 26 February 2019)

This Resource Letter provides a guide to research-based assessment instruments (RBAIs) that can be used in physics classes to assess attitudes and beliefs about physics, epistemologies and expectations, the nature of physics, problem solving, self-efficacy, reasoning skills, and lab skills. We also discuss RBAIs in physics cognate fields, such as mathematics, and observation protocols for standardized observation of teaching. In this Resource Letter, we present an overview of these assessments and surveys including research validation, instructional level, format, and themes, to help faculty find the assessment that most closely matches their goals. This Resource Letter is a companion to "RBAI-1: Research-based Assessment Instruments in Physics and Astronomy," which explicitly dealt with physics and astronomy topics. More details about each RBAI discussed in this paper are available at PhysPort: www.physport.org/assessments.
https://doi.org/10.1119/1.5094139

I. INTRODUCTION

In the first Resource Letter in this series (RBAI-1),[1] we presented 40+ research-based assessment instruments (RBAIs) developed and used by the physics and astronomy education community to assess student understanding of the physics and astronomy content. Here, we present RBAIs used by the physics and science education research communities to examine non-physics-content topics. In our interviews with physics faculty, we have learned that faculty want to assess not only students' conceptual learning but also the skills and attitudes that faculty hope that students gain as a result of their physics course.[2] This Resource Letter is meant to help faculty find and use the assessments that are applicable to their students and their goals beyond the content.

As we look at RBAIs beyond physics topics, we had to decide what to include and not to include because the space of cognate fields is quite large, and researchers in discipline-based education research and the psychological sciences have developed hundreds of RBAIs over decades. In this Resource Letter, we choose to include RBAIs that physics faculty are likely to find most helpful for assessing their own classes. Towards that end, we have not included instruments for programmatic assessment or those intended primarily for use by researchers.

We begin with a general discussion of the RBAIs included in this Resource Letter and their research validation (Sec. II) and then discuss specific RBAIs in several major categories. These RBAIs cover a diverse set of topics including mathematics (Sec. III), attitudes and beliefs, including nature of science and self-efficacy (Sec. IV), problem-solving (Sec. V), scientific reasoning (Sec. VI), laboratory skills (Sec. VII), and observation protocols at a range of levels from high school to graduate school (Sec. VIII).

More details about each of these RBAIs are available at www.physport.org/assessments,[3] where verified educators can download most of the RBAIs. Details available on PhysPort include information about administering, scoring, interpreting results, version history, translations, research validation, and more.

1. "Resource Letter RBAI-1: Research-based assessment instruments in physics and astronomy," A. Madsen, S. B. McKagan, and E. C. Sayre, Am. J. Phys. **85**(4), 245–264 (2017). (E)

2. “Research-based assessment affordances and constraints: Perceptions of physics faculty,” A. Madsen, S. B. McKagan, M. “Sandy” Martinuk, A. Bell, and E. C. Sayre, Phys. Rev.-Phys. Educ. Res. **12**, 010115–1–16 (2016). (E)

3. “PhysPort Assessments,” <www.physport.org/assessments>, PhysPort is a free website developed by the American Association of Physics Teachers in collaboration with Kansas State University and supported by the National Science Foundation. It was previously called “The PER User’s Guide.” At PhysPort, verified educators can learn about download and research-based assessments in physics and related fields, covering content and non-content topics such as attitudes, beliefs, and scientific reasoning, for various courses from high school to graduate levels. (E)

II. RESEARCH-BASED ASSESSMENTS INSTRUMENTS BEYOND PHYSICS TOPICS

Good research-based assessment instruments are different from typical exams in that their creation involves extensive research and development by experts in education research to ensure that the questions measure the constructs that faculty think are important, that the possible responses represent real student thinking and make sense to students, and that students’ scores reliably reflect their understanding. For an overview of the development process for research-based assessments, see Madsen *et al.*[1] Furthermore, good research-based assessments are used to help instructors understand, in aggregate, how their teaching influenced different aspects of their students’ learning/skills/attitudes, and it follows that the results of the class as a whole are more important than individual students’ scores. Exceptions to this include the rubrics and observation protocols discussed below, which can be used to give individual feedback to students and instructors, as well as the Colorado Assessment of Problem-Solving (discussed in Sec. V), which is meant to look at problem-solving skills in individuals.

The assessments and observation protocols discussed in this Resource Letter are all developed using research-based approaches, but because most of them are not assessments of physics or astronomy topics, there are differences in how they are developed, structured, and used.

In the first Resource Letter in this sequence, RBAI-1,[1] most of the assessments were designed to be offered before and after instruction, allowing faculty to assess their instruction by comparing the gain between pre- and post-test scores. For the RBAIs in this Resource Letter (RBAI-2), there is a much larger variety in the format, administration, and interpretation of results. Measuring students’ skills or beliefs is more difficult than measuring their conceptual knowledge. Developers do not just build multiple-choice questions. Instead, they use a variety of assessment formats including asking students to agree or disagree with statements, rubrics to assess a certain skill, open-ended responses, choosing multiple responses, or not scoring questions at all but instead just discussing answers.

To score belief assessments, students are often compared to a normative group of physics experts. For example, if experts disagree with a statement that physics is about memorizing information, then students who also disagree may receive one point, while students who agree with that statement do not. The overall score is a measure of how much students agree with physicists. To track changes over time, we look at “shifts” in students’ scores. Inherent in the design of these kinds of assessments is a dichotomy between the beliefs of experts and novices. If you are concerned that comparing your students’ beliefs and attitudes to experts may invoke a deficit model of students, instead of comparing your students responses to the expert response and calculating an overall score, you could look only at how your students’ beliefs changed as a result of your course, especially for any categories of questions you are interested in. You also could use the assessment statements as the basis for a discussion with your students about their beliefs and attitudes about physics. For more information about attitudes and beliefs surveys and a meta-analysis of results of commonly used surveys, see Madsen *et al.*[4]

Rubrics and observation protocols are usually scored by identifying response patterns or behaviors that are present (or absent) and assigning points to their presence (or absence). The number of points can vary by which rubric or observation protocol you are using and what it focuses on. This is a substantially different scoring system than asking students to fill in a bubble sheet: the person who is scoring the student work or observing the class makes judgments about the work or behavior they observe. Rubrics help faculty and students understand the students’ strengths and areas for growth for a variety of categories, while observation protocols help faculty understand the activities in classrooms for a variety of settings and activities. Rubrics and observation protocols can be powerful forms of formative assessment for both the students and faculty.

In general, the RBAIs with agree/disagree, multiple-choice, or multiple-response formats are the quickest and easiest to score, as scoring can be automated with a spreadsheet or online tool. Rubrics take much longer to use, as the instructor needs to individually rate each students work using the rubric, though the feedback that the rubric provides to the instructor and students is very rich. Similarly, assessments with an open-ended format are more time consuming to score, as the instructor must individually score each student’s open-ended response. A subset of RBAIs discussed in this Resource Letter are available online, as noted in the “Format” column of Tables III through VIII, and discussed in our expert recommendation, “Administering research-based assessment online” on PhysPort.[5] Usually administration and scoring of an RBAI online is quicker and easier than using a paper-and-pencil version.

It can be helpful to look at details of the research validation for each assessment when deciding which RBAI to use. The research validation details do not tell you about the quality of the test as a whole, *only* about the research validation. To make it easier to compare the research validation between RBAIs, we have created a set of research validation categories that we apply in each RBAI (Tables I and II), as we do on PhysPort.org. The details of which categories apply to a particular RBAI are available on the PhysPort assessment page[3] for that RBAI. We have also developed a “research validation level” to help faculty quickly get a sense of the number of research validation categories a particular RBAI fulfills. RBAIs will have a gold level validation when they have been rigorously developed and recognized by a wider research community. Silver-level RBAIs are well validated but are missing some piece, such as validation by the larger community. Bronze-level assessments are those where developers have done some validation but are missing

Table I. Research validation categories for content and belief RBAIs as well as observation protocols.

Categories for content, belief, and reasoning RBAIs	Categories for observation protocols
Questions based on research into student thinking	Categories based on research into classroom behavior
Studied with student interviews	Studied using iterative observations
Studied with expert review	Tested using inter-rater reliability
Appropriate use of statistical analysis	Training materials are tested
Administered at multiple institutions	Used at multiple institutions
Research published by someone other than developers	Research published by someone other than developers
At least one peer-reviewed publication	At least one peer-reviewed publication

several pieces. Finally, a research-based validation means that an assessment is likely still in the early stages. We have developed separate levels of research validation for observation protocols because the development process for these is substantially different than for the other kinds of assessments. Because faculty, and not students, use protocols, it does not make sense to look at student thinking or do student interviews. Instead, when developing observation protocols, it is vital to ensure that the categories of observation are grounded in real classrooms. The protocol is iteratively developed through use in real classrooms, there is a high level of inter-rater reliability (which means that the observers can interpret and apply the protocol similarly), and the training materials for using the protocol have been tested and refined. To reflect the differences between observation protocols and other types of RBAIs, we developed a parallel set of research validation categories for observation protocols (Table I).

The RBAIs are ordered in Tables III through VIII based on their level of research validation, with gold validated RBAIs listed first.

4. "How physics instruction impacts students' beliefs about learning physics: A meta-analysis of 24 studies," A. Madsen, S. B. McKagan, and E. C. Sayre, Phys. Rev. Spec. Top.-Phys. Educ. Res. **11**, 010115 (2015). (E)

5. "Administering research-based assessments online," S. McKagan and A. Madsen, <https://www.physport.org/recommendations/Entry.cfm?ID=93329>. (E)

III. MATHEMATICS ASSESSMENTS

A. Overview of mathematics assessments

RBAIs for mathematics can be used in physics classes to assess students' level of math readiness for a given physics class, or to assess students' understanding of math topics that are covered in physics classes. These tests are often used in concert with or instead of mathematics placement exams developed locally. We discuss three mathematics assessments, developed by mathematics education researchers, that you can use before instruction to get a sense of students' prerequisite mathematics skills and to assess calculus readiness. You could also use these as pre- and post-tests to see how your students' calculus skills improved because of your course. These are the Precalculus Concept Assessment[6] (PCA), the Calculus Concept Inventory[7,8] (CCI), and the Calculus Concept Readiness Instrument[9] (CCR). There are three additional assessments, developed by physicists that assess mathematics topics often taught in physics classes, i.e., vectors and mathematical modeling. These are the Quadratic and Linear Conceptual Evaluation[10] (QLCE), the Test of Vectors[11] (TUV), and the Vector Evaluation Test[10] (VET). You can use these as a pre- and post-test, to both get a sense of what your students know at the start of your course and what they learned because of your course. Other tests exist (e.g., the Basic Skills Diagnostic Test[12] (BSDT)), but there is no published information available about them, and their developers are unavailable for consultation, and/or we cannot access the assessments.

Table II. Determination of the level of research validation for an assessment, as used on PhysPort.org.

# Categories	Research validation level
All 7	Gold
5–6	Silver
3–4	Bronze
1–2	Research-based

The Precalculus Concept Assessment[6] (PCA) is a multiple-choice pre/post assessment of foundational concepts of beginning calculus, including reasoning abilities around the process view of functions, covariational reasoning and computational abilities, understanding of the meaning of function concepts, growth rate of function types, and function representations. The PCA can be used to help a physics faculty member understand their student's calculus readiness. The PCA questions were developed based on a taxonomy of precalculus concepts (The PCA Taxonomy) using an iterative process of developing questions, testing them with students, interviewing students about their responses, and revising the questions and answer choices.

The Calculus Concept Inventory[7,8] (CCI) is a multiple-choice pre/post assessment of the most basic principles of calculus, where questions are conceptual with no computation on the test. The topics covered on the CCI include functions, derivatives, limits, ratios, and the continuum. The CCI was modeled closely after the Force Concept Inventory[13] (FCI), where the questions look trivial to experts, but students in lecture courses score quite poorly on the test. The CCI questions were first developed by a panel of experts who defined the content to be tested and wrote the questions, and then tested iteratively with students and revised. The CCI is not available for download from PhysPort, because we have not been able to access it ourselves, but individual faculty can access the CCI by emailing the developer.[14]

While the PCA was developed using a research-based taxonomy of concepts, the CCI was designed to mimic the FCI. This difference means that students' responses to CCI questions are more likely to surprise physics faculty ("they should have gotten that!") while PCA questions are more likely to present a robust and varied sense of students' understanding of function concepts in a classroom. The CCI is designed for more advanced math skills than the PCA and may be inappropriate for students enrolled in conceptual or algebra-based physics classes; however, in courses which require substantial calculus or differential equations (e.g., intermediate mechanics), it may be a more appropriate pre-test.

Table III. Mathematics assessments.

Assessment	Content	Intended population	Format	Research validation	Purpose
Calculus Concept Inventory (CCI)	Derivatives, functions, limits, ratios, the continuum	Intro college, high school	Multiple-choice	Gold	Assess student understanding of the most basic principles of calculus.
Pre-calculus Concept Assessment (PCA)	Rate of change, function, process view of functions, covariational reasoning	Intro college	Multiple-choice, available online[5]	Gold	Assess essential knowledge that mathematics education research has revealed to be foundational for students' learning and understanding of the central ideas of beginning calculus.
Calculus Concept Readiness (CCR)	The function concept, trigonometric functions, and exponential functions	Intro college	Multiple-choice	Silver	Assess the effectiveness of pre-calculus level instruction or to be used as a placement test for entry into calculus.
Test of Understanding of Vectors (TUV)	Vectors, components, unit vector, vector addition, subtraction and multiplication, dot and cross product	Intro college	Multiple-choice	Silver	Assess students' understanding of vector concepts in problems without a physical context.
Quadratic and Linear Conceptual Evaluation (QLCE)	Graphing, mathematical modeling	Intro college, high school	Multiple-choice	Bronze	Measure student understanding of equations (linear and quadratic) as functional relationships. Also, to measure students' mathematical knowledge in both traditional and reform courses.
Vector Evaluation Test (VET)	Vector addition and subtraction, component analysis, and comparing magnitudes	Intro colleges high school	Multiple-choice, multiple-response, open-ended	Bronze	Measure students' conceptual understanding of vectors.

The Calculus Concept Readiness[9] (CCR) instrument is a multiple-choice pre/post assessment of foundational concepts for introductory calculus, including the function concept, trigonometric functions, and exponential functions. The CCR was developed to assess students' readiness for calculus courses or to assess the effectiveness of pre-calculus courses. Like the PCA, the CCR was developed using a research-based taxonomy of concepts. The CCR is owned by the Mathematical Association of America and is available for a fee through Maplesoft.[15]

At first blush, the PCA, CCR, and CCI cover very similar topics at a very similar level. However, their emphasis is different, and care should be taken to match the test with your students. The CCR surveys students' understanding of a broad base of mathematics concepts from pre-calculus, including both functions and trigonometry, while the PCA focuses only on the mathematics needed to move into calculus (primarily functions) before calculus instruction. The CCI is designed to test the core concepts of calculus and is aimed at students before and after calculus instruction. Both the PCA and the CCR were developed by the same team of researchers using very similar development methods, and the tests have very a similar structure and feel. The CCI was independently developed by a different team using less robust research methods. If you use these as part of a mathematics placement package or to measure their students' mathematics skills, the CCR is recommended because of the trigonometry and solving equations cluster, though you must pay to use it. Physicists are typically not as interested as mathematicians are in the intricacies of how students understand "function" as a concept, devoid of the physical context, so the PCA and CCI may not be as helpful as the CCR for these purposes.

The Quadratic and Linear Conceptual Evaluation[10] (QLCE) is a multiple-choice assessment about relating kinematics to quadratic graphs and equations, relating coefficient changes in linear equations to linear graph changes and vice versa. Some questions have a kinematics context, and some questions have a generic context. The developers created the QLCE because they had heard faculty say that their students "understood the math, but not the concepts," and wanted to see if their physics students did indeed understand these mathematical concepts. There are several sets of questions where students fill in a matrix to answer, so you would need to renumber them for use with Scantron and will need a special Scantron sheet that can take up to 10 answers and multiple responses for each question. These questions were developed based on research into student ideas about quadratic and linear equations and the developers' experience with the concepts with their students.

The Test of Understanding of Vectors[11] (TUV) is a multiple-choice test that assesses introductory physics students' understanding of vector concepts without any physical context. Concepts tested include unit-vector notation, graphical representation of vectors and components, calculation of vector components, vector addition, subtraction and scalar multiplication, and dot and cross product. The TUV questions were developed from students' open-ended responses to questions about vectors, so the multiple-choice answers strongly reflect students' ideas about vectors (both correct and incorrect). The TUV was developed in Mexico in Spanish and then translated into English.

The Vector Evaluation Test[10] (VET) is a multiple-choice, multiple-response (can pick more than one option), and open-ended assessment of vector concepts for introductory physics classes. About a quarter of the questions are asked in a physics context, and the rest are given no physical context. The VET questions were based on the developers' experience with students thinking about vectors.

Both the TUV and VET cover vector decomposition, addition, subtraction, dot products, and cross products, which are the major issues for using vectors in introductory physics. Additionally, the TUV uses both graphical representations and vector-hat representations, so it is possible to compare students' performance across representations. The VET covers coordinate rotation and time changes of kinematics vectors, so it is more appropriate to use this test if you would like to test more topics instead of more representations. Though it is a more thorough test of the topics it does cover, the TUV's reliance on few questions per topic means that scores are still sensitive to the peculiarities of the questions on the test.

6. "The Precalculus Concept Assessment: A tool for assessing students' reasoning abilities and understandings," M. P. Carlson, M. Oehrtman, and N. Engelke, Cogn. Instr. **28**(2), 113–145 (2010). (I)

7. "Development and validation of the Calculus Concept Inventory," J. Epstein, in Proceedings of the Ninth International Conference on Mathematics Education in a Global Community, D. Pugalee, A. Rogerson, and A. Schinck, Eds., pp. 165–170, Springer, New York (2007). (E)

8. "The Calculus Concept Inventory–Measurement of the effect of teaching methodology in mathematics," J. Epstein, Not. Am. Math. Soc. **60**(08), 1018 (2013). (E)

9. "A study of students' readiness to learn calculus," M. P. Carlson, B. Madison, and R. D. West, Int. J. Res. Undergrad. Math. Educ. **1**(2), 209–233 (2015). (I)

10. "Measuring and improving student mathematical skills for modeling," R. K. Thornton, in Proceedings of the GIREP Conference: Modeling in Physics and Physics Education, E. van den Berg, A. L. Ellermeijer, and O. Slooten, Eds., Amsterdam, The Netherlands (2006). (E)

11. "Test of understanding of vectors: A reliable multiple-choice vector concept test," P. Barniol and G. Zavala, Phys. Rev. Spec. Top.-Phys. Educ. Res. **10**(1), 010121 (2014). (I)

12. "Cognitive development in an integrated mathematics and science program," J. Epstein, J. Coll. Sci. Teach. **27**(3), 194–201 (1997). (E)

13. "Force Concept Inventory," D. Hestenes, M. M. Wells, and G. Swackhamer, Phys. Teach. **30**(3), 141–166 (1992). (E)

14. Email jerepst@att.net for a copy of the CCI. The developer of the CCI has passed away, but his assistant is still responding to inquiries at this email address.

15. "Concept Readiness Tests," <https://www.maplesoft.com/products/placement/ccr_test.aspx>. (E)

B. Recommendations for choosing a mathematics assessment

You can use these mathematics assessments before instruction to get a sense of what your students already know, or after instruction if you are implementing new teaching practices to increase student understanding of a given topic and want to assess the effectiveness. Because the QLCE, PCA, CCR, and CCI test overlapping concepts, you should select one of these four that best matches your population and assessment needs. Do not mix-and-match these tests for pre- and post-test because you will have difficulty comparing pre-scores to post-scores. If you are using a test only before instruction to see if your students are ready to take your course or to adjust your teaching to best fit their incoming skills, select a test of more elementary content that might be fully covered in prerequisite classes. If you are using a test before and after instruction, you might select a test that includes some content covered in corequisite courses.

IV. BELIEFS AND ATTITUDES

A. Overview of belief and attitude assessments

There are 14 research-based assessments of students' beliefs and attitudes that we discuss here. We discuss belief and attitude assessment from four categories: students' beliefs about learning physics in general, students' beliefs about specific aspects of physics or their own learning (e.g., labs and problem solving), students' self-efficacy in their physics class, and students' views about the nature of science. There are also additional assessments of motivation, discussed in Lovelace and Brickman[16] that may be of interest, but will not be discussed here.

Since these surveys of beliefs and attitudes do not assess the content covered in any course, they can be used at the high school level and at all levels in the undergraduate and graduate curriculum (unless otherwise noted below). Many of these surveys can be used across disciplines or have versions specifically tailored to other disciplines. Most of these beliefs and attitudes surveys (unless otherwise noted) are meant to be given as a pre-test at the beginning of the semester and post-test at the end of the semester. In order to look at the shifts in belief scores during your course, they are also appropriate to give at other times in the semester (e.g., near exams) or across an entire course sequence.

Belief surveys are carefully designed to measure what students believe about a topic rather than simply whether they like that topic. However, they have several important limitations. First, they can only measure self-reported explicit beliefs, not students' implicit beliefs. For example, a student might say and really believe "When I am solving a physics problem, I try to decide what would be a reasonable value for the answer" but not do that in real life. Second, it may be difficult to distinguish in students' answers whether they are thinking about the structure of the course they are enrolled in or in the practice of learning physics more broadly. Finally, many belief surveys assume a context of a typical physics course that includes elements such as solving problems, reading the textbook, and taking exams and thus may not be appropriate in a very nontraditional physics course or in a context outside of a physics course.

16. "Best practices for measuring students' attitudes toward learning science," M. Lovelace and P. Brickman, CBE Life Sci. Educ. **12**(4), 606–617 (2013). (E)

1. Beliefs about physics learning in general

Many physics faculty care about their students learning to think like physicists but often do not assess this because it is not clear how to do so best. Physics education researchers have created several surveys to assess one important aspect of thinking like a physicist: what students believe that learning physics is all about. These surveys are not about whether students like physics, but about how students perceive the discipline of physics or their physics course. These surveys measure students' self-reported beliefs about physics and their physics courses and how closely these beliefs about physics align with experts' beliefs. There are four assessments about students' beliefs about learning physics in general: The Colorado Learning Attitudes about Science Survey[17] (CLASS), Maryland Physics Expectations Survey[18] (MPEX), Epistemological Beliefs Assessment for Physical Sciences[19,20] (EBAPS), and the Views About Science Survey[21,22] (VASS).

The Colorado Learning About Science Survey[17] (CLASS—pronounced "sea-lass") asks students to agree/disagree with statements about their beliefs about physics and learning physics around such topics as real-world connections, personal interest, sense-making/effort, and problem solving. Students are asked to strongly agree, agree, neutral, disagree, or strongly disagree (5-point Likert scale) with a question statement (Fig. 1). The survey is most commonly scored by collapsing students' responses into two categories ("strongly agree" and "agree" are grouped, "strongly disagree" and "disagree" are grouped) depending on whether they are the same as an expert physicist would give. For an explanation of the reasons for collapsing student responses into two categories, see the "scoring" section of Adams *et al.*[17] An individual student's "percent favorable" score is the average number of questions that they answered in the same way as an expert. It is most common for faculty to look at the shift in their class average percent favorable scores from pre-test to post-test to understand how their course influences students' attitudes and beliefs about physics, on average. One would hope that after a physics course, students' beliefs would become more expert-like, so the class average percent favorable scores would increase from pre- to post-test. The CLASS questions contain only one statement that students can agree or disagree with to help students interpret these questions consistently (as opposed to more than one idea in the same question). The CLASS questions were developed based on questions from the MPEX and VASS. The CLASS added questions about personal interest, aspects of problem solving and the coupled beliefs of sense-making and effort that were not included in the MPEX or VASS.[17]

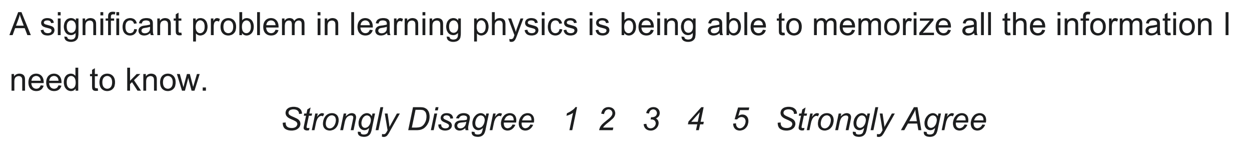
The Maryland Physics Expectations Survey[19] (MPEX) measures students' self-reported beliefs about physics and their physics courses, their expectations about learning physics and how closely these beliefs about physics align with experts' beliefs. The surveys ask students to rank 5-point Likert scale questions about how they learn physics, how physics is related to their everyday lives, and about their physics course. Some of the MPEX questions are very course specific, e.g., they ask about a student's grade in the course. The format and scoring of the MPEX questions are the same as the CLASS questions. The questions on the MPEX were chosen through literature review, discussion with faculty, and the researchers' personal experiences.

The CLASS and MPEX are very similar and several items are the same on both tests. The MPEX and CLASS both ask questions about students' personal beliefs about learning physics, but the MPEX focuses more on students' expectations for what their specific physics course will be like. While the CLASS does not include questions about expectations for the specific course, it does include questions that only make sense in the context of a physics course, e.g., asking about students' belief that they can solve a physics problem after studying that physics topic. The MPEX takes longer to complete than the CLASS, even though it has fewer questions (34 versus 42) presumably because some of the MPEX questions take longer for students to understand and answer because they contain multiple ideas. Both assessments have a strong research validation. The CLASS builds on the MPEX, and has been used more widely, so there is more comparison data available.[4]

The Epistemological Beliefs About Physics Survey (EBAPS) probes students' epistemology of physics, or their view of what it means to learn and understand physics.[19] The EBAPS also contains questions that are course specific (as opposed to being about learning physics in general), for example, one question asks about how students should study in their physics class. The developers tried to ensure that the EBAPS questions do not have an obvious sanctioned answer and have a rich context in order to elicit students' views more successfully.[20] The EBAPS has three question types. Part one contains agree/disagree Likert scale questions, part 2 contains multiple-choice questions, and part 3 gives students two statements and asks them to indicate how much they agree with each (similar to the VASS). The level of sophistication of students' answers is scored using a non-linear scoring scheme where different responses have different weighting depending on how sophisticated the developers determined each answer was. The EBAPS is most appropriate for high school and college level introductory physics courses. The EBAPS questions were developed based on an extensive review of the MPEX and Schommer's Epistemological Questionnaire.[23] The developers synthesized other researchers' ideas to create guiding principles, which they used to write the EBAPS questions.

The main difference between the EBAPS and the CLASS and MPEX is the style of the questions, where the EBAPS has three styles of questions, and the MPEX and CLASS include only agree/disagree questions. The content on the EBAPS, MPEX, and CLASS is similar and all have high levels of research validation.

The Views About Science Survey[21,22] (VASS) is another survey for probing student beliefs about physics and learning physics. The VASS uses a special question format called



A significant problem in learning physics is being able to memorize all the information I need to know.

Strongly Disagree 1 2 3 4 5 Strongly Agree

Fig. 1. Example of a 5-point Likert-scale question on the CLASS.

Table IV. Belief and attitude assessments.

Title	Focus	Intended population	Format	Research validation	Purpose
Beliefs About Physics Learning in General					
Colorado Learning Attitudes about Science Survey (CLASS)	Self-reported beliefs about physics and learning physics	Upper-level, intermediate, intro college, high school	Agree/disagree, available online[5]	Gold	Measure students' beliefs about physics and learning physics and distinguish the beliefs of experts from those of novices.
Maryland Physics Expectations Survey (MPEX)	Beliefs about one's physics course	Upper-level, intermediate, intro college, high school	Agree/disagree	Gold	Probe some aspects of student expectations in physics courses and measure the distribution of student views at the beginning and end of the course.
Epistemological Beliefs Assessment for Physical Sciences (EBAPS)	Epistemological beliefs, structure of knowledge, nature of knowing and learning, real-life applicability, evolving knowledge, source of ability to learn	Intro college, high school	Agree/disagree, multiple-choice, contrasting alternatives	Silver	Probe the epistemological stances of students in introductory physics, chemistry and physical science.
Views About Science Survey (VASS)	Structure and validity of scientific knowledge, scientific methodology, learnability of science, reflective thinking, personal relevance of science	Intro college, high school	Contrasting alternatives	Silver	Characterize student views about knowing and learning science and assess the relation of these views to achievement in science courses.
Beliefs About Physics Learning in a Specific Context					
Colorado Learning Attitudes about Science Survey for Experimental Physics(E-CLASS)	Affect, confidence, math-physics-data connection, physics community, uncertainty, troubleshooting, argumentation, experimental design, modeling	Upper-level, intermediate, intro college	Agree/disagree, available online[5]	Gold	Measure students' epistemologies and expectations around experimental physics.
Attitudes and Approaches to Problem Solving Survey (AAPS)	Attitudes about problem-solving	Graduate, upper-level, intermediate, intro college	Agree/disagree	Silver	Measure students' attitudes and approaches to problem solving at the introductory and graduate level.
Physics Goals Orientation Survey (PGOS)	Goal orientation and motivation in physics	Intro college	Agree/disagree	Silver	Assess students' motivation and goal orientations in university-level physics courses.
Student Assessment of Learning Gains (SALG)	Self-assessment of learning	Intro college	Agree/disagree	Silver	Understand students' self-assessment of their learning from different aspects of the course and their gains in skills, attitudes, understanding of concepts, and integrating information.
Attitudes about Problem Solving Survey (APSS)	Attitudes about problem-solving	Intro college	Agree/disagree	Bronze	Survey students' attitudes towards and views of problem solving.
Nature of Science					
Views of the Nature of Science (VNOS)	Nature of science, theories and laws, tentativeness, creativity, objectivity, subjectivity, social and cultural influences	High school, intro college	Agree/disagree	Silver	Elucidate students' views about several aspects of the nature of science.

Table IV. *Continued*

Title	Focus	Intended population	Format	Research validation	Purpose
Views on Science and Education (VOSE)	Nature of science, theories and laws, tentativeness, creativity, objectivity, subjectivity, scientific method, teaching the nature of science	High school, intro college, intermediate, upper level	Open-ended	Silver	Create in-depth profiles of the views of students or adults about the nature of science and nature of science instruction.
Self-Efficacy					
Sources of Self-Efficacy in Science Courses-Physics (SOSESC-P)	Self-efficacy	Intro college	Agree/disagree	Bronze	Assess students' beliefs that they can succeed in their physics course.
Physics Self-Efficacy Questionnaire (PSEQ)	Self-efficacy	Intro college	Agree/disagree	Bronze	Measure students' self-efficacy in their physics course.
Self-Efficacy in Physics Instrument (SEP)	Self-efficacy	Intro college	Agree/disagree	Bronze	Examine the relationship between physics self-efficacy and student performance in introductory physics classrooms.

contrasting alternative design where students compare and contrast between two viewpoints. For example, one question contains the statement "Learning in this course requires:" with the contrasting alternatives "(a) a special talent" and "(b) a serious effort." Students are asked to compare how much they agree with (a) and (b) by choosing between the following options: (a) $\gg$ (b), (a) $>$ (b), (a) $=$ (b), (a) $<$ (b), or (a) $\ll$ (b). Questions are scored in the same way as the MPEX and CLASS. The VASS can be used in introductory college physics courses and high school physics courses. VASS questions were developed based on an expert/folk taxonomy of student views about science.

The biggest difference between the VASS and the CLASS and MPEX is that the VASS uses the contrasting cases format. The VASS format can be confusing to students if they do not agree that the answer choices given represent opposites. The VASS may be less reliable for measuring expert-like beliefs but still very useful for discussing students' ideas about learning physics. The CLASS and MPEX have more obvious expert-like answers, so their results can give you a better idea of how expert-like your students' views are. Although it may seem that if there is an obvious expert-like answer, students would choose this over reporting their own personal beliefs, *Gray* et al.[24] found evidence that for the CLASS, students answer based on their own personal beliefs. The content of the VASS is very similar to the CLASS and MPEX. Like the MPEX, the VASS has several questions that are course specific.

17. "New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey," W. Adams, K. Perkins, N. Podolefsky, M. Dubson, N. Finkelstein, and C. Wieman, Phys. Rev. Spec. Top.-Phys. Educ. Res. **2**, 010101 (2006). (I)

18. "On the effectiveness of interactive engagement microcomputer based laboratories," E. F. Redish, J. M. Saul, and R. N. Steinberg, Am. J. Phys. **65**(1). (E)

19. "Helping physics students learn how to learn," A. Elby, Am. J. Phys. **69**(7) (2001). (E)

20. "The Idea Behind EBAPS," A. Elby, <http://www2.physics.umd.edu/~elby/EBAPS/idea.htm>. (E)

21. "Views about science and physics achievement: The VASS story," I. Halloun, AIP Conf. Proc. **399**, 605–614 (1997). (E)

22. "Interpreting VASS dimensions and profiles for physics students," I. Halloun and D. Hestenes, Sci. Educ. **7**(6), 553–577 (1998). (E)

23. "The effects of beliefs about the nature of knowledge in comprehension," M. Schommer, J. Educ. Psychol. **82**(3), 498–504 (1990). (I)

24. "Students know what physicists believe, but they don't agree: A study using the CLASS survey," K. E. Gray, W. K. Adams, C. E. Wieman, and K. K. Perkins, Phys. Rev. Spec. Top.-Phys. Educ. Res. **4**, 020106 (2008). (E)

2. *Beliefs about physics learning in a specific context*

There are five assessments about students' beliefs about specific aspects of physics or their own learning, e.g., laboratories and problem solving. These are the Colorado Learning about Science Survey for Experimental Physics[25] (E-CLASS), the Attitudes and Approaches to Problem Solving[26,27] (AAPS), the Attitudes about Problem Solving Survey[28] (APSS), the Physics Goal Orientation Survey[29] (PGOS), and the Self-Assessment of Learning Gains[30,31] (SALG). These surveys have been created for three specific contexts: experimental physics (E-CLASS[25]), problem solving (AAPS[26] and APSS[28]), and goal orientations (PGOS[29]). There are three additional RBAIs that deal with problem solving more generally, and not attitudes about problem solving, discussed below in Sec. V.

The Colorado Learning Attitudes about Science Survey for Experimental Physics[25] (E-CLASS) is designed to measure the influence of a laboratory course on students' epistemologies and expectations around experimental physics. The E-CLASS asks about a wide range of epistemological beliefs, so that it can be used in courses with a wide range of goals. The E-CLASS asks students to rate their agreement with statements by answering for themselves, "What do YOU think when doing experiments for class?" and answering for a physicist, "What would experimental physicists say about their research?" This helps instructors differentiate students' personal and professional epistemologies. The E-CLASS can be used in introductory, intermediate, or upper-level laboratory courses and is administered online through the developer website.[32] The E-CLASS score is calculated using the responses to the questions about students' personal beliefs (not the prompts about what they think a physicist's response is). The E-CLASS score is calculated by giving +1 point for an expert-like response (favorable), 0 points for a neutral response and –1 points for a novice-like response (unfavorable). The total score for the 30 questions can range from –30 to 30 points. The percentage of students who give the expert-like response, and how this changes from pre- to post-test, determines how the course influenced students' beliefs about experimental physics. The E-CLASS questions were developed based on consensus learning goals defined by faculty at the University of Colorado at Boulder for their laboratory curriculum. The questions were modeled after questions on the CLASS and based on common challenges instructors observed students having in laboratory courses.

Two surveys measure students' attitudes and approaches to problem solving in physics. These surveys are important because the way students think about problem solving can affect how they learn this skill, and faculty can target the development of problem-solving skills to help their students improve.

The Attitudes toward Problem Solving Survey[28] (APSS) is a survey of students' attitudes toward problem solving, e.g., how they think about equations, the process they go through to solve problems and their views on what problem solving in physics means. Like other attitude and belief surveys, students are asked to agree with statements using a 5-point Likert scale, strongly (dis)agree and (dis)agree are collapsed, and the percent expert response is calculated as the percentage of questions where students agree with the expert response. In addition to the agree/disagree questions, there are also two multiple-choice questions on the APSS. The APSS is appropriate for introductory college courses. Some of the APSS questions were adopted from the MPEX, while others were newly created.

Like the APSS, the Attitudes and Approaches to Problem Solving[26,27] (AAPS) measures students' agreement with statements about their attitudes and approaches to problem-solving using a 5-point Likert scale. To calculate the average score for a question, +1 is assigned to each favorable response, –1 is assigned to each unfavorable response, 0 is assigned to neutral response, and the overall score is the average of the score for each question. The AAPS can be used at all levels of undergraduate courses and at the graduate level.

Since the AAPS was developed by expanding the APSS, the topics covered and the questions on the AAPS and APSS are quite similar. Fourteen of the questions are the same or very similar between the tests. The AAPS has more questions (33 questions versus 20 questions), so it covers a few more aspects of problem solving than the APSS, including how students feel about problem solving, how they learn from the problem-solving process, use of pictures/diagrams, and what students do while solving a problem. The AAPS also includes questions that target graduate-level problem solving.

The CLASS, MPEX, EBAPS, and VASS also contain questions about students' attitudes and beliefs about problem solving, similar to those on the APSS and AAPS. The AAPS and APSS can specifically target problem-solving beliefs, while the CLASS, MPEX, EBAPS, and VASS ask about a wider range of beliefs and attitudes.

The Physics Goals Orientation Survey[29] (PGOS) is a survey of students' motivations and goal orientations in their physics course. These motivations can influence how students engage in their physics class and how well they learn the material. The PGOS addresses four goal orientations: task orientation (the belief that success is a product of effort, understanding, and collaboration), ego orientation (the belief that success relies on greater ability and attempting to outperform others), cooperation (when students value interaction with their peers in the learning process), and work avoidance (the goal of minimum effort–maximum gain). The PGOS uses a 5-point Likert scale, with 1 point given for strongly disagree, 5 points for strongly agree, and 2–4 points for disagree, neutral, or agree, respectively. The average score for each of the four goal orientations is calculated separately, and there is no overall score calculated. The PGOS is appropriate for introductory and intermediate university physics courses. It can be given as a pre- and post-test to determine how your course may have influenced students' goal orientations. The PGOS questions were taken from a previous survey of goal orientation by Duda and Nicholls[33] and revised so that they would be appropriate for a university-level physics course, with some new questions created. The PGOS was developed in Australia.

The Student Assessment of Learning Gains[30] (SALG) is an online assessment where students self-assess how different parts of their course impacted their learning using a 5-point Likert scale. It is like the student evaluation given at the end of most courses, but the questions only ask students about what they gained from different aspects of the course instead of what they liked. The SALG developers found that students' observations about what they gained from the class were useful to help faculty improve the course, whereas their observations about what they liked were not helpful.[30] You can use the SALG online system[31] to choose questions to include from each of the following categories: understanding of the class content, increase in skills, class impact on attitudes, integration of learning, the class overall, class activities, assignments, graded activities and tests, class resources, the information you were given, and support for you as an individual learner. You can also edit and reorder questions. You can give the SALG at a midpoint in your class to get a sense of which parts of your course could be improved, or at the end to evaluate your students' understanding of how your course supported their learning. The SALG website[31] also has a "baseline instrument" available that can be used at the beginning of a course. The SALG was developed using data from more than 300 student interviews where students discussed what they had gained from certain aspects of a course, and what they liked.

25. "Epistemology and expectations survey about experimental physics: Development and initial results," B. M. Zwickl, T. Hirokawa, N. Finkelstein, and H. J. Lewandowski, Phys. Rev. Spec. Top.-Phys. Educ. Res. **10**, 010120 (2014). (E)

26. "Surveying graduate students' attitudes and approaches to problem solving," A. Mason and C. Singh, Phys. Rev. Spec. Top.-Phys. Educ. Res. **6**, 020124 (2010). (E)

27. "Physics graduate students' attitudes and approaches to problem solving," C. Singh and A. Mason, AIP Conf. Proc. **1179**, 273–276 (2009). (E)

28. "Attitudes toward problem solving as predictors of student success," K. Cummings, S. Lockwood, S. Connecticut, N. Haven, and J. D. Marx, AIP Conf. Proc. **720**(1), 133–136 (2003). (E)

29. "Development of a physics goal orientation survey," C. Lindstrom and M. D. Sharma, Int. J. Innov. Sci. Math. Educ. **18**(2), 10–20 (2010). (E)

30. "Creating a better mousetrap: On-line student assessment of their learning gains," E. Seymour, D. J. Wiese, A. B. Hunter, and S. Daffinrud, in National Meeting of the American Chemical Society, pp. 1–40, San Francisco (2000). (E)

31. "Student assessment of their learning gains," <https://www.salgsite.org>. (E)

32. "E-CLASS: Colorado Learning Attitudes About Science Survey for Experimental Physics," <tinyurl.com/ECLASS-physics>.

33. "Dimensions of achievement-motivation in schoolwork and sport," J. L. Duda and J. G. Nicholls, J. Educ. Psychol. **84**(3), 290–299 (1992). (I)

3. Nature of science

There are two main research-based surveys about the nature of science, the Views on Science and Education Questionnaire[34] (VOSE) and the Views about the Nature of Science Questionnaire[35] (VNOS), which probe students' views about the values and epistemological assumptions of science. These surveys can help faculty understand how their courses and teaching methods influence students' views of the nature of science. These can be especially useful in courses that aim to develop these views, such as courses for pre-service teachers. Both are intended as both a pre- and post-test.

The VOSE[34] is a Likert-scale survey of students' beliefs about the nature of science and beliefs about how you should teach the nature of science. The VOSE addresses seven major topics including tentativeness of scientific knowledge, nature of observation, scientific methods, hypotheses, laws and theories, imagination, validation of scientific knowledge, and objectivity and subjectivity in science. It also includes five questions about students' beliefs about teaching the nature of science. Each question consists of a question statement and 3–9 possible responses, with which students can agree or disagree with using a 5-point Likert scale. There are no right or wrong answers, but each statement corresponds to a particular "position" on one or more subtopics of nature of science. The developer has created an extensive list of coding categories to "create an in-depth profile of a [student's] nature of science views and educational ideas."[34] The coding categories can be found in Chen.[34] Burton[36] developed a numerical system for calculating a numerical score for each issue or topic, by assigning a number between 0 and 4 to a student's response for each item listed under that issue or topic and calculating the average. The VOSE can be used in high school courses and in introductory, intermediate, and upper-level undergraduate courses. The VOSE questions were developed based on questions from the Views on Science-Technology-Society[37] (VOSTS) and VNOS[35] to address concerns about the VOSTS and VNOS being open-ended and hard to administer and score. The VOSE aims to increase the validity of the survey and decrease interpretation biases, as compared to the VOSTS and VNOS.

The Views on the Nature of Science Questionnaire[35] (VNOS) is an open-ended survey of students' ideas about the nature of science, including the empirical, tentative, inferential, creative, theory-laden nature of science, and the social and cultural influences on scientific knowledge. Many of the questions ask students to give an example to support their ideas. In addition to students written responses, the developers encourage faculty to do individual follow-up interviews with students to better understand the meanings of their responses to the questions. Students' responses can be scored as naïve, transitional, or informed based on a rubric for each question. The VNOS can be used with middle school, high school, and introductory college students. The VNOS questions were created by the developers and tested with students and experts.

The VNOS and VOSE cover similar topics around the nature of science. The main difference between them is the format. The VNOS is open-ended while the VOSE asks students to agree/disagree with different options. Because the VNOS is open-ended, it can be time consuming to score and subject to interpretation bias, though conducting interviews with students about their responses reduces the chance of bias in scoring. Another difference between the VOSE and VNOS is that in addition to asking about students' philosophical beliefs about science, the VOSE asks students to agree/disagree with statements about how to teach the nature of science.

Many other multiple-choice instruments to assess students' views of the nature of science were developed in the 1960s, 1970s, and 1980s but were based on researchers' ideas and not on student interviews or research into student thinking.[38] The VOSTS,[37] published in 1992, was the first nature of science instrument to use a student-centered design process, including analysis of student responses and student interviews. However, other researchers found many problems with students' interpretations of the VOSTS.[34,35,39] Both the VOSE and the VNOS were developed in response to these problems.

Surveys about the nature of science, such as the VNOS, have been criticized for measuring only what students say declaratively about the nature of science, which may be quite different from what they do procedurally when engaged in authentic scientific practice.[40] It is worth recognizing that this is an inherent limitation of such surveys.

34. "Development of an instrument to assess views on nature of science and attitudes toward teaching science," S. Chen, Sci. Educ. **90**(5), 803–819 (2006). (I)

35. "Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science," N. G. Lederman, F. Abd-El-Khalick, R. L. Bell, and R. S. Schwartz, J. Res. Sci. Teach. **39**(6), 497–521 (2002). (E)

36. "Student work products as a teaching tool for nature of science pedagogical knowledge: A professional development project with in-service secondary science teachers," E. P. Burton, Teach. Teach. Educ. **29**(1), 156–166 (2013). (I)

37. "The Development of a new instrument: 'Views on Science-Technology-Society' (VOSTS)," G. S. Aikenhead and A. G. Ryan, Sci. Educ. **76**(5), 477–491 (1992). (E)

38. "The influence of history of science courses on students' views of nature of science," F. Abd-El-Khalick and N. G. Lederman, J. Res. Sci. Teach. **37**(10), 1057–1095 (2000). (E)

39. "The nature of science and instructional practice: Making the unnatural natural," F. Abd-El-Khalick, R. L. Bell, and N. G. Lederman, Sci. Educ. **82**(4), 417–436 (1998). (E)

40. "What students say versus what they do regarding scientific inquiry," I. Y. Salter and L. J. Atkins, Sci. Educ. **98**(1) (2014). (E)

4. Self-efficacy

Self-efficacy is a person's situation-specific belief that they can succeed in a given domain.[41] There are three assessments of students' views of their self-efficacy in their physics classes: Sources of Self-Efficacy in Science Courses-Physics[42] (SOSESC-P), Physics Self-Efficacy Questionnaire[43] (PSEQ), and the Self-efficacy in Physics Instrument[44] (SEP). There are numerous other assessments of self-efficacy with differing focuses, e.g., other disciplines and self-efficacy in general. We focus on those specifically developed for physics courses. All three of these assessments ask students to rate their agreement with statements on a five-point Likert scale and are appropriate for introductory college students.

The Sources of Self-Efficacy in Science Courses-Physics[42] (SOSESC-P) assesses students' beliefs that they can succeed in their physics course by asking them to agree or disagree with a series of statements. The questions are divided into four categories, corresponding to four established aspects of self-efficacy: performance accomplishment, social persuasion, vicarious learning, and emotional arousal. These questions ask about students' feelings about different aspects of the course, how the instructor and other students influenced their views of themselves, the students' behavior in the course (paying attention, working hard, etc.), and more. Several of the Likert-scale questions on the SOSESC-P were taken from existing mathematics and general academic surveys of self-efficacy. Additional new questions were written based on the developers' experience with undergraduate science education.

The Physics Self-Efficacy Questionnaire[43] (PSEQ) is a similar survey of students' beliefs that they can succeed in their physics course. The PSEQ has five questions, so it probes only one dimension of self-efficacy. Specifically, the PSEQ focuses on students' confidence in their ability to succeed in their physics course. The questions do not mention specific portions of the course or specific members of the course (other students, instructor, etc.). They simply ask the students about themselves and their own ability in their physics course. Most of the Likert-scale questions on the PSEQ are modified versions of questions from the General Self-Efficacy Scale,[45] while one PSEQ question was written by the developers. The PSEQ was developed in Australia.

The Self-Efficacy in Physics[44] (SEP) instrument is another survey that asks students to agree with statements about their beliefs about their ability to succeed in their physics course. The SEP contains eight questions, which are more specific than those on the PSEQ. These questions ask students how good or bad they are at science/mathematics, if they are good at using computers, and if they believe they can solve two specific mechanics problems. The SEP questions were developed based on a literature review and modeled after self-efficacy questions from surveys in other disciplines.

The SOSESC-P has 33 questions, whereas the PSEQ and SEP have 5 and 8 questions, respectively, so the SOSESC-P probes more dimensions of self-efficacy in more depth than the other surveys. There is a lot more variety in the questions on the SEP than the questions on the PSEQ. The SEP asks students about their belief that they can solve very specific physics problems, their comfort using a computer, and if they consider themselves good at mathematics, whereas the PSEQ questions are about physics in general. All have the same level of research validation.

41. "Self-efficacy: Toward a unifying theory of behavioral change," A. Bandura, in Adv. Behav. Res. Therapy **1**(4), pp. 139–161 (1978). (I)

42. "Engaging students: An examination of the effects of teaching strategies on self-efficacy and course climate in a nonmajors physics course," H. Fencl and K. Scheel, J. Coll. Sci. Teach. **35**(1), 20–25 (2005). (E)

43. "Self-efficacy of first year university physics students: Do gender and prior formal instruction in physics matter?," C. Lindstrøm and M. D. Sharma, Int. J. Innov. Sci. Math. Educ. **19**(2), 1–19 (2011). (E)

44. "The Development of a Physics Self-Efficacy Instrument for use in the introductory classroom," K. A. Shaw, AIP Conf. Proc. **720**(1), 137–140 (2004). (E)

45. Measurement of perceived self-efficacy: Psychometric scales for cross-cultural research, R. Schwarzer, Freie Universität Berlin, Berlin (1993). (I)

B. Recommendations for choosing a belief and attitude assessment

1. General beliefs

Use the CLASS if you want an assessment that is quick to complete, has a large amount of comparison data available, and where the questions are easy for students to understand. Furthermore, use the CLASS if you want to look at categories of questions that were determined through a rigorous statistical analysis, so they reflect students' views of the relationship between questions. The CLASS and MPEX statements refer to the kinds of activities that students do in a traditional introductory physics course, so the questions may not make sense to students if you are teaching in a very nontraditional way. If you have been using the MPEX, EBAPS, or CLASS in the past, you may want to keep using these to compare your results. The MPEX was designed with a

Table V. Problem-solving assessments.

Assessment	Focus	Intended population	Format	Research validation	Purpose
Minnesota Assessment of Problem Solving rubric (MAPS)	Rubric to score written problem solutions	High school, intro college	Rubric	Silver	Assess written problem solutions on five different aspects of problem solving in undergraduate introductory physics courses.
Colorado Assessment of Problem Solving (CAPS)	Detailed understanding of students' problem solving	Graduate, upper-level, intermediate, intro college, high school, middle school	Open-ended	Bronze	Assess students' strengths and weaknesses on 44 different components of the problem-solving process, using a general problem-solving situation that is not tied to any specific discipline.
Assessment of Textbook Problem Solving Ability (ATPSA)	Solving textbook problems	Intro college	Open-ended	Bronze	Gauge students' problem-solving ability in a first-semester calculus-based physics course.

resources perspective, which assumes that students' ideas are not coherent, so if you want an assessment from the resources perspective, use the MPEX.

2. *Specific beliefs*

Use the E-CLASS if you want to measure students' beliefs in the context of experimental physics. Use the APSS if you want to probe your students' attitudes about problem solving, including undergraduate and graduate students. Use the PGOS if you want to understand your students' motivations and goal orientations in their physics course. Use the SALG if you want to understand your students' perspective on which parts of your course helped them learn the most.

3. *Nature of science*

Use the VOSE if you want a multiple-choice assessment that is quick and easy to score. Use the VNOS if you would like to use an open-ended survey to get a more detailed understanding of your students' views on the nature of science.

4. *Self-efficacy*

If you want to measure detailed changes in your students' physics course specific self-efficacy, use the SOSESC-P, as it probes several dimensions of self-efficacy and uses several questions to probe each. If you need a shorter self-efficacy assessment that can be combined with some other assessment, use the five-question PSEQ, which can give you a general sense of your students' belief and confidence in their ability in your course.

V. PROBLEM-SOLVING

A. Overview of problem-solving assessments

Students' ability to solve a problem when there is no solution method obvious to the solver[46] is a key skill that many physics faculty would like their students to develop. Problem-solving can be defined in many ways, e.g., the ability to solve physics textbook problems[47] or a collection of many components that a solver brings to bear to solve any problem, regardless of discipline.[48] Because of the variety of interpretations of what problem solving means, there are also a variety of instruments to measure different aspects of problem solving, including the Minnesota Assessment of Problem Solving rubric[49] (MAPS), the Colorado Assessment of Problem Solving[48] (CAPS), and the Assessment of Textbook Problem Solving Ability[47] (ATPSA). There are also surveys to probe students *attitudes* about problem-solving, rather than their skills (AAPS[26,27] and APSS[28]). These are discussed in Sec. IV A 2. To learn more about the research in problem solving, primarily in physics, see "Resource Letter RPS-1: Research in problem solving."[50]

The Minnesota Assessment of Problem Solving[49] (MAPS) rubric is a rubric that you can use to score your students' written solutions using the following 5 categories of problem-solving: (1) useful description, (2) physics approach, (3) specific application of physics, (4) mathematical procedures, and (5) logical progression. The MAPS rubric is applicable to a wide variety of problem types and introductory physics topics. With this rubric, you score each student's written solution from 1 to 5 for each category, and then, look at the frequency of rubric scores for each category across the students in your class to get a sense of their problem-solving strengths and weaknesses. The MAPS rubric has been used at the high school and introductory college level. This rubric was created based on years of research on student problem solving at the University of Minnesota[51–53] and has been extensively studied for evidence for validity, reliability, and utility.[54]

The Colorado Assessment of Problem-Solving[48] (CAPS) is an open-ended problem-solving assessment which presents a general problem situation from the Jasper Woodbury Series[55] that is not tied to any specific discipline, so that students do not have to understand any particular physics concept in order to complete the assessment. The CAPS consists of a script describing a scenario and questions about how to solve the problems in that scenario. Students' responses to the questions are graded on a continuum using a rubric that assesses 44 different sub-skills of the problem-solving process, to gauge students' strengths and weaknesses in problem solving. There is no overall score, as the CAPS is meant to help you assess which aspects of problem solving an individual student needs more help with. It is appropriate for any level of student (middle school to graduate students). These 44 sub-skills are divided into three categories as follows: (1) knowledge; (2) beliefs, expectations, and motivation; and (3) processes. Use it to give individual guidance to specific students, e.g., undergraduate research student and graduate student. It would not be appropriate to use to assess problem solving as a whole in your class.

The Assessment of Textbook Problem Solving Ability[47] (ATPSA) contains open-ended problems similar to the end of chapter textbook problems. The content covered on the ATPSA is intentionally limited to Newton's laws, energy, and momentum, as these are commonly taught topics in introductory courses. The ATPSA is meant for introductory undergraduate calculus-based mechanics courses, uses right/wrong grading, and can be given as a pre- and post-test, so the overall results can be used to evaluate a course (but not individual students). The ATPSA can help instructors assess the impact of teaching reforms on students' ability to solve traditional physics problems. Basic algebra and trigonometry are required to solve the problems. There is a range of difficulty in the ATPSA questions so that the test can assess students of varying levels, though the level of mathematics required for the questions does not change with the difficulty. There are no questions where a mathematical "trick" is needed. The questions on the ATPSA were created by the test developers.

46. *Thinking, Problem Solving, Cognition*, 2nd ed., R. E. Mayer, W. H. Freeman and Company, New York (1992). (E)

47. "Development of a survey instrument to gauge students' problem-solving abilities," J.Marx and K. Cummings, AIP Conf. Proc. **1289**, 221–224 (2010). (E)

48. "Analyzing the many skills involved in solving complex physics problems," W. K. Adams and C. E. Wieman, Am. J. Phys. **83**(5), 459–467 (2015). (E)

49. "Assessing student written problem solutions: A problem-solving rubric with application to introductory physics," J. L. Docktor, J. Dornfeld, E. Frodermann, K. Heller, L. Hsu, K. A. Jackson, A. Mason, Q. X. Ryan, and J. Yang, Phys. Rev. Phys. Educ. Res. **12**, 010130–1–18 (2016). (E)

50. "Resource Letter RPS-1: Research in problem solving," L. Hsu, E. Brewe, T. M. Foster, and K. A. Harper, Am. J. Phys. **72**(9), 1147 (2004). (E)

51. "Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving," P. Heller, R. Keith, and S. Anderson, Am. J. Phys. **60**(7), 627–636 (1992). (E)

52. "Sex differences in physics learning and evaluations in an introductory course," J. M. Blue, Dissertation, University of Minnesota (1997). (I)

53. "The development of students' problem-solving skills from instruction emphasizing qualitative problem-solving," T. Foster, Dissertation, University of Minnesota (2000). (I)

54. "Development and validation of a physics problem-solving assessment rubric," J. Docktor, Dissertation, University of Minnesota (2009). (I)

55. *The Jasper Project: Lessons in Curriculum, Instruction, Assessment, and Professional Development*, Cognition and Technology Group at Vanderbilt, Lawrence Erlbaum Associates, Mahwah, New Jersey (1997). (E)

B. Recommendations for choosing a problem-solving assessment

If you want to use a standardized method of scoring your students' written solutions to your own physics problems and want to get a better sense of your students' strengths and weaknesses with particular problem-solving skills, use the MAPS rubric. If you have a small number of students (undergraduate research students, graduate students, etc.,), you want to understand their problem-solving strengths and weaknesses in great depth, and you have time to individually go through the problem-solving exercise and associated questions with them, use the CAPS. If you want to assess your students' problem-solving skills on textbook-like problems that cover Newton's laws, momentum, and energy, want something that is standardized so that you can compare over time and to others, and is reasonably easy to score, use the ATPSA.

VI. SCIENTIFIC REASONING

A. Overview of scientific reasoning assessments

Scientific reasoning is an important skill that many faculties would like their students to develop. Most generally we can think of scientific reasoning skills as those needed to conduct scientific inquiry including evidence evaluation, inference, and argumentation to form theories about the natural world.[56] There are two assessments of scientific reasoning that have been used in physics: the Lawson Classroom Test of Scientific Reasoning[57] (CTSR) and the Scientific Abilities Assessment Rubrics[58] (SAARs). The Physics Lab Inventory of Critical Thinking[59] (PLIC) also assesses aspects of students' scientific reasoning skills but focuses more on their reasoning skills as related to labs and is discussed in Sec. VII.

The Lawson Classroom Test of Scientific Reasoning Ability[57] (CTSR) is a multiple-choice pre/post-test with

Table VI. Scientific reasoning assessments.

Assessment	Focus	Format	Intended population	Research validation	Purpose
Lawson Classroom Test of Scientific Reasoning (CTSR)	Proportional thinking, probabilistic thinking, correlational thinking, hypothetico-deductive reasoning	Multiple-choice	Intro college, high school, middle school	Gold	Measure concrete- and formal-operational reasoning.
Scientific Abilities Assessment Rubric (SAAR)	Represent information in multiple ways, design and conduct experiments, communicate scientific ideas, collect and analyze experimental data, evaluate experimental results	Rubric	Intro college, high school	Silver	Assess students' scientific abilities as evidenced in their writing around experiments and design tasks.

questions on conservation, proportional thinking, identification of variables, probabilistic thinking, and hypothetico-deductive reasoning. Lawson describes scientific reasoning as consisting of "a mental strategy, plan, or rule used to process information and devise conclusions that go beyond direct experience."[60] The CTSR was originally intended to help instructors classify students' reasoning abilities as concrete, transitional, or formal, based on the total number of questions they answer correctly. However, this method of classifying students reasoning abilities using their CTSR score is antiquated, oversimplified, and problematic,[61] and it is not clear that the CTSR is measuring only one construct. The validity of the most recent version of the CTSR has been recently studied, and issues were found.[62] Of the 12 pairs of questions on the CTSR, 5 pairs were found to have design problems, e.g., students were answering incorrectly when they did understand the content, or students were confused by details given and misinterpreted the question. The problematic questions came from three clusters of questions: proportional reasoning, control of variables, and hypothetic-deductive reasoning. For a more thorough discussion of the validity of the CTSR, see Bao *et al.*[62] Because of these shortcomings around CTSR questions, the overall score should not be used to classify individual students, and the results should not be used as a stand-alone proxy for your students reasoning abilities. Instead, looking at the change in the overall CTSR score between pre- and post-test for your class can give you a sense of how your course influences students' reasoning abilities because some of the effects of poor question design will probably average out over larger numbers of students. If you do this, you should use the CTSR in combination with other measures of reasoning abilities. Instructors can also use the percentage correct on the CTSR for each cluster of questions to get a sense of their students' strengths and weaknesses around different aspects of scientific reasoning (while taking into account that questions in the proportional reasoning, control of variables, and hypothetic-deductive reasoning clusters have design issues). The Lawson test was developed for high school students but has also been used at the introductory college level. The questions were originally based on demonstrations, where the instructor would perform the demonstration and then ask students questions about it in an interview format. The most recent version has converted these interview questions into a multiple-choice paper and pencil test.

The Scientific Abilities Assessment Rubrics[58] (SAARs) are a set of rubrics used to assess students' levels of competence around seven different scientific abilities, which are as follows:

1. The ability to represent physical processes in multiple ways.
2. The ability to devise and test a qualitative explanation or quantitative relationship.
3. The ability to modify a qualitative explanation or quantitative relationship.
4. The ability to design an experimental investigation.
5. The ability to collect and analyze data.
6. The ability to evaluate experimental predictions and outcomes, conceptual claims, problem solutions, and models.
7. The ability to communicate.

The SAARs are used to assess specific scientific abilities as evidenced in students' written work around experiments or design tasks. The Scientific Abilities Assessment Rubrics outline the different levels of performance (0, missing; 1, inadequate; 2, needs some improvement; and 3, adequate) and include a description of each level, to enable students to self-assess as they work toward developing these abilities. In this way, the SAARs enable formative assessment of students' scientific abilities. Instructors can also use the SAARs to assess their students' acquisition of these scientific abilities by scoring students' laboratory write-ups for a particular experiment or design task from 0 to 3 using the descriptions of the different levels on the rubric. Instructors can then compare the distribution of scores for a particular scientific ability at the beginning and end of the course in hopes of seeing more students scoring "adequate." The SAARs were developed in the context of introductory college courses, though may also be appropriate for high school and intermediate college classes. The list of scientific abilities is based on literature on the history of the practice of physics, a taxonomy of cognitive skills, recommendations of science educators, and an analysis of science-process test items.

56. "The development of scientific thinking skills in elementary and middle school," C. Zimmerman, Dev. Rev. **27**, 172–223 (2007). (E)

57. "The development and validation of a classroom test of formal reasoning," A. E. Lawson, J. Res. Sci. Teach. **15**(1), 1978 (1978). (E)

58. "Scientific abilities and their assessment," E. Etkina, A. Van Heuvelen, S. White-Brahmia, D. T. Brookes, M. Gentile, S. Murthy, D. Rosengrant, and A. Warren, Phys. Rev. Spec. Top.-Phys. Educ. Res. **2**(2), 020103 (2006). (E)

59. "Preliminary development and validation of a diagnostic of critical thinking for introductory physics labs," N. G. Holmes and C. E. Wieman, Phys. Educ. Res. Conf. Proc. 156–159 (2016). (E)

60. "The nature and development of scientific reasoning: A synthetic view," A. E. Lawson, Int. J. Sci. Math. Educ. **2**(3), 307–338 (2004). (I)

61. "What do tests of 'formal' reasoning actually measure?," A. E. Lawson, J. Res. Sci. Teach. **29**, 965–983 (1992). (I)

62. "Validity evaluation of the Lawson classroom test of scientific reasoning," L. Bao, Y. Xiao, K. Koenig, and J. Han, Phys. Rev. Phys. Educ. Res. **14**(2), 20106 (2018). (I)

B. Recommendations for choosing scientific reasoning assessment

Use the CTSR if you want to assess your students' reasoning skills, possibly in conjunction with an appropriate test of their mathematical skill or physics content knowledge and other assessments of reasoning skills, as many shortcomings have been identified in the CTSR, and it relying on the CTSR score as the only measure of your students reasoning skills would be problematic. Do not use this test if you need more detailed information about a specific student or group of students (such as for placement into a particular class), because the design and validity issues with the test do not average out over smaller numbers of students.

Use the SAARs to help your students self-assess their scientific abilities in lab courses. You can also use the SAARs as an instructor to give your students feedback on their competency around specific scientific abilities and sub-abilities and look at how your students' scores change over the course of your class.

VII. LABORATORY SKILLS

A. Overview of laboratory skill assessments

Faculty often assume that during the laboratory portion of a physics course, students develop the ability to gather and evaluate data through experiments. Several assessments of different aspects of lab skills have been created to help instructors evaluate their students' laboratory skills and critical thinking ability at the beginning and end of the course. There are four assessments of lab skills, the Physics Lab Inventory of Critical Thinking[59] (PLIC), the Concise Data Processing Assessment[63] (CDPA), the Physics Measurement Questionnaire[64,65] (PMQ), and the Measurement Uncertainty Quiz[66] (MUQ). There is also an assessment to gauge students' attitudes about experimental physics (E-CLASS[25]), which is discussed in Sec. IV A 2. The Data Handling Diagnostic[67] (DHD) is another assessment of laboratory skills, which will not be discussed further here because the authors did not finish the development and validation of this assessment and advise others to use the CDPA instead of the DHD.

The Physics Lab Inventory of Critical Thinking[59] (PLIC) assesses the way students critically evaluate experimental methods, data, and models and is the newest laboratory skills assessment. The PLIC includes an introduction that describes an experiment using masses and spring and sample laboratory notebook entries for two groups of physicists. The PLIC uses "choose many" multiple-choice questions and Likert-scale questions to assess students' critical thinking around the laboratory notebook entries for this experiment. Students' responses are compared to the "consensus expert response" and "consensus appropriate response" for each question. Because many of the multiple-choice questions allow students to "choose many," the score for each question is between 0 and 4 points, depending on how many "consensus expert responses", "consensus appropriate responses" and "inovice responses" are given. The PLIC has been used in all levels of undergraduate laboratories. The PLIC is still under development. The questions on the PLIC were based on the series of questions an expert posed to himself or herself when conducting an introductory physics experiment.

The Concise Data Processing Assessment[63] (CDPA) is a 10 question multiple-choice pre-post assessment that measures students' understanding of handling data with questions around uncertainty in measurements and the relationships between functions, graphs, and numbers. The CDPA is appropriate to use in any laboratory course with learning goals around data handling. The questions were based on established learning goals for an introductory laboratory course and iteratively refined using student interviews, expert review, and statistical analyses.

Both the PLIC and the CDPA assess students' data analysis skills, but the PLIC also assesses other skills including how students critically evaluate experimental methods, data, and models. The CDPA has 10 multiple-choice questions, where each has its own context, whereas the PLIC has one rich experimental context outlined at the beginning of the assessment, to which all 16 questions refer. Both the PLIC and CDPA have strong research validation.

The Physics Measurement Questionnaire[64,65] (PMQ) is an open-ended pre/post assessment of students' understanding of experimental measurements, including data collection, data processing, and dataset comparison. There is an experimental situation described at the beginning of the assessment, and all the questions refer to this same experimental situation (similar to the PLIC). The questions ask students to reflect on how many measurements they should take, how to report the results of multiple measurements, how to compare sets of measurements, and how to fit a line to experimental data. Because the PMQ questions are open-ended, the answers and explanations are coded according to an established coding scheme, which can be time consuming. In each question or "probe," there is a short conversation between several people, and students are asked to choose which they most agree with and then give a written explanation for their choice. The discussions in each probe are written with concise, simple language in order to be understandable for a wide range of English language levels. The developers use the PMQ results to look at their students' paradigms of measurement as either "point" or "set." A point paradigm would see each measurement as the possible true value, where differences between measurements are a result of environmental factors or experimenter mistakes.[68] In the "set" paradigm, each measurement is an approximation of the true value, and deviations are random and always present. A set of measurements yields the best approximation of the true value, with an associated uncertainty. The questions on the PMQ were based on similar questions from the Procedural and Conceptual Knowledge in Science (PACKS) Project.[69]

The PMQ has a unique format compared to the other laboratory skill assessments, where each question includes a conversation between students, with an open-ended question about the conversation. Furthermore, the scoring of the PMQ is different from the other assessments discussed, and instructors code the responses to understand their students' results. The content and skills assessed on the PMQ are also included in the PLIC, though the PLIC goes into more depth in asking students to evaluate critically experimental methods, data, and models.

The Measurement Uncertainty Quiz[66] (MUQ) is a nonstandard assessment that can be used as the basis of a discussion about precision, significant figures, accuracy, and error propagation with your introductory physics students. The developer explains that it is difficult to create a right/wrong test around the topics of measurement and uncertainty, because even experts may disagree on the correct answer. Because of this limitation, the MUQ questions are an opportunity to discuss with your students why one answer may be better than others. Because the MUQ is for discussion (and is not scored), it is not given as a pre/post-test. The 10 questions on the MUQ are a sample of the open-ended questions given to approximately 100 introductory physics students and 30 experts (graduate physics students and teachers). The most common responses were edited and turned into the multiple-choice options.

The MUQ focuses just on measurement uncertainty, whereas the CDPA also asks about fitting data and relating functions, graphs, and numbers. Both tests use the same question format and have the same number of questions, but the MUQ developers recommend using it to have a discussion with students, instead of using it as a pre/post-test and scoring it, as you would with the CDPA.

63. "Development of the Concise Data Processing Assessment," J. Day and D. Bonn, Phys. Rev. Spec. Top.-Phys. Educ. Res. **7**(1), 010114 (2011). (I)

64. "Point and set reasoning in practical science measurement by entering university freshmen," F. Lubben, B. Campbell, A. Buffler, and S. Allie, Sci. Educ. **85**(4), 311–327 (2001). (E)

65. "First-year physics students' perceptions of the quality of experimental measurements," S. Allie, A. Buffler, L. Kaunda, B. Campbell, and F. Lubben, Int. J. Sci. Educ. **20**(4), 447–459 (1998). (E)

66. "Introductory physics students' treatment of measurement uncertainty," D. L. Deardorff, Dissertation, North Carolina State University (2001). (I)

67. "Diagnostic tests for the physical sciences: A brief review," S. Bates and R. Galloway, New Dir. (6), 10–20 (2010). (E)

68. "Impact of a conventional introductory laboratory course on the understanding of measurement," T. S. Volkwyn, S. Allie, A. Buffler, and F. Lubben, Phys. Rev. Spec. Top.-Phys. Educ. Res. **4**(1), 1–10 (2008). (E)

69. "Investigating in the school science laboratory: Conceptual and procedural knowledge and their influence on performance," R. Millar, F. Lubben, R. Got, and S. Duggan, Res. Pap. Educ. **9**(2), 207–248 (1994). (E)

B. Recommendations for choosing laboratory skill assessment

Use the PLIC to get a rich understanding of your students' skills around critically evaluating experimental methods, data, and models. The PLIC assesses the content covered on the MUQ, CDPA, and PMQ and additional content and skills related to critical thinking around experimentation. If you want a short, simple multiple-choice test of measurement uncertainty and relationships between functions, data, and graphs, use the CDPA. If you are interested in understanding your students' open-ended responses about data collection, processing, and comparison or in looking at your students' paradigms of the measurement as either "point" or "set," use the PMQ. If you want to have a rich conversation about measurement uncertainty with your students, use the MUQ as the basis of the conversation.

VIII. OBSERVATION PROTOCOLS

A. Overview of observation protocols

Faculty in physics departments often observe each other's teaching and give each other feedback to improve teaching. Using an observation protocol for these informal observations can help faculty articulate the goals of these observations and focus on particular aspects of the classroom. Observation protocols can provide data that illustrate what happened in the class, which can be useful for self-reflection and professional development. You can use observation protocols once as a stand-alone activity or to track your own improvement.

Classroom observations using observational protocols can be conducted using segmented, continuous, and holistic procedures.[71] Segmented protocols are those in which the class period is broken up into shorter periods of time, 2-min intervals for example, and then observers note whether they see certain behaviors during that interval or not. At the end of the observation, observers note the number of intervals in which each of the different behaviors happened. Continuous protocols allow observers to indicate what is happening at any given moment in a class, and an observation results in a time-line indicating what happened when. This also allows the different classroom activities to be considered as a certain percentage of overall class-time. Holistic protocols are protocols in which the entire course is considered at once. This is done using a series of questions that the observer responds to at the end of an observation.

There are seven observation protocols that we will discuss here. Four of these protocols focus on recording what is happening in the classroom. These are the Classroom Observation Protocol Undergraduate STEM[72] (COPUS), the Teaching Dimensions Observation Protocol[73,74] (TDOP), the Real-time Instructor Observation Protocol[75,76] (RIOT), and the Student Participation Observation Tool[77] (SPOT). One protocol, the Laboratory Observation Protocol for Undergraduate STEM[71] (LOPUS), focuses on recording what is happening in laboratory courses. One protocol, the Reformed Teaching Observation Protocol[78,79] (RTOP), focuses specifically on assessing the degree of reformed teaching. Finally, one protocol, the Behavioral Engagement Related to Instruction[80] (BERI), looks at the level of student engagement in a class session. All of these observation protocols can be used in high school or college-level courses.

Perhaps the most well-known observation protocol is the Reformed Teaching Observation Protocol[78,79] (RTOP), a holistic paper and pencil observation protocol developed to evaluate the extent to which a classroom uses reform-based teaching techniques, meaning "constructivist, inquiry-based methods."[79] The RTOP developers operationally define "reformed teaching" as "classroom practices that result in a high RTOP score."[79] The RTOP consists of 25 Likert-scale items from three different categories including, "lesson design and technique," "content," and "classroom culture." Observers watch a class session and respond to each item with a maximum of 4 meaning that the item is "very descriptive" of the class to a minimum of zero indicating that the item "never occurred." The RTOP data can be reduced to a single score by adding up the scores for each item. A higher RTOP score means that a class is more reformed, meaning that the course is more active and student-centered. The single RTOP score makes it particularly useful as quantitative evidence of instructor change in practice over time. The RTOP has been used very widely, and RTOP scores have been shown to correlate with conceptual learning gains in college physics courses.[81] There are several questions on the RTOP that evaluate the instructor on the content or design of the lesson, and so, it is more appropriate to use the RTOP with instructors that designed the lesson themselves (and not a teaching assistant who did not have autonomy in deciding what happens in the classroom). The RTOP developers emphasize that RTOP results are not valid unless the observers have gone through several days of training on how to use the instrument. While in-person training is best, there is online training available.[82] The items on the RTOP were developed based on previous research and existing instruments.[78]

The Teaching Dimensions Observation Protocol[73,74] is a segmented observation protocol that aims to record what is happening in the classroom, unlike the RTOP, which is designed to evaluate the degree of reformed teaching. The TDOP looks at three basic dimensions of the classroom including "instructional practices," "student-teacher dialogue," and "instructional technology," and three optional dimensions, including, "potential student cognitive engagement," "pedagogical strategies," and "students' time on task." Each of these dimensions has codes associated with them, and observers memorize the meaning of these codes (28 basic and 11 optional) and circle that code when it happens during each 2-min interval of an observation. Observers can collect data with pencil and paper or with a computerized interface available on the TDOP website.[83] Once data are collected, observers can examine the percentage of intervals that each code (or code category) appears. The TDOP website also automatically creates some charts and graphs for review. TDOP creators recommend that users establish inter-rater reliability and stress that training may take several days depending on how many dimensions are used. Both a TDOP users guide and TDOP scoring sheet are available for download.[83] The codes and categories on the TDOP were developed based on an instrument designed to study inquiry-based middle school sciences courses.[84]

The Classroom Observation Protocol for Undergraduate STEM[72] (COPUS) was developed based on iterative modifications of an early version of the TDOP, so it is also a segmented protocol and is similar in many ways. The COPUS developers aimed to create a more user-friendly version of the TDOP (though the version of the TDOP they were working with had more mandatory categories, and the newer version of the TDOP discussed in this paper has been simplified). COPUS codes are separated into two broad categories, "what teachers are doing" and "what students are doing," with a total of 25 codes using a simplified language. This allows individuals to learn to use COPUS much more quickly than the TDOP or RTOP, in as few as 1.5 hours. The COPUS developers also added some categories that were aligned with best practices in large-enrollment college-level STEM courses, such as discussions motivated by clicker questions. Like the TDOP, observers indicate whether a certain behavior happened or not in each 2-min period using a specialized scoring sheet. The COPUS developers have also recently developed the COPUS profiles online tool[85] that allows a user to upload COPUS data in a spreadsheet in order to create several different visual representations of these data that can be helpful for reflection.

The Real-time Instructor Observation Protocol[75,76] (RIOT) was developed independently from COPUS at the same time, and therefore, the two were developed to fill similar needs but with slightly different focuses. RIOT, which is similar to COPUS and TDOP, allows an observer to categorize what is happening during a classroom observation. Unlike the COPUS and TDOP, the RIOT is a continuous web-based protocol that only follows the instructor and records what they are doing (including if they are interacting with students) but does not record what students are doing independently of the instructor. The categories for RIOT are organized by the types of interactions that are possible with students in the classroom, "talking at students," "talking with students," "observing students," and "not interacting with students." An observer clicks on icons representing these categories in order to indicate that a certain interaction is occurring as they are observing a classroom and continuously clicks new observations as they are observed. The web program records timestamps for each observation. The RIOT was originally developed as a part of a Teaching Assistant (TA) pedagogy course to help new graduate student TAs understand how to interact with students in an active learning environment, so it is useful for helping faculty as well as teaching assistants understand and improve their teaching. Like COPUS, RIOT requires little training to use. The RIOT categories were developed based on observations of classrooms using the Collaborative Learning through Active Sense-making in Physics (CLASP) curriculum[86] at University of California at Davis and emergent behaviors seen there.

The Student Participation Observation Protocol[77] (SPOT) is an observation protocol very similar to the RIOT in that it is web-based and continuous and has the same developers, but there are a few key differences in the content and layout. SPOT had a more rigorous development process than RIOT, as categories are backed by research on student-centered learning in science classrooms. SPOT categories represent the observable features of seventeen of the best practices in active learning.[77] Different from RIOT, but similar to the COPUS, the SPOT records what both the instructor and students are doing (whereas RIOT focuses on the instructor) and is organized by class "mode" referring to how the instructor and students are interacting with each other at any given time. The class can be in "small-group mode," where students are working in small groups, "whole class mode," where students are watching a lecture, movie, or demo, and "independent mode" where students are working silently and independently (such as when they are taking an exam). In each mode, different codes are available to describe different behaviors of instructors and students. SPOT is optimized for courses that include some traditional lecture elements in order to better classify how participation happens, and who is participating. For example, during a lecture where the instructor may interact by asking or answering questions, SPOT allows an observer to classify student responses as either shouted-out, asking a question, answering a question, contributing an idea, or via whole-class choral response. SPOT also allows the observer to keep track of individual students using a map interface based on where they are sitting in the room. This can help instructors determine if many students are participating, or if it is the same five or six each time. Since SPOT is web-based like RIOT, it also generates colorful figures useful for self-reflection. To see examples of these figures, see the PhysPort assessment pages for the RIOT[87] and SPOT.[88]

The Laboratory Observation Protocol for Undergraduate STEM[71] (LOPUS) was developed to categorize student and teacher actions in laboratory settings. The LOPUS creators started their development with a draft of the COPUS, then reviewed the literature and watched video of laboratory classes, to determine new behaviors that should be added to the LOPUS, which were not included in the COPUS. Like the COPUS, the LOPUS is a segmented protocol and organized into two broad categories of instructor behaviors and student behaviors, but LOPUS also has a third category that captures the content of student and teacher verbal interactions in laboratory classes, and who (teacher or student) initiated the interaction. For example, someone viewing an instructor lecturing about data analysis would use the pair of codes: "Lec" (indicating that the instructor is lecturing) and a qualifying code from this third category, "Ana" (indicating that the conversation is about data analysis and calculations). The LOPUS team also cut some of the codes from the COPUS that they found were highly correlated with each other, in order to cut back on the number of codes an observer needed to memorize. The LOPUS is available in a web-based format through the General Observation and Reflection Platform (GORP).[89] The platform auto-creates charts and plots that are useful for reflection.

The Behavioral Engagement Related to Instruction[80] (BERI) protocol is a segmented observation protocol to measure student behavioral engagement, defined as on-task behavior, in large university classes. The BERI can help an instructor figure out which parts of their class resulted in higher student engagement. The BERI protocol outlines six engaged behaviors, for example, listening, writing, and engaged instructor interactions, and six unengaged behaviors, for example, settling in/packing up, being off-task, or disengaged computer use. The observer chooses a group of ten students and sits near them. During the class, the observer cycles through each of the 10 students and records, on a printout of instructor notes, if each student was engaged or disengaged during part of the class. The BERI observation protocol categories were developed based on observations of large classes to determine which student behaviors could be defined as engaged and disengaged.

The BERI protocol focuses particularly on student engagement, whereas the other protocols discussed above have a more general focus. The COPUS, LOPUS, TDOP, and SPOT all record student behaviors during the class, but they do not label these behaviors as engaged or disengaged.

Several of the observation protocols mentioned here have been incorporated into web-based tools to make them easier to use. See our expert recommendation on PhysPort[70] for more information on accessing these online protocol tools.

70. "Which observation protocols are available online?," C. Paul and A. Madsen, <https://www.physport.org/recommendations/Entry.cfm?ID=110649>. (E)

71. "Characterizing instructional practices in the laboratory: The laboratory observation protocol for undergraduate STEM," J. B. Velasco, A. Knedeisen, D. Xue, T. L. Vickrey, M. Abebe, and M. Stains, J. Chem. Educ. **93**(7), 1191–1203 (2016). (I)

72. "The classroom observation protocol for undergraduate stem (COPUS): A new instrument to characterize university STEM classroom practices," M. K. Smith, F. H. M. Jones, S. L. Gilbert, and C. E. Wieman, CBE Life Sci. Educ. **12**(4), 618–627 (2013). (E)

73. "Toward a descriptive science of teaching: How the TDOP illuminates the multidimensional nature of active learning in postsecondary classrooms," M. T. Hora, Sci. Educ. **99**(5), 783–818 (2015). (I)

74. "Teaching Dimensions Observation Protocol (TDOP) user's manual," M. T. Hora, A. Oleson, and J. J. Ferrare, p. 28, Wisconsin Center for Education Research, University of Wisconsin-Madison (2013). (I)

75. "Using the Real-time Instructor Observing Tool (RIOT) for reflection on teaching practice," C. Paul and E. West, Phys. Teach. **56**(3), 139–143 (2018). (E)

76. "Variation of instructor-student interactions in an introductory interactive physics course," E. A. West, C. A. Paul, D. Webb, and W. H. Potter, Phys. Rev. Spec. Top.-Phys. Educ. Res. **9**(1), 010109 (2013). (E)

77. "Observable features of active science," K. Roseler, C. A. Paul, M. Felton, and C. H. Theisen, J. Coll. Sci. Teach. **47**(6), 83–92 (2018). (E)

78. "Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol," D. Sawada, M. D. Piburn, E. Judson, J. Turley, K. Falconer, R. Benford, and I. Bloom, Sch. Sci. Math. **102**(6), 245–252 (2002). (I)

79. "Reforming physics instruction via RTOP, " D. MacIsaac and K. Falconer, Phys. Teach. **40**, 479–485 (2002). (E)

80. "A new tool for measuring the behavioral engagement in large university classes," E. S. Lane and S. E. Harris, J. Coll. Sci. Teach. **44**(6), 83–91 (2015). (E)

Table VII. Laboratory skill assessments.

Assessment	Focus	Intended population	Format	Research validation	Purpose
Physics Lab Inventory of Critical Thinking (PLIC)	Evaluating experimental methods, generating and evaluating conclusions based on data, comparing measurements with uncertainty, evaluating data fitted to a model	Upper-level, intermediate, intro college	Agree/disagree, multiple-choice, multiple-response, available online [5]	Silver	To assess how students critically evaluate experimental methods, data, and models.
Concise Data Processing Assessment (CDPA)	Measurement and uncertainty, relationships between functions, numbers, and graphs	Graduate, upper-level, intermediate, intro college	Multiple-choice	Silver	To probe student abilities related to the nature of measurement and uncertainty and to handling data.
Physics Measurement Questionnaire (PMQ)	Measurement and uncertainty	Intro college	Open-ended	Silver	To probe students understanding of measurement and uncertainty using open-ended sample discussions.
Measurement Uncertainty Quiz (MUQ)	Measurement and uncertainty	Intro college	Discussion of multiple-choice	Bronze	To discuss measurement and uncertainty concepts with students, and why one answer might be better than the others.

81. "Effect of Reformed Courses in Physics and Physical Science on Student Conceptual Understanding," K. Falconer, S. Wyckoff, M. Joshua, and D. Sawada, in Annual Conference of the American Educational Research Association, Seattle, WA (2001). (E)

82. "Using RTOP," <http://physicsed.buffalostate.edu/AZTEC/RTOP/RTOP_full/using_RTOP_1.html>. (I)

83. "Welcome to the TDOP," <http://tdop.wceruw.org>. (E)

84. "Implementing immersion: Design, professional development, classroom enactment and learning effects of an extended science inquiry unit in an urban district," E. Osthoff, W. Clune, J. Ferrare, K. Kretchmar, and P. White, Madison, WI (2009). (E)

85. "COPUS Profiles," <http://www.copusprofiles.org/>. (E)

86. "Sixteen years of Collaborative Learning through Active Sense-making in Physics (CLASP) at UC Davis," W. Potter, D. Webb, E. West, C. Paul, M. Bowen, B. Weiss, L. Coleman, and C. De Leone (2012). (E)

87. "Real-time Instructor Observing Tool (RIOT)", <www.physport.org/assessments/RIOT>. (E)

88. "Student participation observation tool (SPOT)", <www.physport.org/assessments/SPOT>. (E)

89. "Tools for evidence-based action," <http://tea.ucdavis.edu>. (E)

B. Recommendations for choosing an observation protocol

While all the observation protocols discussed here are potentially useful for self-reflection and professional development, we particularly recommend the COPUS and RIOT for these purposes based on their short training times, and resources for self-training. Use the COPUS for your professional development and self-reflection if you are particularly interested in what specific pedagogical tools are used in the classroom (e.g., students making a prediction, instructor showing a demonstration). Use RIOT if you are more concerned with what the instructor is doing more generally, and their interactions with students (e.g., the instructor is explaining content, the instructor is listening to a question). If you are interested in evaluating how reformed a course is, especially if you want to apply a numeric score to this evaluation to compare to national results, and you can attend a training, use the RTOP. If you want a detailed account of what pedagogical actions take place in a classroom and have time for training, use the TDOP. Use SPOT if you have questions about the frequency, type, and diversity of student participation in the classroom. Use the LOPUS if you are interested in lab environments and use the BERI if you are particularly interested in how your students' level of engagement in class depends on what you are doing in class.

IX. SURVEY OF FACULTY TEACHING PRACTICE

There are two surveys of faculty instructional practices that are commonly used in physics, The Teaching Practices Inventory[90] (TPI) and the Postsecondary Instructional Practices Survey[91] (PIPS). Both are research-based surveys that ask faculty to self-report on their teaching and the kinds of things that go on in their classrooms and the durations. Researchers use these surveys to characterize the self-reported teaching practices of faculty, though the results could also be used by faculty themselves to illustrate their teaching practices in tenure and promotion documents. Since this Resource Letter focuses on assessments that give faculty new information that they can use to understand and improve what is happening in their specific course, and these self-

Table VIII. Observation protocols.

Title	Focus	Intended population	Format	Research validation	Purpose
Reformed Teaching Observation Protocol (RTOP)	Degree of reformed teaching	All levels	Holistic	Gold	To assess the degree to which reformed teaching is occurring in classrooms.
Teaching Dimensions Observation Protocol (TDOP)	Instructor and student classroom behaviors	All levels	Segmented, available online[70]	Gold	To classify instructor and student behaviors in STEM classrooms.
Classroom Observation Protocol Undergraduate STEM (COPUS)	Instructor and student classroom behaviors	All levels	Segmented, available online[70]	Silver	To classify instructor and student behaviors in STEM classrooms.
Real-time Instructor Observation Protocol (RIOT)	Instructor-student classroom interactions	All levels	Continuous, available online[70]	Silver	To classify instructor interactions with students in STEM classrooms.
Student Participation Observation Tool (SPOT)	Instructor and student classroom behaviors	All levels	Continuous, available online[70]	Bronze	To classify instructor and student behaviors in STEM classrooms.
Laboratory Observation Protocol for Undergraduate STEM (LOPUS)	Instructor and students' lab behavior	All levels	Segmented, available online[70]	Bronze	To classify instructor and student behaviors in STEM labs.
Behavioral Engagement Related to Instruction (BERI)	Student engagement	All levels	Segmented	Bronze	To quantitatively measure student engagement in large university classes

report surveys of teaching practice are about faculty perceptions of their teaching more generally, we will not discuss these surveys in more detail.

90. "The Teaching Practices Inventory: A new tool for characterizing college and university teaching in mathematics and science," C. Wieman and S. Gilbert, CBE-Life Sci. Educ. **13**, 552–569 (2014). (E)

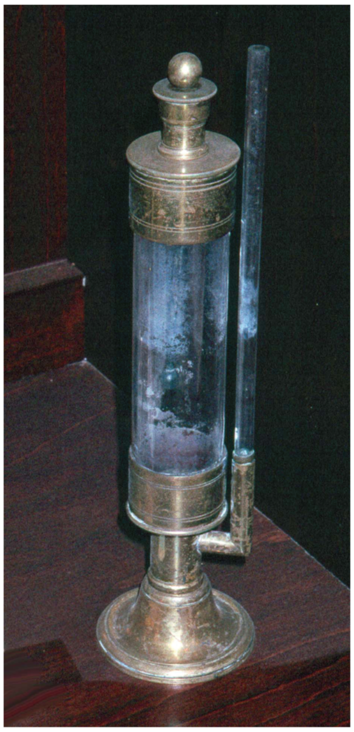
91. "Introducing the Postsecondary Instructional Practices Survey (PIPS): A concise, interdisciplinary, and easy-to-score survey," E. M. Walter, C. R. Henderson, A. L. Beach, and C. T. Williams, CBE Life Sci. Educ. **15**(4), 1–11 (2016). (E)

X. CONCLUSION

This paper summarizes major RBAIs in non-physics-content areas: mathematics (Table III), beliefs and attitudes (Table IV), problem solving (Table V), scientific reasoning (Table VI), laboratory skills (Table VII), and observation protocols (Table VIII). In contrast with the previous Resource Letter in this series (RL: RBAI-1), this collection of RBAIs is generally used to augment our picture of student learning in physics rather than investigate their understanding of specific physics topics. RBAIs in this collection are useful at all points in the high school and undergraduate curriculum.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of the other members of the PhysPort team and KSUPER who worked on this project: John D. Thompson, Jaime Richards, Devon McCarthy, and Brian Danielak. This work was partially supported by NSF Grant Nos. PHYS-1461251, DUE-1256354, DUE-1256354, DUE-1347821, and DUE-1347728.



Kinnersley Air Thermometer

This device was developed by Ebenezer Kinnersley, a colleague and contemporary of Benjamin Franklin in Philadelphia. It uses the expansion of air as an electric spark passes through it to give a measure of the energy of the spark. Inside the larger glass tube is a pair of discharge knobs, and water fills up the larger tube up to the level of the lower knob. When the spark from a Leiden jar heats and expands the air in this tube, water is forced into the smaller tube; the amount by which the water level increases in this latter tube is therefore a measure of the energy of the electric spark. This apparatus is in the collection of Transylvania University in Lexington, Kentucky. It was probably bought from Pixii in Paris in 1839 for 18 francs. (Picture and Text by Thomas B. Greenslade, Jr., Kenyon College.)