Development and Uses of Upper-division Conceptual Assessments

Bethany R. Wilcox,[1] Marcos D. Caballero,[2] Charles Baily,[3] Homeyra Sadaghiani,[4] Stephanie V. Chasteen,[1] Qing X. Ryan,[1] and Steven J. Pollock[1]

[1]*Department of Physics, University of Colorado, 390 UCB, Boulder, CO 80309*
[2]*Department of Physics and Astronomy & CREATE for STEM Institute, Michigan State University, East Lansing, MI 48824*
[3]*School of Physics and Astronomy, University of St. Andrews, St. Andrews, Fife KY16 9SS Scotland, UK*
[4]*Department of Physics and Astronomy, California State Polytechnic University, Pomona, CA 91768*

The use of validated conceptual assessments alongside conventional course exams to measure student learning in introductory courses has become standard practice in many physics departments. These assessments provide a more standard measure of certain learning goals, allowing for comparisons of student learning across instructors, semesters, institutions, and pedagogies. Researchers at the University of Colorado Boulder have developed several similar assessments designed to target the more advanced physics of upper-division classical mechanics, electrostatics, quantum mechanics, and electrodynamics courses. Here, we synthesize the existing research on our upper-division assessments and discuss some of the barriers and challenges associated with their development, validation, and implementation as well as some of the strategies we have used to overcome these barriers.

PACS numbers: 01.40.Fk, 01.40.G-, 01.40.gf, 01.50.Kw

I. INTRODUCTION

Research-based conceptual assessments represent one of the most commonly adopted tools to come out of the Physics Education Research (PER) community in the last several decades. At the introductory level, the Force Concept Inventory [1] is arguably the most well known of these assessments; however, many other instruments, spanning multiple topical areas, have been developed (see Ref. [1–3] for examples and Ref. [4] for a more comprehensive list). Historically, these assessments have had a number of significant impacts on physics education at the introductory level. For example, they provide a measure of some aspects of student learning that are often not captured by conventional exams. They also represent a standardized and validated tool for evaluating the effectiveness of classroom strategies across instructors, institutions, and time. By providing a measure of student learning across courses and pedagogies, these assessments can support both pedagogical and institutional changes that enhance student learning [5, 6].

Fewer conceptual assessments have been developed to target upper-division physics content, in part because conceptual assessment at the advanced undergraduate level presents some uniques challenges. For example, advanced physics content requires students to employ sophisticated mathematical tools and techniques. This increased emphasis on mathematics makes it more difficult, and perhaps less desirable, to create assessments that focus only on students' conceptual understanding. Additionally, the increased complexity of the physics content makes it challenging to construct clear, level-appropriate questions that can be answered within a reasonably short time frame. The relatively small body of existing research on students' difficulties also makes it more difficult to create questions that specifically target areas where students struggle. Various logistical constraints on the development of standardized assessments also represent a more significant barrier at the upper-division level than the introductory level. For example, less consistency in content coverage between different instructors and institutions makes it more difficult to create a single instrument that matches the learning goals of a majority of courses/instructors. Additionally, small class sizes at the upper-division level hinder efforts to collect enough early-implementation data to achieve sufficient statistical power to ensure the validity and reliability of a new instrument.

Despite these challenges, several conceptual assessments have been developed for the upper-division level, targeting a range of content areas that include (but are not limited to): sophomore classical mechanics [7], junior electricity and magnetism [8–10], quantum mechanics [11–13], and several engineering assessments targeting thermodynamics [14–16] and waves [17]. Additionally, several assessments developed for the introductory and sophomore level have been used productively as pre/post tests at the upper-division level [18, 19]. Note that, here, we are using the term 'conceptual assessment' broadly and include in this category assessments that target aspects of mathematical thinking (rather than procedural mathematics) and strategic processes and practices (e.g., identifying the correct solution method).

In this paper, we will focus on four assessments created at the University of Colorado Boulder (CU) as examples of the development and uses of upper-division conceptual assessments. The goals of this paper are to: (1) present an overview of CUs upper-division conceptual assessments including motivation, approaches, development, and current status of each instrument (Sec. II) while highlighting similarities and differences between our approach and that of others; (2) summarize examples of outcomes from each assessment (Sec. III); and (3) discuss the implementation of these assessments in

Assessment	Format	# of Q's	Level and Content	Standard Text
Colorado Classical Mechanics/Math Methods Instrument (**CCMI**) [7]	FR	11*	Sophomore mechanics up to (but not including) the Lagrangian and Hamiltonian formulations	Taylor [20] Ch. 1-5 ‡
Colorado Upper-division Electrostatics Diagnostic (**CUE**) [6]	FR †	17	Junior electrostatics with minimal magnetostatics	Griffiths [21] Ch. 1-5
Colorado UppeR-division ElectrodyNamics Test (**CURrENT**) [10, 22]	FR	6	Junior electrodynamics up to (but not including) relativistic electrodynamics	Griffiths [21] Ch. 7-9
Quantum Mechanics Assessment Tool (**QMAT**) [13]	FR †	14	Junior quantum mechanics focusing on solutions to the time-independent Schrödinger equation	Griffiths [23] Ch. 1-4
Electromagnetics Concept Inventory (**EMCI**) [8]	MC	23**	Electrodynamics for junior-level engineers including both fields and waves	NA
Symmetry and Gauss's Law Conceptual Evaluation (**SGLCE**) [18]	MC	25	Conceptual understanding of symmetry and Gauss's Law at the introductory level	NA
Quantum Mechanics Survey (**QMS**) [11]	MC	31	Junior quantum mechanics in one spacial dimension	NA
Quantum Mechanics Visualization Instrument (**QMVI**) [12]	MC	25	Introductory through graduate quantum with an emphasis on visualization	NA
Quantum Mechanics Conceptual Survey (**QMCS**) § [19]	MC	12	Quantum mechanics as appropriate for sophomore-level modern physics	NA

TABLE I. Brief overview of specific upper-division conceptual assessments. The top section includes CU's four named upper-division assessments, and the bottom section includes similar information for several of the major alternative instruments that have been developed for or used at the upper division level. Each assessment is classified as either multiple-choice (MC) if the final numerical score comes from only multiple-choice or multiple-response questions, or free-response (FR) otherwise.
§ The QMCS was developed at CU, but was targeted at introductory quantum mechanics and thus is not included as one of CU's upper-division assessments
† There is a multiple-choice adaptation of the QMAT (called the QMCA) and a multiple-response version of the CUE (called the CMR CUE); these versions will be discussed in Sec. II C.
* Only 9 questions on the CCMI contribute to the overall numerical score. The remaining 2 questions target the use of specific mathematical tools (Fourier Series and separation of variables) and are outside the scope of most classical mechanics courses.
** The EMCI is split into two, 23 question versions: one targeting fields and one targeting waves.
‡ Coverage also includes Newton's Universal Law of Gravitation following Ch. 5 of Thornton and Marion [24].

the classroom, including barriers and possible solutions (Sec. IV). We will not be presenting new findings that, for example, compare learning outcomes or unpack student difficulties, but rather we will present an overview of conceptual assessment for upper-division courses.

II. UPPER-DIVISION CONCEPTUAL ASSESSMENTS AT CU

Over the past 8 years, the PER group at CU has developed four conceptual assessments for the upper-division level: the Colorado Classical Mechanics/Mathematical Methods Instrument (CCMI), the Colorado Upper-division Electrostatics Diagnostic (CUE), the Colorado UppeR-division ElectrodyNamics Test (CURrENT), and the Quantum Mechanics Assessment Tool (QMAT). The development of these instruments was motivated, in part, by a need for a research-based and validated measure of the success of our course transformation efforts [6, 10, 13, 25] with respect to learning goals developed for each course. These learning goals were developed in conjunction with a broad cross-section of CU physics faculty as part of the Science Education Initiative's model for course transformation [26]. The goals represent a consensus of what these faculty want students to be able to do after completing our upper-division courses [25]. Several examples of these consensus learning goals are given below (see Ref. [27] for the full list of learning goals for each course).

Math/physics connection: Students should be able to translate a physical description of an [upper-division] problem to a mathematical equation necessary to solve it.

Communication: Students should be able to justify and explain their thinking and/or approach to a problem or physical situation.

Problem-solving techniques: Students should be able to choose and apply the problem-solving technique that is appropriate to a particular problem.

These course-scale learning goals are tightly linked to the physics content; however, they also emphasize

more meta-level outcomes related to the problem-solving strategies and habits of mind characteristic of professional physicists. In particular, we were interested in designing our assessments to target specific learning goals that were not typically assessed by traditional exams (e.g., conceptual understanding, justifying your reasoning, etc.). Although they were developed locally, these goals are not specific to the courses taught at CU, and feedback from external faculty suggest that these learning goals are valued and relevant more broadly in the U.S. physics community.

Each assessment was designed to target topics in one of the canonical core upper-division content areas (e.g., electrostatics, quantum mechanics, etc.); however, they are not designed to provide a comprehensive assessment of all material. The goal was instead to focus on a smaller subset of the material in order to provide a litmus test for student achievement with respect to our learning goals. A brief overview of each of our four named assessments is given in Table I along with information on several other assessment instruments that target the same core content areas. Comparisons of the development and validation of these instruments will be discussed later.

A. Content Coverage

1. Electricity and Magnetism

For the first half of junior electricity and magnetism, three potential assessment instruments are (see Table I): the CUE, the Symmetry and Gauss's Law Conceptual Evaluation (SGLCE), and the Electromagnetics Concept Inventory - Fields (EMCI - Fields). The SGLCE was designed to assess introductory physics students' ideas about symmetry and Gauss's Law [18], but preliminary testing with upper-division undergraduates and graduate students suggest that this instrument is challenging even for more advanced students. However, as an introductory assessment, the SGLCE does not include any of the more advanced electrostatics topics (e.g., solutions to Laplace's Equation). The EMCI - Fields was designed to target electrostatics, magnetostatics, and time-varying electromagnetic fields for junior engineering courses [8]. The content coverage of the CUE is similar to that of the EMCI, but the CUE does not include time-dependence, as this is typically not included in a first semester electricity and magnetism course in physics departments [9].

For the second half of junior electricity and magnetism, two assessments are available: the CURrENT, and the Electromagnetics Concept Inventory - Waves (EMCI - Waves). The EMCI - Waves focuses exclusively on the propagation and generation of electromagnetic waves, with a strong emphasis on engineering applications (e.g., waveguides, transmission lines, etc.) [8]. The CURrENT picks up where the CUE leaves off with time-variation, electromagnetic waves, and Maxwell's Equations [22]. Neither of these instruments includes relativistic electrodynamics. Note that Notaros *et al.* [8] have also crafted a 25 question combined EMCI that would be appropriate for a single semester electromagnetism course and covers a sampling of topics from both the Waves and Fields versions of the assessment.

2. Quantum Mechanics

A relatively large number of assessments have been developed for quantum mechanics including (see Table I): the QMAT, the Quantum Mechanics Survey (QMS), the Quantum Mechanics Visualization Instrument (QMVI), and the Quantum Mechanics Conceptual Survey (QMCS). Of these, only the QMAT and QMS were specifically developed for the upper-division level, and both target measurement, solutions to the Schrödinger equation in one dimension, and time-evolution from a wavefunctions-first perspective [11, 13]. The QMVI was designed to provide a longitudinal measure of students' understanding from introductory up through graduate quantum mechanics [12] with a specific emphasis on visualization. The longitudinal focus of the QMVI means that it includes a number of topics not typically covered until graduate quantum mechanics [19]. Lastly, the QMCS was developed to target sophomore-level, introductory quantum mechanics (i.e., modern physics). While the developers suggest that the QMCS may be particularly valuable as a pre-post measure in more advanced courses, they also note that many faculty see the QMCS as too basic for the upper-division level [19]. This is also why, though the QMCS was developed at CU, it is not included as one of CU's upper-division assessments.

3. Classical Mechanics and Thermodynamics

There are a number of assessments that target mechanics at the introductory level (e.g., [1, 2]; however, we are aware of only one published instrument for mechanics beyond the introductory level (i.e., sophomore-level classical mechanics), the CCMI. The CCMI was developed to target mechanics up to (but not including) the Lagrangian and Hamiltonian formulations, and also includes gravitation [7]. Additionally, while there are several engineering focused thermodynamics inventories available [14–16], we are not aware of any published, physics-centric thermodynamic instruments.

B. Development

The development of each of CUs four upper-division assessments followed the same basic process (See Fig. 1). The first draft of each assessment was generated in faculty working groups facilitated by PER specialists/postdocs, who then further developed and refined

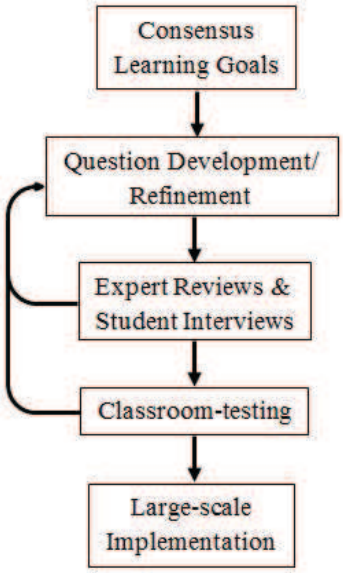


FIG. 1. A schematic of the design process used to develop CU's upper-division conceptual assessments.

the instruments. Initial question development was focused on addressing course-scale learning goals [27] articulated through collaborative discussions with CU physics faculty who had taught each course [25] (see examples at the beginning of Sec. II). These learning goals guided all stages of the assessment development including content coverage and format. For example, these consensus learning goals motivated one of the more unique aspects of CU's assessments: the free-response format. Because our learning goals emphasized students' ability to synthesize, generate, and justify their responses, we determined that an open-ended format would be more valued by faculty. Early drafts were also informed by known student difficulties identified through informal observations (e.g., in-lecture discussions or in small group work) and (when available) research on students' understanding.

When developing upper-division assessments, one of the key challenges is in creating level-appropriate questions that can be answered within a reasonably short time frame. For example, we quickly found that students are easily slowed by complicated calculations or tasks that required remembering exact formulae. To avoid excessive time spent on calculations, we used a number of different strategies such as asking students only to provide and justify a solution method rather than having them actually work through a problem. Another strategy that proved effective was asking students only to provide the sign of a certain quantity or whether it was zero or nonzero, rather than asking them to determine the value of that quantity.

At times, differences in the targeted content had an impact on the types of questions included in each assessment. For example, a significant component of electrostatics is the development of multiple problem-solving techniques all aimed at calculating the same physical quantity (i.e., the potential). To capture this, a large fraction of the questions on the CUE deal with selecting the appropriate problem-solving method. Alternatively, electrodynamics has far less ambiguity in solution method and thus the CURrENT does not include this type of question. Additionally, recognizing the physical meaning of abstract mathematical quantities (e.g., the wavefunction) is a particularly critical skill in quantum mechanics. Thus, more of the questions on the QMAT targeted students' interpretation of various quantities. In all cases, the advanced and specialized nature of these assessments makes it inappropriate for the full instruments to be used as pretest measures. Instead, pretests consist of a small subset of the questions on the full instrument which can be answered using ideas and techniques covered in earlier courses.

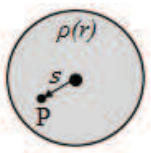
After a preliminary draft was completed, each assessment was reviewed by multiple experts in either physics content or assessment. Expert reviews establish the content validity of the assessments by ensuring that: (1) the physics content was accurate and clearly expressed, and (2) this content was valued by experts and perceived as appropriate for the upper-division level. The assessments were revised and refined based on this feedback. For example, early drafts of the assessments were often too long, and expert feedback was vital to shortening the instruments by identifying and eliminating questions that were least reflective of the goals of the course.

Revised drafts were then given to a small number (5-15) of volunteer undergraduates in an interview setting. Student interviews establish the face validity of the assessments by confirming that students were interpreting the prompts and figures as we intended. Interviews were conducted with individual students in a think-aloud format where interviewees were asked to articulate their reasoning as they worked through the problems on the assessment. When necessary, the assessments were modified to ensure that a student's responses reflected their knowledge of the physics content and not their understanding (or lack thereof) of the question.

Following expert reviews and student interviews, the assessments were tested as in-class post-tests in at least one semester of the associated course. Student scores during the classroom testing phase were analyzed to identify questions that were too difficult or too easy, or that did not discriminate between high and low achieving students (see Sec. II D). These items were either removed or modified, and the new version was re-tested with students in interviews and in-class implementations. Classroom testing is also critical to ensuring that the majority of students can complete the assessment within a 50 minute period. For example, early tests of the CURrENT and CCMI indicated that the instruments were too long and several questions/subparts were removed as a result. The final version of each assessment was a result of iterative refinement based on expert reviews, student interviews, and student performance in classroom tests (see Fig. 1).

Give a brief outline of the EASIEST method you would use to solve the problem.
DO NOT SOLVE the problem, we just want to know:
(1) The general strategy (half credit) and
(2) Why you chose that method (half credit)

Q7. *A solid non-conducting sphere, centered on the origin, with a non-uniform charge density that depends on the distance from the origin, $\rho(r) = \rho_o e^{-r^2/a^2}$ where a is a constant. Find E (or V) at point P.*



Q7 Rubric

Answer (3 pts)	Correct answer is Gauss's Law +1 point for saying direct integration
Explanation (2 pts)	Full credit requires some explanation of why (not just how) Gauss's Law is used. This would include some mention of the Gaussian surface used or the symmetry (such as charge distribution depends only on "r", or E field is radial). **+1** point if the correct Gaussian surface is drawn **+0-1** point for explaining how to solve by Gauss's Law If answer "direct integration" must give explanation of how they would solve this integral. **0.5** for a poor explanation of how they would go about it (e.g., writing down Coulomb's Law)

FIG. 2. An example question from the CUE along with the associated scoring rubric. This style of complex rubric was used for the CUE and QMAT and graders must undergo specific training to use these rubrics to produce reliable scores.

The available literature on other upper-division assessments (see lower half of Table I) suggests they were developed using a similar iterative design cycle. One notable difference is the central role that our explicitly-articulated learning goals played in the design process. These more meta-level goals guided us towards developing questions that not only targeted content knowledge, but also assessed reasoning and problem solving skills (e.g., Fig. 2). Alternatively, literature on the development of other assessments focuses on content coverage, typically determined through textbook reviews and faculty surveys. Specific questions are often developed to target known student difficulties; however, alignment of the questions and overall instrument with explicitly-articulated learning goals is not typically discussed for other assessments.

C. Scoring

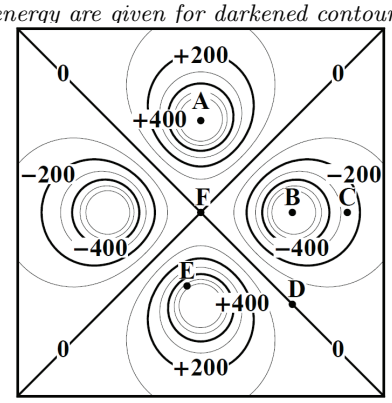
During the classroom testing phase of developing these assessments, it is necessary to develop scoring rubrics. For open-ended conceptual assessments this process is particularly challenging as any rubric must allow multiple graders to produce valid and reliable scores. We have utilized two distinct styles of grading rubrics with our conceptual assessments (discussed below).

The CUE and QMAT were the first of the CU assessments to be developed, and both assessments are characterized by fairly open-ended questions (e.g., Fig. 2). These open-ended questions have the potential to elicit a large variety of student responses: however, creating clear, reliable rubrics for such questions requires complex and nuanced grading schemes that reflect subtle differences in students' responses. To create these rubrics, common student ideas on each question were identified from student responses during classroom testing, and the developers agreed on scores for these common answers. A detailed grading scheme describing specific point allocations for a variety of student responses was developed to reflect these consensus scores. An example of this style of rubric is given in Fig. 2.

Early tests of inter-rater reliability for the complex grading rubrics for the CUE and QMAT (see Fig. 2) showed that some amount of training was necessary for new graders to produce consistent scores. For the case of the CUE, training involves a new grader independently scoring 10-15 example CUE exams and comparing their results to established scores. Conservative measures of the inter-rater reliability of the final version of the CUE scoring rubric indicated a substantial degree of inter-rater reliability when defining agreement as no more than a 10% difference between scores on each question by different graders [9]. While effective at ensuring reliable scores between different graders, the training process is time consuming (roughly 5 hrs) and thus often represents a significant barrier to faculty adoption.

When work on the CUE began, we were not aware of any existing examples of such a complex validated rubric for a free-response conceptual assessment that could be used reliably by independent graders. While the work with the CUE ultimately demonstrated that creating this type of rubric was possible, its development required dedicated work by a PER post-doc and multiple iterations of reliability testing. Efforts to create an equally reliable grading rubric for the QMAT were less successful in part because the QMAT intentionally provides multiple opportunities for even more broadly open-ended responses (e.g. "describe what happens to the real and imaginary parts as time goes by, with words and pictures", "list important qualitative features", "give an example of a state which..."). While these questions elicited rich and informative student responses, they made creation of unambiguous grading criteria more difficult. Reliability testing for the QMAT rubric was not completed before the post-doc responsible for the QMAT left at the end of the initial

Q3) *Below is a plot of the potential energy, in Joules, of a particle free to move on a 2-d plane. Values of the potential energy are given for darkened contours.*



(a) For which of these points (A–F) is the particle in stable equilibrium?
(b) Please explain how you decided on the above answer.
(c) Rank the magnitude of the gradient at the above points, from largest to smallest. (If some points have gradients with equal magnitude, please make that clear in your answer.)

Part A (1 pt)

Full credit, 1	Correct	Only (B)
No credit, 0	Incorrect	Any other responses

Part B (1 pt)

Full credit, 1	Correct	Reasonable answer to *correct response only*, for example: - Lowest potential energy, valley/well analogy, stable against small pushes
No credit, 0	Incorrect	Any other responses

Part C (1 pt)

Full credit, 1	Correct	Either E>C>D>A=B=F or E>C>D=A=B=F **(it is unclear if there's a small non-zero gradient at D)** Give credit even if signs are missing (e.g., E,C,D,(A,B,F))
Half credit, 0.5	Incomplete answer	Correct but missing 1 location (e.g., E>C>A=B=F)
No credit, 0	Incorrect	Any other responses

FIG. 3. An example question from the CCMI along with the associated scoring rubric. This style of simple rubric was used for the CCMI and CURrENT and graders need minimal training to produce reliable scores.

2-year appointment, at which point, further development was put on hold. Recent efforts have shifted the nature of the instrument by redesigning it in a multiple choice format (discussed briefly below).

Informed by the difficulties encountered in the development and use of the complex scoring rubrics for the CUE and QMAT, questions on the CCMI and CURrENT were intentionally designed to support less ambiguity in scoring [7]. This necessitated a shift away from the more open-ended questions that characterize the CUE and QMAT. Instead, the CCMI and CURrENT questions (e.g., Fig. 3) were designed to elicit a more constrained set of student responses while still capturing aspects of student reasoning. The scoring rubrics for the CCMI and CURrENT emphasize identifying correct elements and have fewer opportunities for partial credit than in the rubrics for the CUE and QMAT. An example of this style of rubric is given in Fig. 3.

These rubrics were developed and validated using the same process as the rubrics for the CUE and QMAT; however, the more 'all-or-nothing' focus makes the CCMI and CURrENT rubrics considerably simpler [22, 28]. Typically, these grading schemes do not award points based on intermediate steps or for partially correct or incomplete responses. Minimal training is required to achieve a high degree of inter-rater reliability using these rubrics [22, 28]. One potential trade-off of the simpler CCMI rubric is that it is not necessary to include as many examples of common incorrect responses that can help an instructor recognize or anticipate common student difficulties. To counteract this, we have begun creating a second 'difficulties' rubric for the CCMI [28]. This rubric is not designed to provide numerical scores, but instead to describe common student difficulties with each of the questions and present examples of what these difficulties look like.

Motivated in part by pressure from external institutions to simplify the scoring process further, we have recently begun exploring the viability of various multiple-choice versions of these assessments. To date we have developed multiple-response and multiple-choice versions of the CUE and QMAT. Distractors for both of these instruments were constructed from common responses to the free-response versions. The new version of the CUE utilizes a novel 'select all' format, which we are calling coupled multiple-response (CMR) [29]. The CMR CUE was specifically designed to match the original free-response CUE as closely as possible and uses nearly identical questions and prompts. The new adaptation of the QMAT, called the Quantum Mechanics Concept Assessment (QMCA) [30, 31], includes 31-items and was developed by author H.S. The QMCA was initially developed as a multiple-choice version of the 14-item QMAT; however, further refinement resulted in the removal of some QMAT questions and the addition of several new items.

D. Validation

Once reliable scoring rubrics were developed and sufficient data collected during classroom testing, we generated test statistics to establish the validity and reliability of our new assessment instruments. Two common perspectives on test development are Classical Test Theory (CTT) [32] and Item Response Theory (IRT) [33]. The

Assessment	Years of Active Work	Status of Instrument	Validation Statistics?
CCMI	3	Near final	Yes [7]
CUE	5	Finalized	Yes [9]
CURrENT	3	Near final	Yes [22]
QMAT	2	Archived	No
CMR CUE	2	Near final	Yes [29]
QMCA	2	Finalized	Yes [31]

TABLE II. Status of the development and validation of CU's upper-division assessments. The bottom two assessments are newly developed multiple-response and multiple-choice adaptations of the CUE and QMAT.

majority of conceptual assessments in physics, both at the introductory and upper-division level have been validated using CTT, while only a small number have been developed or analyzed using IRT [34–37]. One significant drawback of CTT is that all test statistics are population dependent. As a consequence, there is no guarantee that test statistics calculated for one student population (e.g., physics students at a community college) will hold for another population (e.g., physics students at a university). For additional discussion of the limitations of CTT, see Ref. [38].

IRT was specifically designed to address the shortcomings of CTT and all item and student parameters are independent of both population and test form [33]. However, there are several significant drawbacks to IRT as a potential tool to develop upper-division physics assessments. Even the most simple dichotomous IRT models require large N (>100) to produce reliable estimates of item and student parameters [33, 39]. This number increases for more complex models that, for example, include item discrimination parameters, or for instruments with polytomous scoring [39]. The small class sizes typical of upper-division physics would necessitate classroom testing at multiple institutions, possibly over multiple semesters, to collect this volume of data. Due in large part to the logistical barriers to IRT, the development and validation of our upper-division assessments was guided by CTT.

CTT posits several characteristics of a high quality assessment and a number of test statistics that quantify how well an instrument matches these characteristics. For polytomously scored assessments, these statistics include [32]: item difficulty as measured by the average score on each individual item, item discrimination as measured by Pearson Correlation Coefficients of item scores with the rest of the test, internal consistency as measured by Cronbach's Alpha [40], and whole test discrimination as measured by Ferguson's Delta [3]. For dichotomously scored assessments, several of the test statistics used are slightly different (see Ref. [32] for an overview).

Because work on each of the CU upper-division assessments began at different times, each is currently at a slightly different stage of development and validation (see Table II). The CUE is the oldest of the assessments and has had nearly 5 years of continuous work including development and data collection. The CUE is available in its final form with full validation statistics [9]. Development of the CCMI and CURrENT began roughly 3 years ago and are both in the final stages of classroom testing. These instruments will only undergo minor revision before final publication. All preliminary test statistics indicate that both assessments are valid and reliable [7, 22]. Development of the QMAT began shortly after the CUE and continued for roughly 2 years. However, development of the QMAT was put on hold before classroom testing was completed, and validation statistics were never published for this instrument. Work on the multiple-response CUE and the QMCA (the multiple-choice adaptation of the QMAT) began roughly 2 years ago. The multiple-response CUE is in the final stages of validation [29] and the QMCA is available in its final form with full validation statistics. [31].

As an example of test validation using CTT, we summarize here the validation statistics for those assessments listed in Table II (other than the QMAT); published statistics for other upper-division assessments (Table I) also tend to fall within the same ranges. Overall student performance across courses and institutions is between 45-55% for all of our instruments. These averages, while low compared to traditional course exams, allow for considerable discrimination between high and low performers. Consistent with this, all of our instruments have Ferguson's Delta values of $\delta \geq 0.97$, where anything above 0.9 is considered acceptable [3]. Additionally, all items on these assessments have item-test correlation coefficients above the standard cutoff ($r \geq 0.2$ [3]), demonstrating a satisfactory degree of item discrimination. In terms of internal consistency, all of our assessments have Cronbach's Alpha values of $\alpha \geq 0.75$ with the exception of the CURrENT which has $\alpha = 0.69$ when treating numbered questions as items (N=6) and $\alpha = 0.72$ when treating numbered sub-parts as items (N=15). While α for the CURrENT is closer to the standard threshold ($\alpha \geq 0.7$), it has also been shown that having fewer test items tends to drive Cronbach's Alpha downward [40]. The CURrENT, with only 6 questions or 15 sub-parts, is most susceptible to this tendency. Thus, all of CU's upper-division assessments provide valid and reliable measures of student learning for the tested population of students.

III. USES OF CU'S UPPER-DIVISION ASSESSMENTS

Once developed, student performance on these assessments can be used for a variety of different purposes by researchers, administrators, and individual instructors.

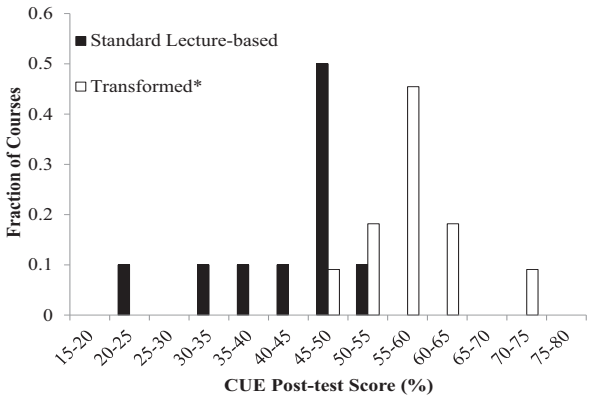


FIG. 4. Histogram of average course score on the CUE across 7 institutions demonstrating improved performance for courses using CU's transformed materials (11 courses, 9 CU; 329 students, 312 CU) relative to courses using only traditional lecture (10 courses, 5 CU; 303 students, 191 CU).
* All transformed courses used some or all of CU's transformed electrostatics materials

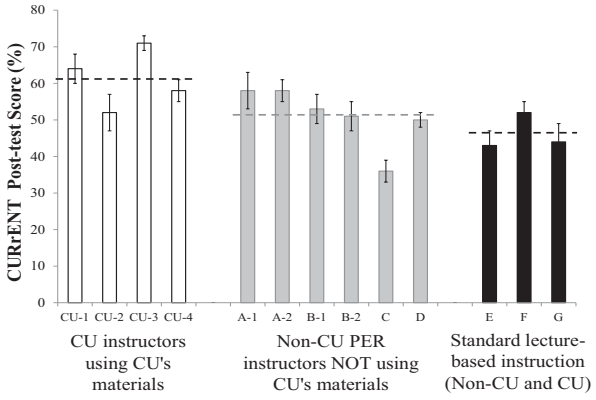


FIG. 5. Histogram of average course score on the CURrENT across 6 institutions demonstrating preliminary indications of improved performance for CU's transformed (N=124) and PER (N=172) courses relative to courses using only traditional lecture (N=50). Overall averages by course are marked by the dashed lines.

A. As a measure of student learning

As with conceptual assessments at the introductory level, the most common motivation for using upper-division assessments is as a standardized measure of student performance that can be compared across time, courses, instructors, institutions, and pedagogies. Indeed, one of the primary motivators for the development of our instruments was a need to assess the effectiveness of our upper-division course transformation efforts relative to other instructional practices [26].

For example, data on average scores on the CUE across 21 courses and 7 institutions demonstrate that transformed electrostatics courses score significantly higher on the CUE post-test (see Fig. 4). Using students as data points, transformed courses averaged 56.6 ± 1% and traditional courses scored 45.7 ± 1%. Treating courses as data points, these averages shift to 58.0 ± 2% and 42.3 ± 3% respectively.

While we have considerably less data available from the CURrENT than the CUE, scores from 13 courses at 6 institutions also show preliminary indications that transformed curricular materials improve student learning as measured by the CURrENT (see Fig. 5). Treating the courses as data points, CU transformed courses average 61 ± 4% and courses taught by PER instructors but not using CU's transformed course materials average 51 ± 3%. This represents consistent improvement when compared with an average of 46 ± 3% from courses taught using only traditional lecture. However, the standard-lecture based sample in these data is small and more data collection will be necessary to more robustly establish the impact of our course transformations on student learning.

While a discussion of the effectiveness of CU's transformed curricular materials is not the goal of this paper, it is worth noting a potential concern that we are 'teaching to the test.' Given that both our instruments and our transformed materials were designed to address our explicit learning goals, it is perhaps not surprising that students using our materials score higher on the assessments. However, faculty at CU and elsewhere, including those using standard lecture-based instruction, have consistently indicated that they agree that our learning goals represent what they want students to know, and that the assessments target a subset of these learning goals. Our assessments reflect the concepts and goals that upper-division instructors value and provide information about how well their students understand those concepts and achieve those goals.

B. For administrative purposes

Student performance on these instruments can also be used for administrative purposes. For example, comparative data on student learning using different pedagogies have been a necessary (though not sufficient) condition in supporting faculty at CU and elsewhere in their efforts to incorporate interactive engagement techniques into their upper-division courses, and to sustaining the use of our transformed curricular materials at CU [41]. It has also become common at CU for instructors to include their students' performance on these instruments as part of their tenure and promotion cases as evidence of reflective teaching practices. Additionally, the CU Physics Department has presented these assessments to deans and potential donors as evidence of improvements to undergraduate STEM education at CU. We have also been told informally that scores have been included in annual departmental reports at other institutions.

C. To investigate student difficulties

In addition to using scores as comparative measures, our assessments have also been used to gain insight into the nature of common student difficulties. The free-response format provides a particularly rich data source for identifying and characterizing topics that students find particularly challenging. Examples of this from CU span a number of topics including: Gauss's Law [42], Ampere's Law [43], divergence of vector fields [44], Taylor series [7, 45], quantum energies/time development [13], and electric potential [46]. In some cases, investigation of difficulties identified by one of these assessments has inspired broader research efforts. For example, early classroom testing of the CCMI revealed that our students often struggled not only with *how* to use Taylor Series, but also knowing *when* they were appropriate. Subsequent investigation of this difficulty helped to inform the development of an analytical framework for student use of mathematical tools in physics that specifically attends to how students determine which tool is appropriate [45].

As an example of this insight into student difficulties, analysis of preliminary data from the QMAT showed that our students had significant difficulties with the relationship between the Hamiltonian and the time evolution of quantum states similar to those reported previously [47]. Roughly half of these students agreed with the statement that applying the Hamiltonian to an arbitrary state gives information on how the state will evolve in time (QMAT Q4); however, only a quarter could also justify why the statement was true [13]. Many students who disagreed with this statement focused on the lack of time dependence in the Hamiltonian itself. Similarly, only a third of our students both disagreed with the statement that a system in an eigenstate of an arbitrary operator would stay in that state until disturbed (QMAT Q12), and could also justify why it was incorrect. Data from the QMCA also suggest that concepts related to time evolution are particularly challenging, and that the difficulty may arise, in part, from an over-generalization of the unique properties of energy measurements to physical observables whose corresponding operators do not commute with the Hamiltonian.

In addition to providing insight into the nature of student difficulties, standardized conceptual assessments can also be used to determine and to compare the relative prevalence of these difficulties across institutions and pedagogies. The QMCA provides a striking example of this. Questions on the QMCA/QMAT can be grouped into five main concept frames: measurement, the time independent Schrödinger equation, wavefunctions and boundary conditions, time evolution, and probability/probability density. The overall patterns of students' QMCA scores on these five topics are strikingly similar across 10 institutions. The greatest variation appears in students' scores related to quantum measurement, whereas scores in the other four categories are practically the same [31]. This observation supports the existence of several wide-spread common student difficulties regardless of student population and type of institution.

There are additional examples of our assessments being used to compare the prevalence of student difficulties between institutions. Researchers at Oregon State University (OSU) have looked at responses to a subset of CUE questions from students at both OSU and CU to identify ways in which the two student populations differ in terms of both scores and prevalent difficulties [48]. In particular, they examined one question on the CUE that is most easily solved using superposition of either the electric field or potential. They found that at both institutions, students often did not identify superposition as the correct solution method, or explicitly referred to the superposition of charges instead of fields. However, they also found that students at OSU were less likely to use the term 'superposition', and were more likely to use the superposition of electric potential than students at CU. These differences likely reflect differences between the CU and OSU curriculum, as the OSU curriculum does not emphasize the term 'superposition' and presents the electric potential before the electric field [49].

IV. IMPLEMENTATION OF CU'S UPPER-DIVISION ASSESSMENTS

Since their development, all of our upper-division assessments have been administered at multiple universities in the US. We have encountered a number of barriers and challenges to consistent use of these assessments and have explored a number of strategies to minimize these barriers.

A. Barriers and Challenges

One barrier to large-scale implementation of these assessments is faculty/instructor resistance to the assessments themselves. As standardized assessment is not a normal part of upper-division physics instruction, some instructors are hesitant to give an assessment that could reflect poorly on their instruction or be used to set one faculty member up against others.

Faculty can also be discouraged by the logistical requirements of these assessments. Instructors must dedicate class time to give these assessments, typically a full 50 minute class period at the end of the semester. This can seem particularly onerous if the instructor is being asked to give the assessment by an outside source (e.g., a department chair or physics education researcher). Once given, the free-response format of our assessments also makes them challenging and time consuming to grade. Busy faculty are often unable to dedicate the necessary time to grade these assessments.

Instructors who administer these assessments can also experience some resistance from students. For example,

students often want to use the assessments as a study tool. However, because these assessments are difficult and time consuming to develop, keeping them secure is particularly important. For this reason, we actively discourage instructors from providing solutions for their students or allowing them to take the test with them. Students can find these restrictions frustrating, particularly if no opportunity is given for them to review and ask questions about the assessment.

Another challenge we have encountered both with the development and large-scale implementation of our upper-division assessments is the relative lack of consistency between the content coverage and pace of advanced physics courses compared to introductory courses. The exact content of the upper-division physics curriculum can vary significantly from institution to institution and even from instructor to instructor. It is also not unusual for instructors to feel more ownership of these advanced courses and thus there is a greater degree of customization of each course. This makes it difficult to create a one-size-fits-all assessment that accurately reflects the content coverage and emphasis of the majority of courses. While we believe our assessments are representative of broader courses in that they were designed to match canonical textbooks and consensus learning goals, some external institutions have argued that the instruments favor the particular content and teaching styles at CU [48].

B. Strategies and Solutions

We have implemented a number of strategies to minimize the barriers and challenges documented above. To minimize faculty resistance to the assessments themselves, we solicited faculty involvement early in the development process to ensure they have the opportunity to help shape the instruments so that they value student outcomes on these measures. To reduce some of the logistical barriers, we have consistently offered to help faculty with grading each of these assessments. As a more sustainable strategy, our newer assessments (the CCMI and CURrENT) were both explicitly designed to have simple grading rubrics that are fast and straightforward to use. This helps to minimize faculty concern about being able to grade these assessments. Even more recently, we have begun developing multiple-choice and multiple-response versions of these assessments that allow for fast and objective grading. To date, the CUE and QMAT have been converted into two different easily-gradable formats; detailed discussion of these new versions can be found in Refs. [29–31, 50].

We have also developed strategies to minimize student resistance. Framing the tests as valuable but low-stakes measures of students' understanding that can be used to help them prepare for the final exam can be effective at promoting student buy-in. When possible, we also provide individualized feedback for each student, which includes their overall score relative to the class average. Additionally, offering a few extra office hours the final week of classes where students can come discuss and review their exams (without taking them home) can also help to encourage students to see these instruments as useful preparation for the final.

Variable content coverage between courses is a more challenging barrier to address as it is in many ways a characteristic of upper-division physics instruction, rather than the assessments themselves. However, this issue was particularly important for the CCMI, as the classical mechanics course at CU is a joint math methods course as well. To address this, the CCMI includes two optional questions in addition to the 9 core questions. These two optional questions target several of the mathematical methods emphasized in the CU course but are not included in the score on the assessment because they are not representative of broader classical mechanics courses. The issue of variable content coverage was also addressed early in the development of the CURrENT during a summer working group in which faculty from external institutions participated in discussions concerning the appropriate scope for the instrument. This greater insight into what the E&M 2 course looked like at other institutions directly motivated several restrictions in the content coverage of the CURrENT.

V. SUMMARY & DISCUSSION

Over the past decade, an increasing number of standardized and validated conceptual assessments have been developed that specifically target physics content beyond the introductory level. Specific topic areas include classical mechanics, electricity and magnetism, quantum mechanics and thermodynamics. In this paper, we identified and compared many of these assessment instruments based on format, content coverage, and development. We then provided a detailed review of four instruments created at CU as an example of the development, validation, and uses of upper-division conceptual assessments in physics. We also discussed some of the barriers to implementing these assessments in the classroom as well as some strategies and solutions to overcoming these barriers.

Of the published assessment instruments discussed here, all were developed using a similar iterative design cycle involving initial development, expert reviews, student interviews, and preliminary classroom testing (Fig. 1). At CU, the initial development phase was heavily influenced by consensus learning goals that emphasized more meta-level outcomes related to 'thinking-like-a-physicist.' These meta-level goals were the primary motivation for the unique free-response format of CU's assessments. Initial development of other assessments (Table I) focused instead on achieving appropriate content coverage without explicit discussion of non-content related learning goals.

Consistent with the literature published at the introductory level, the majority of the upper-division conceptual assessments described here (Table I) were validated using Classical Test Theory, and test statistics are available for all but two (the EMCI and QMAT). In all cases, the majority of the statistics for a given instrument fell within accepted ranges, indicating that each offers a valid and reliable measure of student learning within the tested populations and contexts.

There are a variety of examples of the uses of these assessment tools: as comparative measures of student learning across instructors, institutions, and time; and as sources of insight into student difficulties. The latter use is particularly true for the four open-ended assessments from CU, as the free-response format allows for generation and identification of new student difficulties rather than primarily providing data on the prevalence of known student difficulties, as on a multiple-choice instrument. However, we have also encountered a number of barriers to both small and large-scale implementation of conceptual assessments in the upper-division including: faculty resistance, student resistance, and logistical constraints. In some cases, we have implemented strategies to reduce these barriers (e.g., creating simple grading rubrics and multiple-choice versions to simplify the grading process).

While many of the barriers to conceptual assessment at the upper-division level are also, at least to some extent, barriers at the introductory level, one issue that is particularly acute at the advanced undergraduate level is the issue of variable course coverage. Reduced consistency in content coverage and emphasis between instructors and institutions makes it challenging to create assessment instruments that are appropriate for a broad range of courses. This is reflected in, for example, the relatively large number of assessments available for advanced quantum mechanics, each with slightly different focus and scope. Barring a national standardization of the upper-division physics curriculum, which we see as unlikely and undesirable, one potential solution to this issue, requiring large-scale coordination of both the PER and broader physics communities, would be to create banks of questions that can be used by individual instructors to craft course-appropriate assessments. This strategy is similar to what has been done for large-scale testing in K-12 (i.e., SAT or ACT testing) and would require the use of Item Response Theory to validate all potential items.

Ongoing work with CU's upper-division assessments includes completing final classroom tests of the CCMI and CURrENT, as well as the CMR CUE and QMCA. Particular emphasis is being placed on expanding classroom testing beyond the developing institutions in order to more robustly establish the validity of these assessments for a broader spectrum of physics students. Future work may include leveraging these assessments as longitudinal measures of student learning, creating new assessments for additional topical areas (e.g., thermodyamics), and/or converting the CCMI and CURrENT to multiple-choice or multiple-response formats to further facilitate large-scale use.

The translation of CU's free-response assessments into multiple-choice/multiple-response versions was motivated entirely by a desire to increase the scalability and usability, and is not an indication that we see the free-response versions as obsolete. The logistical advantages of the multiple-choice formats come with significant trade-offs (e.g., reduced insight into details of student thinking and exclusion of unanticipated responses). Ultimately, which version of the assessment should be used in any given context is dependent on both the kind of information an instructor or researcher wants to capture (e.g., comparative scores vs. deeper insight into student reasoning) as well as the logistical constraints of the specific course/program (e.g., class size). Thus, there is value in having both formats available for use in different contexts.

At both the introductory and upper-division level students' scores on standardized conceptual assessments should be interpreted carefully. Performance on these assessments, while valuable for a number of reasons described above, does not represent a 'catch-all, end-all' measure of student learning outcomes. When interpreting scores, explicit attention should be paid to how well the goals of any individual course align with CU's course-scale goals which are the foundation for each assessment. Moreover, while our assessments were specifically designed to target meta-level learning goals like problem-solving skills, in practice it is often not possible to distinguish between, for example, a student who does not understand the content from one who simply cannot articulate their (correct) reasoning. Additionally, because of the need for specialized language and mathematical techniques, the pre- and post-test versions of our upper-division conceptual instruments are often distinct, making it less meaningful to report or interpret normalized learning gains as is standard at the introductory level.

Even with an explicit emphasis on CU's meta-level learning goals, our upper-division conceptual assessments are still heavily content focused. Yet, there are many skills and characteristics related to a student's development as a physicist that extend beyond content knowledge that are rarely, if ever, assessed directly. For example, the capacity for independent learning, the ability to read and write scientific publications, and the ability to work collaboratively are just a few characteristics of successful physicists that we ultimately want our physics majors to internalize. We argue that operationalizing and assessing these implicit goals represents an important outstanding issue for the PER community to consider. An open questions is, can we begin to craft assessments that more accurately reflect the full range of learning outcomes we value for our physics majors?

ACKNOWLEDGMENTS

Particular thanks to the PER@C group and the faculty and students who contributed to the development of these assessments. This work was funded by the Science Education Initiative, an NSF-CCLI Grant DUE-1023028 and an NSF Graduate Research Fellowship under Grant No. DGE 1144083.

[1] D. Hestenes, M. Wells, and G. Swackhamer, *Force concept inventory*, The Physics Teacher **30**, 141 (1992), URL `http://link.aip.org/link/?PTE/30/141/1`.

[2] R. K. Thornton and D. R. Sokoloff, *Assessing student learning of newtons laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula*, Am. J. Phys. **66** (1998).

[3] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, *Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment*, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006), URL `http://link.aps.org/doi/10.1103/PhysRevSTPER.2.010105`.

[4] http://www.ncsu.edu/per/TestInfo.html (2014).

[5] R. R. Hake, *Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses*, Am. J. Phys. **66**, 64 (1998), URL `http://link.aip.org/link/?AJP/66/64/1`.

[6] S. V. Chasteen, S. J. Pollock, R. E. Pepper, and K. K. Perkins, *Transforming the junior level: Outcomes from instruction and research in e&m*, Phys. Rev. ST Phys. Educ. Res. **8**, 020107 (2012).

[7] M. Caballero and S. Pollock, in *Physics Education Research Conference 2013* (Portland, OR, 2013), PER Conference, pp. 81–84.

[8] B. M. Notaros, in *Antennas and Propagation Society International Symposium, 2002. IEEE* (IEEE, 2002), vol. 1, pp. 684–687.

[9] S. V. Chasteen, R. E. Pepper, M. D. Caballero, S. J. Pollock, and K. K. Perkins, *Colorado upper-division electrostatics diagnostic: A conceptual assessment for the junior level*, Phys. Rev. ST Phys. Educ. Res. **8**, 020108 (2012).

[10] C. Baily, M. Dubson, and S. Pollock, in *Physics Education Research Conference 2012* (Philadelphia, PA, 2012), vol. 1513 of *PER Conference*, pp. 54–57.

[11] G. Zhu and C. Singh, *Surveying students' understanding of quantum mechanics in one spatial dimension*, Am. J. Phys. **80**, 252 (2012).

[12] E. Cataloglu and R. Robinett, *Testing the development of student conceptual and visualization understanding in quantum mechanics through the undergraduate career*, American Journal of Physics **70**, 238 (2002).

[13] S. Goldhaber, S. Pollock, M. Dubson, P. Beale, and K. Perkins, in *Physics Education Research Conference 2009* (Ann Arbor, Michigan, 2009), vol. 1179 of *PER Conference*, pp. 145–148.

[14] K. C. Midkiff, T. A. Litzinger, and D. Evans, in *Frontiers in Education Conference, 2001. 31st Annual* (IEEE, 2001), vol. 2, pp. F2A–F23.

[15] R. A. Streveler, R. L. Miller, A. I. Santiago-Román, M. A. Nelson, M. R. Geist, and B. M. Olds, *Rigorous methodology for concept inventory development: Using the'assessment triangle'to develop and test the thermal and transport science concept inventory (ttci)*, International Journal of Engineering Education **27**, 968 (2011).

[16] D. Evans, G. L. Gray, S. Krause, J. Martin, C. Midkiff, B. M. Notaros, M. Pavelich, D. Rancour, T. Reed-Rhoads, P. Steif, et al., in *Frontiers in Education, 2003. FIE 2003 33rd Annual* (IEEE, 2003), vol. 1, pp. T4G–1.

[17] T. Thoads and R. J. Roedel, in *Frontiers in Education Conference, 1999. FIE'99. 29th Annual* (IEEE, 1999), vol. 3, pp. 13C1–14.

[18] C. Singh, *Student understanding of symmetry and gausss law of electricity*, Am. J. Phys. **74** (2006).

[19] S. B. McKagan, K. K. Perkins, and C. E. Wieman, *Design and validation of the quantum mechanics conceptual survey*, Phys. Rev. ST Phys. Educ. Res. **6**, 020121 (2010), URL `http://link.aps.org/doi/10.1103/PhysRevSTPER.6.020121`.

[20] J. R. Taylor, *Classical mechanics* (University Science Books, 2005), ISBN 9781891389221, URL `http://books.google.com/books?id=P1kCtNr-pJsC`.

[21] D. J. Griffiths, *Introduction to electrodynamics* (Prentice Hall, 1999), ISBN 9780138053260, URL `http://books.google.com/books?id=M8XvAAAAMAAJ`.

[22] Q. Ryan, C. Astolfi, C. Baily, and S. Pollock, in *Physics Education Research Conference 2014* (Minneapolis, MN, 2014), vol. In press of *PER Conference*.

[23] D. J. Griffiths and E. G. Harris, *Introduction to quantum mechanics*, vol. 2 (Prentice Hall New Jersey, 1995).

[24] J. B. Marion and S. T. Thornton, *Classical dynamics of particles and systems* (Brooks/Cole Cengage Learning, 2003).

[25] R. Pepper, S. Chasteen, S. Pollock, and K. Perkins, in *Physics Education Research Conference 2011* (Omaha, Nebraska, 2011), vol. 1413 of *PER Conference*, pp. 291–294.

[26] S. V. Chasteen, B. R. Wilcox, M. D. Caballero, K. K. Perkins, S. J. Pollock, and C. Wieman, *Educational transformation in upper-division physics: The science education initiative model, outcomes, and lessons learned*, Phys. Rev. ST Phys. Educ. Res. **In review** (2014).

[27] http://per.colorado.edu/sei (2014).

[28] L. Doughty and M. D. Caballero, in *Physics Education Research Conference 2014* (Minneapolis, MN, 2014), vol. In press of *PER Conference*.

[29] B. R. Wilcox and S. J. Pollock, *Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics*, Phys. Rev. ST Phys. Educ. Res. **10**, 020124 (2014), URL `http://link.aps.org/doi/10.1103/PhysRevSTPER.10.020124`.

[30] H. Sadaghiani, J. Miller, S. Pollock, and D. Rehn, in *Physics Education Research Conference 2013* (Portland, OR, 2013), PER Conference, pp. 319–322.

[31] H. Sadaghiani and S. Pollock, *Quantum mechanics concept assessment: Development and validation study*, Phys. Rev. ST Phys. Educ. Res. **In press** (2015).

[32] P. Engelhardt, in *Getting Started in PER* (2009), vol. 2.

[33] L. Ding and R. Beichner, *Approaches to data analysis of multiple-choice questions*, Phys. Rev. ST Phys. Educ.

Res. **5**, 020103 (2009).

[34] L. Ding, *Seeking missing pieces in science concept assessments: Reevaluating the brief electricity and magnetism assessment through rasch analysis*, Phys. Rev. ST Phys. Educ. Res **10**, 010105 (2014), URL http://link.aps.org/doi/10.1103/PhysRevSTPER.10.010105.

[35] J. S. Aslanides and C. M. Savage, *Relativity concept inventory: Development, analysis, and results*, Phys. Rev. ST Phys. Educ. Res. **9**, 010118 (2013), URL http://link.aps.org/doi/10.1103/PhysRevSTPER.9.010118.

[36] J. A. Marshall, E. A. Hagedorn, and J. O'Connor, *Anatomy of a physics test: Validation of the physics items on the texas assessment of knowledge and skills*, Phys. Rev. ST Phys. Educ. Res. **5**, 010104 (2009), URL http://link.aps.org/doi/10.1103/PhysRevSTPER.5.010104.

[37] M. Planinic, L. Ivanjek, and A. Susac, *Rasch model based analysis of the force concept inventory*, Phys. Rev. ST Phys. Educ. Res. **6**, 010103 (2010), URL http://link.aps.org/doi/10.1103/PhysRevSTPER.6.010103.

[38] C. S. Wallace and J. M. Bailey, *Do concept inventories actually measure anything?*, Astronomy Education Review **9**, 010116 (2010).

[39] R. J. De Ayala, *Theory and practice of item response theory* (Guilford Publications, 2009).

[40] J. M. Cortina, *What is coefficient alpha? an examination of theory and applications.*, Journal of applied psychology **78**, 98 (1993).

[41] S. Chasteen, R. Pepper, S. Pollock, and K. Perkins, in *Physics Education Research Conference 2011* (Omaha, Nebraska, 2011), vol. 1413 of *PER Conference*, pp. 139–142.

[42] R. Pepper, S. Chasteen, S. Pollock, and K. Perkins, in *Physics Education Research Conference 2010* (Portland, Oregon, 2010), vol. 1289 of *PER Conference*, pp. 245–248.

[43] C. S. Wallace and S. V. Chasteen, *Upper-division students' difficulties with ampère's law*, Phys. Rev. ST Phys. Educ. Res. **6**, 020115 (2010), URL http://link.aps.org/doi/10.1103/PhysRevSTPER.6.020115.

[44] C. Baily and C. Astolfi, in *Physics Education Research Conference 2014* (Minneapolis, MN, 2014), vol. In press of *PER Conference*.

[45] B. R. Wilcox, M. D. Caballero, D. A. Rehn, and S. J. Pollock, *Analytic framework for students use of mathematics in upper-division physics*, Phys. Rev. ST Phys. Educ. Res. **9**, 020119 (2013), URL http://link.aps.org/doi/10.1103/PhysRevSTPER.9.020119.

[46] R. E. Pepper, S. V. Chasteen, S. J. Pollock, and K. K. Perkins, *Observations on student difficulties with mathematics in upper-division electricity and magnetism*, Phys. Rev. ST Phys. Educ. Res. **8**, 010111 (2012), URL http://link.aps.org/doi/10.1103/PhysRevSTPER.8.010111.

[47] E. Gire and C. Manogue, in *Physics Education Research Conference 2008* (Edmonton, Canada, 2008), vol. 1064 of *PER Conference*, pp. 115–118.

[48] J. Zwolak, M. B. Kustusch, and C. Manogue, in *Physics Education Research Conference 2013* (Portland, OR, 2013), PER Conference, pp. 385–388.

[49] http://www.physics.oregonstate.edu/portfolioswiki/courses (2014).

[50] B. Wilcox and S. Pollock, in *Physics Education Research Conference 2013* (Portland, OR, 2013), PER Conference, pp. 365–368.