Assessing learning outcomes in middle-division classical mechanics: The Colorado Classical Mechanics and Math Methods Instrument

Marcos D. Caballero,[1,2,*] Leanne Doughty,[3] Anna M. Turnbull,[1,4] Rachel E. Pepper,[5] and Steven J. Pollock[6]

[1]*Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA*
[2]*Department of Physics and Centre for Computing in Science Education, University of Oslo, N-0316 Oslo, Norway*
[3]*School of Education and Human Development, University of Colorado Denver, Denver, Colorado 80204, USA*
[4]*Lyman Briggs College, Michigan State University, East Lansing, Michigan 48824, USA*
[5]*Department of Physics, University of Puget Sound, Tacoma, Washington 98416, USA*
[6]*Department of Physics, University of Colorado Boulder, Boulder, Colorado 80309, USA*

(Received 13 June 2016; published 19 April 2017)

Reliable and validated assessments of introductory physics have been instrumental in driving curricular and pedagogical reforms that lead to improved student learning. As part of an effort to systematically improve our sophomore-level classical mechanics and math methods course (CM 1) at CU Boulder, we have developed a tool to assess student learning of CM 1 concepts in the upper division. The Colorado Classical Mechanics and Math Methods Instrument (CCMI) builds on faculty consensus learning goals and systematic observations of student difficulties. The result is a 9-question open-ended post test that probes student learning in the first half of a two-semester classical mechanics and math methods sequence. In this paper, we describe the design and development of this instrument, its validation, and measurements made in classes at CU Boulder and elsewhere.

DOI: 10.1103/PhysRevPhysEducRes.13.010118

I. INTRODUCTION

In recent years the physics education research (PER) community has placed a strong emphasis on improving student learning in upper-division courses for physics majors [1–4]. Many research studies have shown the wide variety in students' understanding of particular physics concepts and practices during and after instruction [5–15]. Systematic efforts to assess student understanding on a broader scale have been useful in facilitating this effort [16]. These systematic assessments of student understanding at the upper-division highlight common and persistent student difficulties that can both inform curricular and pedagogical innovations and help form the basis for research efforts. Furthermore, these measures of student performance provide an indicator of the effectiveness of different pedagogies and curricula and can be used by instructors and departments to improve course offerings over time.

In fact, over the last 40 years, the awareness created by assessments of student learning using conceptual inventories has helped to drive widespread transformation of introductory lecture courses [17–19]. The use of these conceptual inventories has also helped the physics community identify persistent difficulties and provided the means to compare learning outcomes between different pedagogical and curricular reforms as well as across many institutions and implementations [20–27].

Over the last decade, the Department of Physics at the University of Colorado Boulder (CU) has worked to transform their upper-division lecture courses to more student-centric instruction [28,29]. This transformation process has involved the development of faculty-consensus learning goals [30], the use and development of instructional materials [31,32], and research to identify student difficulties [5,11,12,33], which has informed refinements to both the aforementioned learning goals and instructional materials. In recent years, upper-level assessments in the areas of quantum mechanics [34] and electricity and magnetism [35,36] have been developed to, in part, understand the impact of these transformations on student understanding.

In this paper, we present the Colorado Classical Mechanics and Math Methods Instrument (CCMI) that is both grounded in the history of this work and opens a new space for upper-level physics assessments—middle-division classical mechanics and mathematical methods

*Corresponding author.
caballero@pa.msu.edu

(CM 1). The central goal of this paper is to present a valid and reliable assessment that can be used in a variety of classical mechanics implementations. While we have begun using this instrument to unpack specific student difficulties [12,37], we have not fully investigated the CCMI's utility in this area. The CCMI (Sec. II) consists mostly of open-ended questions that probe students' use of the sophisticated skills and practices outlined in faculty-consensus learning goals. In Sec. III, we present the development of the CCMI including the design of the questions and the measures that provide evidence of validity. We discuss the design and structure of the grading rubric as well as measures of reliability in Sec. IV. In Sec. V, we present statistical results from its implementation at CU and other institutions through the lens of classical test theory. Finally, in Sec. VI, we discuss implementation, measurement, and possible uses.

II. THE COLORADO CLASSICAL MECHANICS AND MATH METHODS INSTRUMENT

The Colorado Classical Mechanics and Math Methods Instrument is a 9-question (with a total of 22 parts) open-ended test that focuses on topics taught in the first half of a two-semester classical mechanics sequence. This first course concludes before a discussion of the calculus of variations; hence, the Lagrangian and Hamiltonian formulations of mechanics are absent from the test. The CCMI focuses on core skills and commonly encountered problems. Students solve a variety of problems such as determining the general solution to common differential equations (e.g., $\ddot{x} = -A^2x$) finding equilibria and sketching net forces on a potential energy contour map and decomposing vectors in Cartesian and plane-polar coordinates. We have designed the CCMI to be given in a standard 50 min lecture period. To accompany the longer post test, we have developed a short (15–20 min) pretest that contains a subset of three problems taken from the post test. Figure 1 shows a sample CCMI question that appears on both the pre- and the post test. Table I contains the full listing of questions on the CCMI. The complete CCMI and the accompanying support documents are available online [38].

In designing the CCMI, decisions were made about which topics to include, which learning goals to assess, and how to allot points for particular questions and parts. In the sections that follow (Secs. III–V), we articulate how and why those decisions were made, but suffice it to say the CCMI is limited in its scope and its ability to serve as an assessment of classical mechanics. However, these limitations are no less (or more) relevant to the other assessments that are widely used in introductory [18,19,39] and advanced [35] courses. Each assessment represents only a small slice of the course that is being assessed; this is particularly true of the Force Concept Inventory, which is used widely even in courses that span the spectrum of mechanics implementations [24,26] and there is still value in its use.

III. DEVELOPMENT AND CONTENT VALIDATION

The development of the CCMI followed the process established by Chasteen *et al.* [35], who was recently reviewed by Wilcox *et al.* in their paper describing the uses and development of upper-level physics assessments [16]. Broadly speaking, the process involves establishing and prioritizing assessable learning goals, crafting questions that are tested with students using think-aloud interviews [40], and validating questions based on student and faculty input.

A. Development history

At CU, CM 1 is a blended classical mechanics and mathematical methods course that forms the first half of a two-semester sequence in classical mechanics. In recent years, this course was transformed from lecture-based instruction to more active and student-centric instruction [29]. The early part of this transformation involved the

Learning goals evaluated: *Students should be able to*:
· choose appropriate area and volume elements to integrate over a given shape.
· translate the physical situation into an appropriate integral to calculate the gravitational force at a particular point away from some simple mass distribution.

Q9 Consider an infinitely thin cylindrical shell with non-uniform mass per unit area of $\sigma(\phi, z)$. The shell has height h and radius a, and is not enclosed at the top or bottom.

(a) What is the area, dA, of the small dark gray patch of the shell which has height dz and subtends an angle $d\phi$ as shown to the right?
(b) Write down (BUT DO NOT EVALUATE) an integral that would give you the MASS of the entire shell. Include the limits of integration.

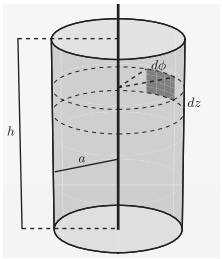

FIG. 1. Certain topic-scale learning goals are evaluated by the CCMI questions. The sample question appears on the CCMI pre- and post tests; vector calculus is a prerequisite for CM 1. This question constitutes 9% of the total post-test score.

TABLE I. Questions appearing on the CCMI. The full instrument is available online [38]. Questions 10 and 11 are both optional (*) and multiple-choice questions (MCQ).

Q no.	Pts	Short name	Description	Cohen's kappa	Pearson coeff.
Q1	3	Common differential equations	Context: 1D, linear, homogenous differential equations Tasks: Write the general solution to the differential equations $\ddot{x} = -A^2x$ (part a) and $dy/dt = By$ (part b). Describe a physical situtation where $d^2z/dt^2 = B$ is applicable (part c)	0.42	0.54
Q2	2	Taylor approximation	Context: Gravitation Task: Given $\Delta g = GM_E/(R-d)^2 - GM_E/R^2$, explain how you would determine an approximate formula for Δg if d is small.	0.63	0.25
Q3	5	Potential energy map	Context: Potential energy plot of a particle free to move on a 2D plane. Tasks: Where is the particle in stable equilibrium (parts a and b)? Rank the magnitude of the gradient at points on the plot (part c). Draw vectors that represent the force at those points (part d)	0.75	0.50
Q4	5	Damped harmonic oscillator	Context: Expression, $a_1\ddot{x} + a_2\dot{x} + a_3x = 0$, and a corresponding graph for the motion of mass on a spring. Tasks: Identify the units of a_1, a_3 (parts a and b), and resketch the solution if a_3 is smaller (part c). What would a $g(t)$ in lieu of "0" represent (part d)?	0.57	0.47
Q5	3	Simple harmonic oscillator	Context: Simple harmonic motion Tasks: In simple harmonic motion, what is restoring force proportional to (part a)? Write an expression for position as a function of time (part b). Draw potential energy as a function of position (part c).	0.88	0.59
Q6	6	Vector decomposition	Context: Ball sliding in the bottom of a sawed off sphere. Tasks: Draw the vectors $\hat{r}$ and $\hat{\theta}$ (part a). Express the velocity vector in the x-y and r-θ coordinate systems (part b). Check your answer (part c).	0.45	0.56
Q7	2	Resonance	Context: Mass on a frictionless spring attached to a driving force with a small amount of friction in the system. Tasks: Sketch the amplitude of the oscillation of the mass as a function of the driving frequency.	0.90	0.46
Q8	4	Writing a differential equation	Context: Particle confined to move between two objects that attract it. Tasks: Given description of the position and forces, write down a differential equation that describes the position of the particle as a function of time.	0.67	0.54
Q9	3	Writing an integral	Context: Infinitely thin cylindrical shell with non-uniform mass per unit area. Tasks: Write down the infinitesimal area, dA (part a). Write down an integral that would give you the mass of the entire shell (part b).	0.78	0.50
Q10*	2	Fourier series	Context: Graph of periodic function. Tasks: Which Fourier series could be the correct expansion for the given function? (MCQ)	…	…
Q11*	1	Laplace's equation	Context: Function of two variables. Tasks: How would you separate U to solve Laplace's equation in Cartesian coordinates, $\partial^2U/\partial x^2 + \partial^2U/\partial y^2 = 0$, using separation of variables? (MCQ)	…	…

development of consensus learning goals by a group of faculty. A series of faculty meetings were held to develop consensus course-scale learning goals and to articulate the topical content coverage of the course [30]. Overall 19 faculty (4 PER, 15 non-PER) participated in at least one of the 7 meetings, with an average of 9 faculty at each meeting [30]. Course-scale learning goals focus on how the student develops over the whole semester. For example, students in CM 1 are consistently exposed to the connection between math and physics. Relevant course-scale learning goals for this area include "Students should be able to translate a physical description of a sophomore-level classical mechanics problem to a mathematical equation necessary to solve it. Students should be able to explain the physical meaning of the formal and/or mathematical formulation of and/or solution to a sophomore-level physics problem. Students should be able to achieve physical insight through the mathematics of a problem."

After the development of course-scale learning goals, a set of specific, topic-scale learning goals were drafted. To develop these topic-scale learning goals, we utilized field notes collected during lectures, weekly homework help sessions, and faculty meetings. A further set of faculty meetings were held in which the topic-scale learning goals were agreed upon. In these meetings, several topic-scale learning goals were selected to be assessed on the CCMI as articulated in the learning goals [30].

These topic-scale learning goals combined content coverage that faculty had defined and the mathematical and problem-solving skills characteristic of upper-division coursework. For example, "Students should be able to use Newton's laws to translate a given physical situation into a differential equation" and "Students should be able to project a given vector into components in multiple coordinate systems, and determine which coordinate system is most appropriate for a given problem." These course-scale and topical-scale learning goals are available in the Supplemental Material [41].

These topic-scale (measurable) learning goals formed the basis for the development of the CCMI. While these learning goals were developed by CU faculty, and are specific to CM 1, we believe that many are applicable to the mathematical methods and classical mechanics courses offered at other universities because (i) the goals were developed by a mix of traditional and PER physics faculty with many more traditional faculty contributing, and (ii) the topical coverage was drawn from the first five chapters of a standard classical mechanics textbook [42]. Moreover, faculty from five other institutions have given the CCMI in their courses and were interviewed to obtain feedback on the learning goals assessed by the CCMI as well as the CCMI itself. Since these interviews, faculty at more than 20 institutions have given the CCMI to their students. These interviews led to changes in coverage and scoring of the CCMI.

As the topic-scale learning goals were developed, CU faculty discussed which ones were most fundamental to student learning, that is, which goals (when met by students) would be taken as evidence of learning in CM 1, which goals formed the basis for future learning (e.g., in future physics coursework), and, thus, which goals should be assessed on a standardized instrument. When a topic-scale goal was deemed by faculty to be assessment worthy, a draft assessment item was written by the postdoctoral researcher facilitating these conversations with input from faculty. Sixteen open-ended questions were initially written. Some of these questions were adapted from exam or clicker questions written by CU faculty in previous semesters. All questions were informed by observed student difficulties [32]. The early versions of these questions were entirely open ended and were developed to draw out student ideas about the particular concepts and skills that would be assessed on the final instrument.

The earliest version of the CCMI contained 16 questions—more than could be answered in a single 50 min class period. Thus, the CCMI was split into two 11-question versions with some number of overlapping questions; each version was given to half the students in the class. One benefit of developing this instrument at a large, research-intensive university is a large population of students taking CM 1—in some semesters, more than 100 students have been enrolled in CM 1 at CU. Through a number of administrations of early versions of the questions, feedback from faculty and students, as well as timed testing, the CCMI was trimmed to an 11-question, open-ended assessment that could be administered in a 50 min period. Question were culled from the active list for any of several reasons including that they overlapped with other questions in terms of the primary learning goals they assessed, that they were not producing a good variation in student performance, or that they were clearly idiosyncratic to the implementation at CU. Following this internal development period, the CCMI was offered in a "beta" version to faculty teaching courses like CM 1 at other institutions. Administration of the CCMI at these other institutions provided additional feedback on the content coverage and scoring of the CCMI.

Interviews were conducted with these beta testers to learn more about their courses, their needs, and their view of the CCMI. These interviews were prompted by concerns about certain questions from the Colorado Upper-Division Electrostatics Assessment from colleagues using it at other institutions [43]. Prior to these interviews, faculty were given a copy of the CCMI and the accompanying learning goals (see Fig. 1) to review. The "CM 1" courses that our interviewees taught ranged from quite similar to CM 1 (e.g., a 2 semester sequence classical mechanics) to quite compressed compared to CM 1 (e.g., a 1 semester course on classical mechanics that surveys all common topics including Lagrangian Hamiltonian dynamics and the orbit

equation). While there was a substantial diversity among the topical coverage among the courses taught by these faculty, most agreed that 9 of the 11 questions were covered well enough in their courses to be included as part of the assessment. However, for most faculty, 2 questions, which deal with Fourier series and Laplace's equation, were covered superficially or not at all in their courses. As a result, the CCMI consists of 11 questions—9 core questions that count towards the overall score that can be compared across institutions, and 2 optional questions that may be used at institutions where such topics are taught.

B. Content validation of the CCMI

In designing the CCMI, we took the approach that an assessment of student learning should address the topics that traditional physics faculty value. This serves to validate the instrument in the sense that the questions being asked of students cover the topics in the way that faculty desire. Further, this process serves to generate buy-in to use the instrument. Second, the instrument needs to be interpretable by students, that is, students need to be able to interpret each question consistently and in the ways that instructors expect. Below, we detail how we established the validity of the CCMI through discussions with faculty and think-aloud interviews with students.

1. Expert validation

As the basis for the questions were the expert-developed learning goals [30], the instrument was grounded in the topics deemed essential by faculty. Draft questions were developed from these learning goals; some were inspired by existing course materials (clicker questions, exam questions, etc.) and others were crafted from scratch. Once a complete set of questions was drafted, faculty at CU and elsewhere were consulted individually to obtain their feedback on the instrument. The CCMI was sent to faculty before meeting with the postdoctoral researcher for a semistructured interview. The instrument and subsequent questions were positioned to the interviewed faculty in the following way:

> *"Does this question ask about the kinds of things you want students in your CM 1 class to learn?"*
> *"If a student in your CM 1 class correctly solved this question, would you say that student demonstrated an understanding of this topic? Why?"*
> *"If a student performed well on this instrument, would you expect them to have performed well in your CM 1 class? Why?"*

As faculty spoke on these different topics, follow-up questions were asked to elucidate the meaning behind faculty's answers. In all, nine faculty (4 at CU and 5 elsewhere; all non-PER faculty) were interviewed for between 50 and 90 min. Individual faculty input was often aligned with each other, likely because these interviews took place following the discussion of learning goals. But, there were conflicting comments at times. For example, most faculty interviewed agreed that the instrument should focus on conceptual aspects of CM 1 while one or two faculty desired students to perform calculations on certain questions (e.g., Taylor series) because they believed that to be the only way to judge student learning on those particular topics. Where there was disagreement between interviewed faculty, we sided with the majority. Hence, the CCMI focuses on more conceptual aspects of CM 1. Faculty input was critical to deciding which questions to prune from the 16-question version of the CCMI. Discussion with faculty lead to ranking questions by "most important for students to understand after completing CM 1."

2. Student validation

Questions on the CCMI were further shaped by conducting think-aloud interviews with students while they solved the CCMI. The interviews served two purposes: (i) to ensure that the wording of the questions was clear for students (i.e., that students would interpret questions as asked), and (ii) to collect student reasoning for correct and incorrect answers in order to help shape the grading rubric, which had not yet been fully designed. Eight CU students who had recently completed CM 1 earning grades ranging from A to C were interviewed (in two cohorts) for 60–90 min as they solved the CCMI. Following a think-aloud protocol [44], students narrated their thoughts while solving each question. The interviewer took notes identifying how each student read each question, what reasoning was brought to bear on each question, and where there were points of confusion or issues of clarity. If at any time the student struggled to answer a question, the interviewer suggested they make their best attempt given what they understand. Following a student's completion of the CCMI, the interviewer followed-up question by question with the student about their reading of the questions and their reasoning through their answer. The interviewer also discussed the correct solution to each question with most students as they were often interested in how well they performed. These interviews and notes were analyzed for salient themes that addressed issues of clarity and student reasoning after the first cohort of students completed the interviews.

The most prevalent issues were addressed by the first round of editing by the development team. For example, no interviewed student in the first cohort knew how to answer the Taylor series question. Discussion with the interviewees indicated a mismatch between our intent (i.e., explaining the importance of the small parameter in the expansion) and their experience (i.e., not ever being asked to think explicitly about the small parameter). Questions were redrafted before conducting interviews with the next cohort

of students. In this second set of interviews, the majority of questions elicited the expected responses and underlying reasoning. Those questions that still had some issues were positioned to the students as

> *"In this question, we are trying to get you to work with this idea (e.g., Taylor series) in this way (e.g., identifying the small parameter in the expansion), how would you know to do that?"*

Students' responses to questions of this kind provided the final edits to the previously problematic questions.

IV. GRADING THE CCMI

Scoring student responses to an assessment reliably undergird the value of the assessment to students and faculty. The rubric for the CCMI was informed by the lessons learned from our development of other upper-level assessments [16] as well as experienced and anticipated challenges for faculty users. To navigate this rubric, we have found it useful to define questions (numbered items that assess topic scale learning goals), parts (lettered subquestions that assess some narrower aspect of the learning goal or scaffold the question), and points (the numeric score allotted to parts, questions, and the whole assessment).

A. Rubric rationale

With the validity of the CCMI established, we turned to scoring student responses to provide an indication of student performance. It is important for independent assessments of student learning, such as the CCMI, that independent graders achieve consistent results. Therefore, the scoring rubric needs to capture the variety of student responses and indicate how each response is scored. There are a number of possible approaches to supporting graders in this work. For example, in the electrostatics context, the Colorado Upper-Division Electrostatics Diagnostic (CUE) took the approach of training graders to attend to both students' final answers and the nuances of student responses [35]. As such, graders were not only providing a consistent score for student work, but also attending to the details of student difficulties. The training was not intended to be overly prohibitive (∼8 h), but there was not much interest outside the PER community to learn to grade the CUE. Thus, researchers at CU have continued to provide a grading service to the physics community. In order to facilitate grading and promote wider use of the CUE, Wilcox *et al.* developed a multiple-choice version of the CUE that can be delivered online [36]. This work leveraged the large body of CUE responses collected over the years to develop an updated set of questions and a logical grading model that has proven quite successful—reproducing similar results to the original CUE.

While the CCMI has recently had significant interest from faculty at a number of institutions, the initial work to develop the rubric could not leverage a large body of responses. Thus, we decided to separate the two roles of the assessment into two rubrics: (i) a grading rubric that allows for scoring student work from a "high performance" perspective [45,46], and (ii) a difficulties rubric that helps to uncover the prevalence of student difficulties in CM 1 [37]. The grading rubric is intended for faculty with no training to grade their students' responses consistently and have confidence that their scoring is meaningful. The difficulties rubric that we are developing is intended for researchers (or faculty) who intend to dig deeper into student reasoning and requires some amount of training. In this paper, we discuss only the grading rubric as our central goal is to present the CCMI and not yet the difficulties that it might uncover.

The approach to grading the CCMI that we have used focuses on the students' final answer and points are taken away for errors in that answer. Graders need only to attend to one part of a student's answer and can score based on more salient features of the student's final response. This grading approach is taken by both the CCMI and the Colorado UppeR-division ElectrodyNamics Test (CURrENT) [46]. The development of the grading rubric was grounded in patterns in students' responses to CCMI questions, which formed the basis for categories in the grading rubric [12,37]. The grading rubric describes how points are deducted for different errors, providing examples where necessary (it does not list all the possibilities). The illustrative errors are those commonly seen in students' answers. The allocation of points on each question and the partial credit awarded for some responses are based on faculty "rankings" of the relative importance of the learning goal each question assesses and the relative importance of the errors. While the rubric was in development, faculty were asked to rank questions based on the prompt, "If a student answers this question correctly, they have demonstrated that they have learned an important topic in CM 1." While there was variation among the responses, there were some questions that faculty clearly deemed more important and thus they were allotted a higher fraction of the overall points. The rubric that we present here is the final version based on a few scoring variations that included grading student work in a more nuanced fashion (as the CUE does) and a correct answer only fashion (as the FCI does). The more nuanced rubric was deemed too much of a barrier for faculty to adopt themselves and the correct answer rubric produced little variation in student performance. As an example, the rubric used to grade the question shown in Fig. 1 appears as Fig. 2. Large scale ($N > 500$) use of the rubric on students' responses at CU and other institutions resulted in changes to the wording of the rubric and the addition of new examples. While a different design for scoring student work might be used, in our design, we considered asking

Question 9 (Writing an Integral) - Total points: 3:

Part A: 1 point

Full credit (1)	Correct	$ad\phi dz$
Partial credit (0.75)	Wrong length scale	$rd\phi dz$
No credit (0)	Incorrect	No credit for any other responses

Part B: 2 points

Full credit (2)	Correct	Correct integral form: • Integrals must be over $d\phi$ and dz; dA must agree with part A • Limits on the $d\phi$ integral must be $[0, 2\pi]$ • Limits on the dz integral must be either $[0, h]$ or $[-h/2, h/2]$ • Mass density ($\sigma(\phi, z)$ or σ) must appear in the integral; Substituting ρ is OK
Partial credit (1.75) Minus 0.5 points (1 max)	"Symmetric" integral Incorrect 2D integral	Integral over $d\phi$ is from $[0, \pi]$ but is multiplied by 2 Problematic limits or kernel: • Limits on the integral are incorrect (e.g., $[-h, h]$ or $[0, \pi]$) • Kernel of the integral is incorrect (e.g., missing $\sigma(\phi, z)$ or a)
No credit (0)	Incorrect	No credit for any other responses (e.g., 1D or 3D integral)

FIG. 2. Grading rubric for the question appearing in Fig. 1. The format focuses the grader's attention on the final response provided by the student. The grading rubric was not designed to elucidate details of student difficulties, but rather to capture the common final responses provided by students and score them accordingly.

traditional faculty to grade the assessment and how we might achieve consistent results across untrained graders. Our grading procedure does produce consistent results.

B. Intergrader reliability

Through a series of analyses, we established the reliability of our grading rubric. Our work follows the analysis conducted by Chasteen *et al.* to establish a reliable grading rubric for the CUE [35], but also makes use of an untrained grader who was asked to use the completed rubric to score student responses. The two graders (one untrained) scored responses from 100 students to all 11 questions on the CCMI. The resulting scores assigned to individual responses were compared as well as the overall score for a given students' CCMI. The resulting analysis demonstrated that an untrained grader can score students' responses to the CCMI reliably using the grading rubric.

First, the average overall difference in CCMI scores assigned to students between a trained and untrained grader is less than 5% (3.5% $\pm$ 2.7%) of the total points. Figure 3 demonstrates that the graders agreed on a total score within 10% for all but two students, and for the majority of students (79%) the graders were below 5% disagreement.

While this difference on total score is an intuitive measure of agreement, a more rigorous test of agreement is one that excludes the possibility that graders agree by chance. Cohen's kappa characterizes the agreement between two (or more) graders [47,48] while attempting to remove chance agreement from the calculation. However, there are concerns with using Cohen's kappa, where partial credit is awarded, where the scales between items differ, and where the total number of possible scores is high. Furthermore, it is worth noting that Cohen's kappa is a relatively conservative measure of agreement [49]. Chasteen *et al.* provide additional discussion of the issues

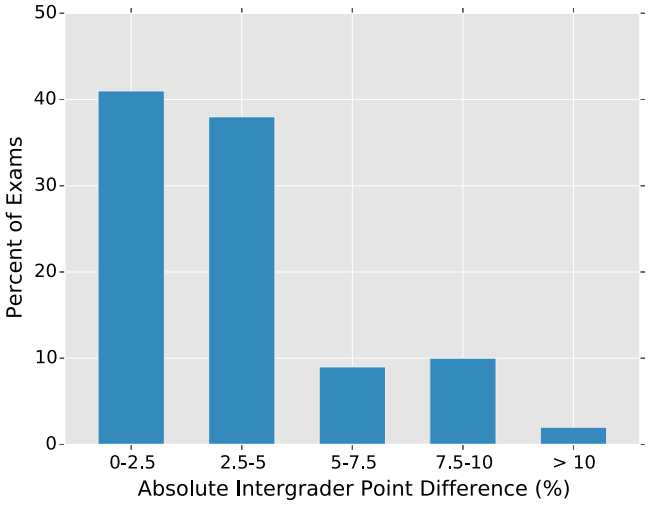

FIG. 3. The absolute difference in CCMI scores assigned by a trained and untrained grader is presented. The average difference on total CCMI score between the trained and untrained grader is 3.5% $\pm$2.7%. The graders agreed to 5% on overall score for 79% of the exams.

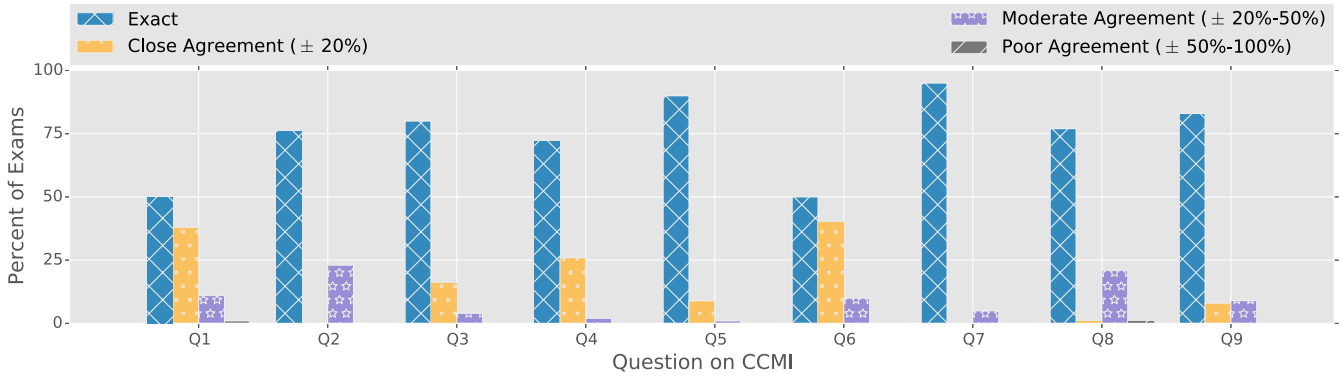

FIG. 4. For each question, the score differences between a trained and untrained grader are shown. These percentage agreement between graders was binned as exact (same score, blue), close (within ±20%, orange), moderate (within ±20%–50%, purple), and poor (more than ±50%, gray) agreement. For all questions, exact agreement was the most prevalent form of agreement (≥50% of exams for all questions).

associated with Cohen's kappa for determining reliability on an upper-level assessment [35].

As expected (and previously observed by Chasteen *et al.*), agreement across all possible point distributions is low ($\kappa = 0.23$). It is unlikely that each grader will agree on the overall points awarded to each student, but it is fairly likely for graders to agree within a few points (Fig. 3). Like the CUE, calcuating Cohen's kappa for scores binned into two-point intervals (∼5%) provides evidence of moderate agreement ($\kappa = 0.47$). When binned into four-point intervals (∼10%), we obtain evidence of substantial agreement ($\kappa = 0.64$). Hence, within differences of 5%, we find reasonable agreement between trained and untrained graders.

While this overall agreement is reasonable, it may be that specific questions may contribute to these differences more than others. That is, it might be that some combination of a specific question and the rubric describing how to score that question is unreliable. By determining Cohen's kappa for each question on the CCMI (see Table I), we find some evidence that questions 1 (common differential questions) and 6 (vector decomposition) might be contributing to these overall discrepancies. This message is further bolstered by the evidence provided in Fig. 4 where we have shown the percent agreement between a trained and untrained grader on each question. Here, we define "exact" to be the same score for the students' response while "close," "moderate," and "poor" represent agreement to ±20%, ±20%–50%, and ±50%–100%, respectively.

These analyses provide evidence of a robust and reliable grading rubric, but we acknowledge that due to our design there is some information lost, particularly if the CCMI rubric is compared to the CUE rubric. Because of the focus on final answers, information about student difficulties that would be captured in a more detailed rubric is lost. We are developing a separate difficulties rubric to address this issue [37]. However, what is gained (speed, accuracy, and adoption) by this approach to grading should not be understated.

V. STATISTICAL VALIDATION OF THE CCMI

To establish an assessment as a valid and reliable instrument, further analysis into specific properties of the test must be conducted. Recently, this kind of work has shifted towards using response modeling techniques such as Item Response Theory (IRT) [50,51]. While IRT is quite robust and used widely, the body of data needed to use it reliably is more than we have been able to collect. Over the last several years, we have collected data from five CM 1 courses at CU ($N = 244$) and from eleven similar courses at nine other institutions ($N = 218$). There are simply not as many users or students taking upper-level assessments of this type. Hence, we make use of Classical Test Theory [52]—following the analysis conducted by Chasteen *et al.* [35] and Wilcox *et al.* [36].

A. Internal consistency

An assessment of student learning should be internally consistent. If the assessment aligns with the goals of instruction, students who perform well on a single question should perform well on other questions. Essentially, each question should provide consistent information about a student performance (on the average). It is typical to use Cronbach's alpha to investigate internal consistency—estimating the reliability of scores or the "unidimensionality" of the assessment. We determined Cronbach's alpha treating each part of a question as an item because the total number of test items on CCMI is small. We find that the CCMI is a highly internally reliable assessment ($\alpha = 0.83$). The acceptable range for α is above 0.7, with greater than 0.8 being *good* [53].

B. Criterion validity

If aligned well with the learning goals for a course, we expect that an independent assessment of student learning (i.e., the CCMI) should correlate with other assessments of student learning (e.g., final exams). Students' exams are the

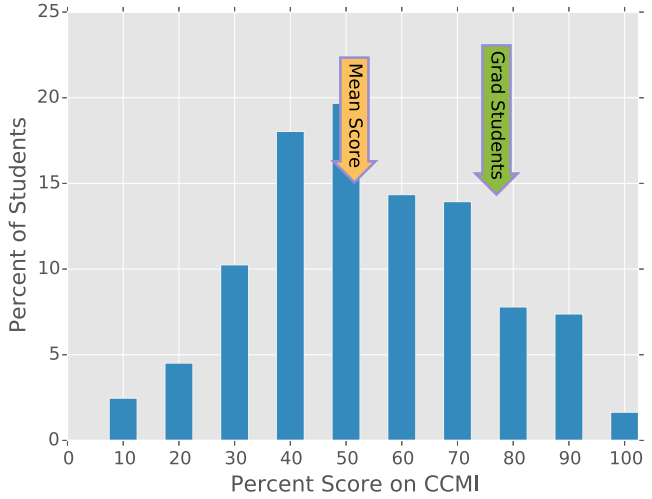

FIG. 5. A histogram of student post-test scores on the CCMI is presented ($N = 462$; CU and other institutions). The average score for students taking the CCMI is indicated (orange arrow): $49.0\% \pm 1.0\%$ as well as the performance by first-year physics graduate students at CU Boulder (green arrow): $74.5\% \pm 3.4\%$ ($N = 5$).

most similar measure to the CCMI. Like exams, the CCMI is completed individually in timed and controlled environments. But, unlike exams, it does not affect students' grades. Each class at CU took three exams: two regular hour exams and a final. The averages of those three exams were normalized [using a z score, $z = (x - \bar{x}/\sigma)$] to allow comparisons of different instructors. CCMI post-test scores were strongly correlated with these z-scored exam averages ($r = 0.71$); a linear model can thus account for 50% of the variance in exam scores associated with CCMI scores. This result provides evidence that the CCMI is well aligned with the learning goals of the course as assessed by final exams. Similarly high correlations were observed on the CUE [35].

C. Item-test correlation

We expect that the performance on individual items to correlate well with the overall score on the instrument (minus the item being tested). This correlation is expected from the premise that the whole assessment is a measure of a large construct—knowledge of CM 1 concepts—and that construct has underlying features—e.g., Taylor series—that will be more or less learned in similar amounts. We use Pearson's r (linear correlation) to determine how well each item connects to the rest of the CCMI (see Table I for values of r for each item). We find that all items are above the established threshold ($r \sim 0.2$) for item-test correlation. However, we note that question 2 (Taylor series) correlates much less than the rest of the items do with the whole instrument.

D. Discrimination

An assessment of student learning should be able to separate students who demonstrated low understanding from those who demonstrated high understanding. Ferguson's delta is the typical measure of discrimination used for assessments of this type—it provides a measure of how broadly the scores are distributed across the possible scores. In calculating Ferguson's delta, we used the total number of points on the assessment rather than the number of items as each question is worth a different number of points. We find that the CCMI has excellent discrimination on a per-point basis ($\delta = 0.99$). A test with $\delta > 0.9$ is considered to have good discrimination [54].

E. Additional analyses of item difficulty

While Ferguson's delta is a typical measure, it might not be an intuitive measure of discrimination. In Fig. 5, we provide the histogram of student performance on the CCMI, which shows the mean score to be $49.0\% \pm 1.0\%$ ($N = 462$). Indeed, the CCMI is a difficult assessment. First-year graduate students at CU earned an average score of $74.5\% \pm 3.4\%$ ($N = 5$). In Fig. 6, we provide a visualization of the difficulty of each item. The mean and median score for each item are plotted along with the score that the top 30% of scores lies at or above.

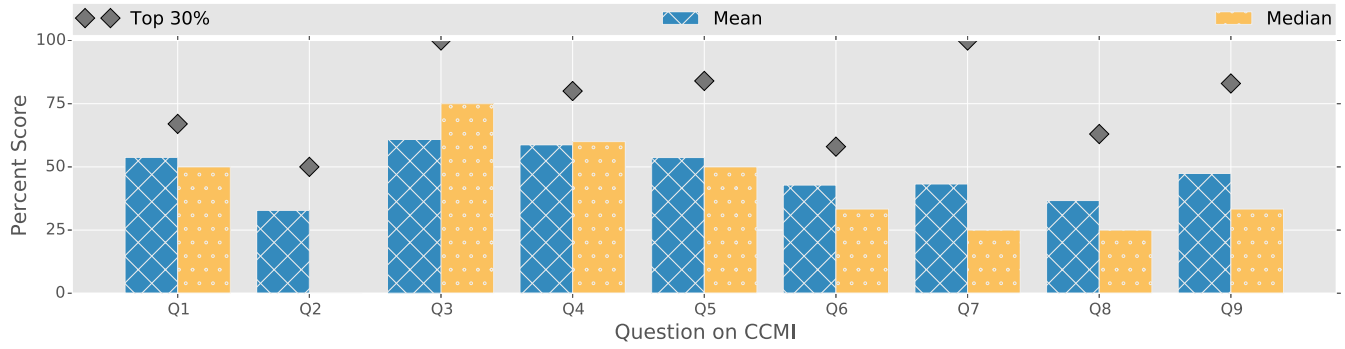

FIG. 6. The mean—blue bar—and median—orange bar—of student performance on each question is provided. The median score for question 2 is zero. In addition, we show the score—gray square—the score that separates the top three deciles from the bottom seven deciles, that is, the score which the top 30 percent of scores lies above.

VI. DISCUSSION AND CONCLUSION

In summary, we have developed an assessment for classical mechanics and mathematical methods courses for which we have established validity and developed a reliable grading rubric. Scores on the CCMI correlate well with other measures of student understanding (i.e., in-class exams) and internal measures of validity, reliability, and discrimination are well within the acceptable scope for such an assessment. While it may appear that the instrument is quite specialized, use by and feedback from faculty at more than 20 institutions have shaped the assessment to cover a broad range of course offerings. Faculty teaching courses quite similar to CM 1 (e.g., a 2 semester sequence classical mechanics) as well as those courses that are quite compressed compared to CM 1 (e.g., a 1 semester course on classical mechanics that surveys all common topics including Lagrangian Hamiltonian dynamics and the orbit equation) have used the CCMI. That feedback from faculty informed both the design and use of rubric developed to analyze student work on the CCMI. The design of the rubric for the CCMI separated the two traditional roles of assessment in physics education—(i) gaining a reliable understanding of student performance on specific topics, and (ii) identifying persistent student difficulties [16]. The former role was presented in this paper as the grading rubric, which demonstrated reliability even when used by an untrained grader. A rubric to address the latter is in development [37] and will be the subject of future work.

The CCMI was designed to serve a variety of purposes. Most simply, it is an independent measure of student understanding after instruction in a classical mechanics course. Student performance on specific topics as well as performance across the instrument can serve as a secondary and standardized measure of student understanding after a classical mechanics course. These measures can be used by faculty to improve different aspects of their instruction as they see fit. Most faculty who have used the CCMI have used it for this purpose. Faculty have reviewed their score reports to identify strengths and weaknesses in their instruction based on their interpretation of students'scores as well as to provide direct feedback to their students.

At a slightly higher scale, the CCMI may serve as a tool for departments looking to assess their physics program. It is becoming increasingly incumbent upon physics departments to demonstrate some form of independent assessment, and the CCMI (along with other standardized instruments) can help serve this purpose. Unlike course final exams, the CCMI is a standardized instrument, which invites comparison over time, between curricula, and across institutions. As such, student performance on the CCMI could be part of a more comprehensive departmental assessment.

From a cultural perspective, the CCMI offers opportunities for new (and seasoned) faculty to push on norms for teaching evaluation in their own tenure and promotion cases. Faculty teaching classical mechanics courses can demonstrate their commitment to quality instruction by including student performance on CCMI in their teaching portfolios. These kinds of independent assessments are critical to understanding how student learning is being affected by instruction beyond the typical collection of course syllabi and student responses to end-of-course evaluations.

While we have developed a valid and reliable assessment for classical mechanics that can serve a number of purposes, we have accepted certain limiting factors in our design. Given the constraints of administration (i.e., a 50 min lecture period), the content coverage of the CCMI is limited (Table I). Not every instructor will agree on which topics should appear on an assessment for classical mechanics—making it impossible to satisfy each instructor's needs. To address the issue of topical coverage, we drew from consensus learning goals [30] that were developed by traditional physics faculty. In designing the questions for the CCMI, we worked with these faculty to prioritize the learning goals and, thus, the topics that were evaluated on the CCMI. Furthermore, we collected feedback from instructors across the country to ensure that the CCMI meet most of their needs. It was in this work that two questions on the CCMI were designated optional as these topics were not covered to the degree they were at CU. In a sense, we have developed an assessment that serves as the "common denominator" for many implementations of classical mechanics.

A second limitation is our focus on students' final answers for the grading rubric, which underemphasizes the process by which the student obtained the answer, and, moreover, can make it difficult to judge the prevalence of specific student difficulties. The purpose of this answer-focused grading rubric was to streamline the process by which faculty can obtain information on student performance on the CCMI. For example, a significant challenge for the CUE has been to train new graders to reliably score student responses to the CUE, which informed our decision to simplify the process so that an untrained grader using the rubric could score student responses reliably and have confidence that they had done so (Figs. 3 and 4). Our current grading rubric has achieved this.

To deal with this limitation, we are developing a rubric that helps categorize difficulties that manifest on the CCMI [37]. This rubric is being informed by research into student understanding of classical mechanics [11,12]. However, it is worth noting that there is still much that can be learned from scoring the CCMI as we have done: the most prevalent incorrect answers are represented in the grading rubric as partially correct answers (Fig. 2). In fact, our research into student's approaches to vector decomposition [12] was informed by results from grading the vector decomposition problem on the CCMI. Hence, some information about the prevalence of certain kinds of student

difficulties are captured by the grading rubric. Wilcox *et al.* solved the problem of reliably scoring an independent assessment differently by adapting the CUE to a multiple-choice version with a logical scoring system that could be offered online or with scantrons [36]. This work benefited from the large body of student responses to the CUE collected over the years. Now that we have completed the development of the CCMI and collected a similarly large body of student responses, we are exploring the possibility that the CCMI might be adapted into a multiple-choice, machine-gradable format.

ACKNOWLEDGMENTS

We gratefully acknowledge the generous contributions of CU faculty, especially A. D. Marino, J. L. Bohn, K. P. McElroy, and collaborating faculty elsewhere. Particular thanks to the members of PER@C and PERL@MSU who have provided feedback and guidance on this work over the years. We also greatly appreciate the help of our student participants. This work was supported by University of Colorado's Science Education Initiative and Michigan State University's College of Natural Science.

[1] C. A. Manogue, P. J. Siemens, J. Tate, K. Browne, M. L. Niess, and A. J. Wolfer, Paradigms in physics: A new upper-division curriculum, Am. J. Phys. **69**, 978 (2001).

[2] C. Singh, Student understanding of quantum mechanics, Am. J. Phys. **69**, 885 (2001).

[3] 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).

[4] S. V. Chasteen, S. J. Pollock, R. E. Pepper, and K. K. Perkins, Thinking like a physicist: A multi-semester case study of junior-level electricity and magnetism, Am. J. Phys. **80**, 923 (2012).

[5] 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).

[6] C. S. Wallace and S. V. Chasteen, Upper-division students' difficulties with Ampère's law, Phys. Rev. ST Phys. Educ. Res. **6**, 020115 (2010).

[7] C. Singh, Student difficulties with quantum mechanics formalism, AIP Conf. Proc. **883**, 185 (2007).

[8] E. B. Pollock, J. Thompson, and D. Mountcastle, Student understanding of the physics and mathematics of process variables in pv diagrams, AIP Conf. Proc. **951**, 168 (2007).

[9] T. I. Smith, J. R. Thompson, and D. B. Mountcastle, Student Understanding of Taylor Series Expansions in Statistical Mechanics, Phys. Rev. ST Phys. Educ. Res. **9**, 020110 (2011).

[10] B. R. Wilcox and S. J. Pollock, *Proceedings of the 2014 Physics Education Research Conference, Minneapolis, MN* (AIP, New York, 2015), pp. 271–274.

[11] 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).

[12] A. Turnbull, L. Doughty, V. Sawtelle, and M. D. Caballero, *Proceedings of the 2015 Physics Education Research Conference, College Park, MD* (AIP, New York, 2015), pp. 335–338.

[13] C. A. Manogue, K. Browne, T. Dray, and B. Edwards, Why is Ampère's law so hard? A look at middle-division physics, Am. J. Phys. **74**, 344 (2006).

[14] C. Singh and E. Marshman, Review of student difficulties in upper-level quantum mechanics, Phys. Rev. ST Phys. Educ. Res. **11**, 020117 (2015).

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

[16] B. R. Wilcox, M. D. Caballero, C. Baily, H. Sadaghiani, S. V. Chasteen, Q. X. Ryan, and S. J. Pollock, Development and uses of upper-division conceptual assessments, Phys. Rev. ST Phys. Educ. Res. **11**, 020115 (2015).

[17] D. Hestenes, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[18] R. Thornton and D. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula, Am. J. Phys. **66**, 338 (1998).

[19] 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).

[20] L. C. McDermott and E. Redish, Resource letter: PER-1: Physics education research, Am. J. Phys. **67**, 755 (1999).

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

[22] D. E. Meltzer and R. K. Thornton, Resource Letter ALIP–1: Active-Learning Instruction in Physics, Am. J. Phys. **80**, 478 (2012).

[23] 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).

[24] M. A. Kohlmyer, M. D. Caballero, R. Catrambone, R. W. Chabay, L. Ding, M. P. Haugan, M. J. Marr, B. A. Sherwood, and M. F. Schatz, Tale of two curricula: The performance of 2000 students in introductory electromagnetism, Phys. Rev. ST Phys. Educ. Res. **5**, 020105 (2009).

[25] M. D. Caballero, E. F. Greco, E. R. Murray, K. R. Bujak, M. Jackson Marr, R. Catrambone, M. A. Kohlmyer, and M. F. Schatz, Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study, Am. J. Phys. **80**, 638 (2012).

[26] L. Ding and M. D. Caballero, Uncovering the hidden meaning of cross-curriculum comparison results on the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **10**, 020125 (2014).

[27] J. Von Korff, B. Archibeque, K. A. Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: a 50k-student study, arXiv:1603.00516.

[28] S. Chasteen, K. Perkins, P. Beale, S. Pollock, and C. Wieman, A Thoughtful Approach to Instruction: Course transformation for the rest of us, J. Coll. Sci. Teach. **40**, 24 (2011).

[29] S. V. Chasteen, B. Wilcox, M. D. Caballero, K. K. Perkins, S. J. Pollock, and C. E. Wieman, Educational transformation in upper-division physics: The science education initiative model, outcomes, and lessons learned, Phys. Rev. ST Phys. Educ. Res. **11**, 020110 (2015).

[30] R. E. Pepper, S. V. Chasteen, S. J. Pollock, and K. K. Perkins, Facilitating faculty conversations: Development of consensus learning goals, AIP Conf. Proc. **1413**, 291 (2012).

[31] B. S. Ambrose, Investigating student understanding in intermediate mechanics: Identifying the need for a tutorial approach to instruction, Am. J. Phys. **72**, 453 (2004).

[32] S. Pollock, R. Pepper, and A. D. Marino, Issues and progress in transforming a middle-division classical mechanics/math methods course, AIP Conf. Proc. **1413**, 303 (2012).

[33] M. D. Caballero, B. R. Wilcox, R. E. Pepper, and S. J. Pollock, *Proceedings of the 2012 Physics Education Research Conference, Philadelphia, PA* (AIP, New York, 2013).

[34] H. R. Sadaghiani and S. J. Pollock, Quantum mechanics concept assessment: Development and validation study, Phys. Rev. ST Phys. Educ. Res. **11**, 010110 (2015).

[35] 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).

[36] 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).

[37] L. Doughty and M. D. Caballero, *Proceedings of the 2014 Physics Education Research Conference,Minneapolis, MN* (AIP, New York, 2015), pp. 71–74.

[38] All referenced course materials are available at http://www.colorado.edu/physics/EducationIssues/ClassicalMechanics/. The CCMI can be accessed by verified educators on PhysPort https://www.physport.org/assessments/assessment.cfm?I=71&A=CCMI.

[39] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[40] C. Bowen, Think-Aloud Methods in Chemistry Education: Understanding Student Thinking, J. Chem. Educ. **71**, 184 (1994).

[41] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.13.010118 for course and topic scale learning goals that were used in the development of the CCMI.

[42] J. R. Taylor, *Classical Mechanics* (University Science Books, New York, 2005).

[43] J. P. Zwolak, M. B. Kustusch, and C. A. Manogue, Re-thinking the Rubric for Grading the CUE: The Superposition Principle, arXiv:1307.0902.

[44] M. E. Fonteyn, B. Kuipers, and S. J. Grobe, A description of think aloud method and protocol analysis, Qualitative health research **3**, 430 (1993).

[45] M. D. Caballero and S. J. Pollock, *Proceedings of the 2013 Physics Education Research Conference Portland, OR* (AIP, New York, 2014), pp. 81–84.

[46] C. Baily, M. Dubson, and S. J. Pollock, Research-Based Course Materials and Assessments for Upper-Division Electrodynamics (E&M II), AIP Conf. Proc. **1513**, 54 (2013).

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

[48] J. Cohen, Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit, Psychol. Bull. **70**, 213 (1968).

[49] R. L. Brennan and D. J. Prediger, Coefficient kappa: Some uses, misuses, and alternatives, Educ. Psychol. Meas. **41**, 687 (1981).

[50] F. M. Lord, *Applications of Item Response Theory to Practical Testing Problems* (Routledge, London, 1980).

[51] S. E. Embretson and S. P. Reise, *Item Response Theory* (Psychology Press, London, 2013).

[52] L. M. Crocker, J. Algina *et al.*, *Introduction to Classical and Modern Test Theory* (JSTOR, Orlando, 1986), Vol. 6277.

[53] L. J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika **16**, 297 (1951).

[54] P. Kline, *Handbook of Psychological Testing* (Routledge, London, 2013).