Assessing student written problem solutions: A problem-solving rubric with application to introductory physics

Jennifer L. Docktor,[1,2,*] Jay Dornfeld,[1,3] Evan Frodermann,[1] Kenneth Heller,[1] Leonardo Hsu,[4] Koblar Alan Jackson,[5] Andrew Mason,[1,6] Qing X. Ryan,[1] and Jie Yang[1]

[1]*School of Physics and Astronomy, University of Minnesota–Twin Cities, Minneapolis, Minnesota 55455, USA*
[2]*Department of Physics, University of Wisconsin–La Crosse, La Crosse, Wisconsin 54601, USA*
[3]*Robbinsdale Armstrong High School, Plymouth, Minnesota 55441, USA*
[4]*Department of Postsecondary Teaching and Learning, University of Minnesota–Twin Cities, Minneapolis, Minnesota 55455, USA*
[5]*Physics Department, Central Michigan University, Mount Pleasant, Michigan 48859, USA*
[6]*Department of Physics and Astronomy, University of Central Arkansas, Conway, Arkansas 72035, USA*

(Received 26 May 2015; published 11 May 2016)

Problem solving is a complex process valuable in everyday life and crucial for learning in the STEM fields. To support the development of problem-solving skills it is important for researchers and curriculum developers to have practical tools that can measure the difference between novice and expert problem-solving performance in authentic classroom work. It is also useful if such tools can be employed by instructors to guide their pedagogy. We describe the design, development, and testing of a simple rubric to assess written solutions to problems given in undergraduate introductory physics courses. In particular, we present evidence for the validity, reliability, and utility of the instrument. The rubric identifies five general problem-solving processes and defines the criteria to attain a score in each: organizing problem information into a Useful Description, selecting appropriate principles (Physics Approach), applying those principles to the specific conditions in the problem (Specific Application of Physics), using Mathematical Procedures appropriately, and displaying evidence of an organized reasoning pattern (Logical Progression).

DOI: 10.1103/PhysRevPhysEducRes.12.010130

I. INTRODUCTION

Solving complex, nonroutine problems has been recognized as essential for all citizens, especially those in technical fields [1,2]. Because physics courses serve as a foundation for disciplines both inside and outside of STEM, they are an ideal venue for teaching problem solving.

It is common for physics instructors to use problem solving as a mechanism for teaching physics content and assessing if that content has been learned. In addition, instructors often have a goal of helping students construct physics knowledge [3] or introducing students to the culture of science [4], both of which also require a generalized problem-solving framework. Indeed, physics education researchers have developed different types of problems [5–8], curricular add-ons [9], intelligent tutoring systems [10], and entire curricula [11,12] designed to scaffold the learning of problem solving. However, one significant barrier to the development and dissemination of such materials has been the lack of a validated, standard, general, and easy-to-use assessment tool for measuring the quality of students' problem solving.

In this paper, we describe the development and testing of a simple rubric for assessing students' problem solving based on their written problem solutions. In Sec. II, we first discuss the research on problem solving and its assessment on which our work is based. Section III describes the design of the rubric, including its categories and scoring criteria and how it is related to previous work. Section IV presents evidence for the validity and reliability of the instrument, using a framework from the *Standards for Educational and Psychological Testing* [13] and Messick [14]. Finally, we discuss some applications of the problem-solving rubric in Sec. V. More detail on the development of the rubric and many of the studies described in this paper can be found in Ref. [15].

II. BACKGROUND

A. Research on expert and novice problem solving

The process of problem solving occurs when a person needs to resolve a situation where they do not know a specific set of actions they can use to reach that resolution [16]. For example, Martinez [17] summarizes the cognitive science definition of problem solving as "the process of moving

*Corresponding author.
jdocktor@uwlax.edu

toward a goal when the path to that goal is uncertain." Thus, whether a particular task represents a problem depends on the solver's experience and perception of the task [17]. It is important to distinguish problems from "exercises," which are situations for which a person, in principle, knows the actions necessary to reach a solution. The distinction between a problem or an exercise has nothing to do with whether it is easy or hard, or whether it is time consuming to resolve. What may be a problem for one person may be an exercise for another [18,19]. The primary difference between the two is that a problem requires the solver to make decisions about linking their knowledge in novel ways.

Extensive research has found two major differences between expert and novice problem solvers: their knowledge organization and their problem-solving process [20–22]. Experts organize their knowledge in interconnected chunks, hierarchically grouped around a small number of fundamental principles [23,24] and have organized decision-making processes that help them choose relevant principles for solving a problem [25]. Many of these processes are so automated that they are not explicitly displayed or even recognized by the expert. In contrast, novices have fragmented knowledge, and their decision-making processes are often narrowly context related and formula driven.

An expertlike problem-solving process is distinguished by the initial performance of a qualitative analysis to constrain the problem and categorize it based on fundamental principles [26]. Experts then apply a principle or a small number of principles to the problem in an organized manner, using self-monitoring strategies to assess their progress, as well as their final solution [27]. In other words, an expert problem solution consists of many decisions guided by metacognition. Novices, on the other hand, typically focus on specific quantities in the problem and try to match those with mathematical procedures. While novices might perform a rudimentary qualitative analysis when required, they usually do not do so spontaneously [28].

Thus, in the domain of physics problem solving, the process of moving toward expertise involves the integration of students' conceptual knowledge with their problem-solving framework [29]. The organization of one's knowledge affects one's problem-solving process and vice versa. Indeed, improving students' generalized problem-solving skills results in an improvement in their conceptual knowledge [8,30,31].

B. Research on assessing problem solving

Research studies to assess students' physics problem solving have focused on students' problem-solving processes [26,32], metacognition [33], knowledge organization [24], strategic knowledge for choosing and applying useful principles to physics problems and ability to debug existing problem solutions [34], and problem-solving skills that seem to be independent of a student's physics content knowledge [35].

These studies have often relied on grades given by classroom instructors, percentage of problems answered correctly [36,37], or time spent solving a problem [38]. Although convenient and easy to collect with large numbers of students, these measures can give an inadequate description of a student's problem solving, since none of them directly assesses the characteristics that are known to distinguish experts from novices and the correlation between such indirect measures and problem-solving expertise is unknown.

It is especially important to distinguish the process of assessing problem solving from grading. When grading, instructors take into account many factors that are not necessarily indicative of expertlike problem solving such as grading time, fairness, motivating students, and the correctness of the end result or intermediate results [39]. It is common practice to heavily weight obtaining a correct answer, even when it is not obvious that the reasoning that led to the correct answer is either correct or complete. Indeed, expert problem solvers, when faced with a real problem, can be expected to generate an incorrect answer a significant fraction of the time [17]. Furthermore, grading credit is often given to encourage the use of particular artifacts modeled for students during the course. Even though these features, such as drawing a free-body diagram or solving equations algebraically before using numbers, may be beneficial to student learning, their existence is not necessarily indicative of expertlike thought. The use of these features to determine a grade can vary substantially depending on the problem and instructor [39]. In contrast, an assessment of the expertness of problem solving should focus exclusively on the nature and quality of the reasoning leading to the result.

One popular method to investigate problem solving is think-aloud interviews, in which students verbalize their thinking processes as they attempt a problem [40]. Such interviews are the gold standard of problem-solving assessment. Some practical difficulties with this method include the time involved to prepare and conduct them, the large amount of data generated from even a few interview transcriptions, the artificial setting in which they take place, and the complicated nature of the data analysis. Although interviews give valuable insight into student problem solving, they are difficult or impossible to use in a large-statistics study.

Rubrics are another research method for assessing complex behavior in authentic situations. Because rubrics can be used to quantify different dimensions of performance and include standards of attainment for each of those dimensions, they facilitate consistent scoring across assignments and raters [41,42]. There is a long history of trying to develop rubrics to evaluate problem solving in physics [11,27,43–50]. These rubrics tend to focus on similar features, such as the solver's choice of useful principles, the use of those principles, the use of representations and

mathematics, and evidence of an organized solution plan. Although valuable, these rubrics vary in their use of consistent criteria across dimensions and are often complex and difficult to use.

III. ASSESSMENT INSTRUMENT DESIGN

A. Design criteria

Our goal was to develop an instrument that would be practical to use for quantitatively assessing the problem solving of a large number of students in authentic situations. Based on the considerations discussed above, we decided to develop a rubric, the Minnesota Assessment of Problem Solving (MAPS), that could be applied relatively easily to normal classroom activities such as students' written problem solutions from homework or tests.

In general, a rubric is a scoring guide designed to rate the performance of a complex task such as reading comprehension, writing, or problem solving. It can be *holistic*, having a single dimension that makes an overall judgment about the quality of the work, or *analytic*, in which multiple dimensions of the work are assessed separately. In either case, each dimension is scored on multiple levels of proficiency based on a description characterizing the behavior corresponding to that level. Because complex tasks have many possible paths toward a resolution and the resolution is often context dependent, the descriptions of the levels require the rater to decide on the meaning of terms such as "major" vs "minor," and "most" vs "few," as well as others such as "unorganized," "difficult to follow," and "shows support." The use of such language is a reflection of the complexity of the work it is designed to assess, as well as of the diversity of learning contexts from which that work might be drawn [41,42]. For example, what is considered a minor error at the beginning of a high school physics class might be considered a major error at the end of a university physics class. Such flexible language distinguishes a "rubric" from a "checklist," which is a scoring tool that depends on the presence or absence of a particular artifact and does not necessarily require disciplinary or pedagogical expertise on the part of the rater. The full MAPS instrument including category descriptions and criteria for each of its levels is given in Table I.

Our design requirements were that the rubric should be as follows:

(i) *Easy to use.*—A rubric is easier to score and interpret if it minimizes the number of categories and the complexity of the scoring. Our rubric was constructed to span the majority of the space distinguishing novice from expert problem-solving practices using as few categories as possible. Metacognitive processes that are often not directly expressed in written work are not directly addressed by the rubric. For example, the details of planning a solution and the revision and refinement of that plan are difficult to assess from written work because students often do not write down those processes unless explicitly instructed to do so. Experts rarely write down their plan explicitly. However, because these processes affect the overall coherence and consistency of the solution, they can be implicitly assessed by a combination of rubric categories. Affective qualities such as motivation, interest, and beliefs about physics that are not usually evident from written work have been omitted. Finally, care has been paid to make the language used in the rubric as consistent as possible across its categories.

(ii) *Usable in authentic situations.*—The rubric focuses on written work and is applicable to solutions spanning the range of problem types and topics in typical physics classes. It is purposefully not tied to any particular pedagogy, problem-solving framework, or strategy. For example, it does not key on specific features, such as using any particular representation or solving for a symbolic solution first. Although such features may be important to an instructor or even most instructors, they do not necessarily distinguish between novice and expert problem solvers.

(iii) *Have evidence for its validity, reliability, and utility.*— In particular, the assessment rubric should quantify the major differences between novice and expert problem solving as defined in the research literature and agree with the expectations of physics instructors. We have tested the rubric extensively for its consistency across multiple raters, many different types of problems and solutions, and its usefulness for distinguishing between novice and expert solutions. These studies are described below.

It is important to stress the difference between grading and assessment of problem solving. The rubric is intended to be useful to researchers and curriculum designers by providing them with a tool to assess students' problem solving, without the difficulty of conducting large numbers of students interviews or relying on indirect measures such as problem correctness or solution time. Although it is not designed to be used by instructors for assigning grades, the rubric could be used by instructors for assessing the impact of their curriculum on their students' problem-solving performance, as well as determining how well their grading correlates with expertlike problem-solving behavior by their students.

B. Categories and scoring

After testing many different combinations of categories and scoring on different types of written problem solutions and with different raters, we arrived at the five categories of the MAPS rubric below as being the easiest and most reliable to implement at large scale.

TABLE I.

	5	4	3	2	1	0	NA(problem)	NA(solver)
USEFUL DESCRIPTION	The description is useful, appropriate, and complete.	The description is useful but contains minor omissions or errors.	Parts of the description are not useful, missing, and/or contain errors.	Most of the description is not useful, missing, and/or contains errors.	The entire description is not useful and/or contains errors.	The solution does not include a description and it is necessary for this problem /solver.	A description is not necessary for this *problem.* (i.e., it is given in the problem statement)	A description is not necessary for this *solver.*
PHYSICS APPROACH	The physics approach is appropriate and complete.	The physics approach contains minor omissions or errors.	Some concepts and principles of the physics approach are missing and/or inappropriate.	Most of the physics approach is missing and/or inappropriate.	All of the chosen concepts and principles are inappropriate.	The solution does not indicate an approach, and it is necessary for this problem/ solver.	An explicit physics approach is not necessary for this *problem.* (i.e., it is given in the problem)	An explicit physics approach is not necessary for this *solver.*
SPECIFIC APPLICATION OF PHYSICS	The specific application of physics is appropriate and complete.	The specific application of physics contains minor omissions or errors.	Parts of the specific application of physics are missing and/or contain errors.	Most of the specific application of physics is missing and/or contains errors.	The entire specific application is inappropriate and/or contains errors.	The solution does not indicate an application of physics and it is necessary.	Specific application of physics is not necessary for this *problem.*	Specific application of physics is not necessary for this *solver.*
MATHEMATICAL PROCEDURES	The mathematical procedures are appropriate and complete.	Appropriate mathematical procedures are used with minor omissions or errors.	Parts of the mathematical procedures are missing and/or contain errors.	Most of the mathematical procedures are missing and/or contain errors.	All mathematical procedures are inappropriate and/or contain errors.	There is no evidence of mathematical procedures, and they are necessary.	Mathematical procedures are not necessary for this *problem* or are very simple.	Mathematical procedures are not necessary for this *solver.*
LOGICAL PROGRESSION	The entire problem solution is clear, focused, and logically connected.	The solution is clear and focused with minor inconsistencies	Parts of the solution are unclear, unfocused, and/ or inconsistent.	Most of the solution parts are unclear, unfocused, and/ or inconsistent.	The entire solution is unclear, unfocused, and/ or inconsistent.	There is no evidence of logical progression, and it is necessary.	Logical progression is not necessary for this *problem.* (i.e., one-step)	Logical progression is not necessary for this *solver.*

Useful Description assesses a solver's process of organizing information from the problem statement into an appropriate and useful representation that summarizes essential information symbolically, visually, and/or in writing. A problem description could, but does not necessarily require, specifying known and unknown information, assigning appropriate symbols for quantities, stating a goal or target quantity, drawing a sketch or picture of the physical situation, stating qualitative expectations, drawing an abstracted physics diagram or graph, defining coordinate axes, and/or choosing a system. It does not require any specific representation that a student might include to gain partial credit with a particular instructor.

The term "useful" means that the description was used in the solution by that particular problem solver. The term "description" was chosen to be consistent with other uses of the term [11,51] and to avoid the multiple interpretations of the term "representation" [26,52,53]. This category is similar to the stages of "Understanding the problem" [54] or of "Representing the problem" [52] in some problem-solving frameworks.

Physics Approach assesses a solver's process of selecting appropriate physics concepts and principles to use in solving the problem. Here, the term "concept" is used to mean a general physics idea, such as a vector, or specific ideas such as momentum and velocity. The term "principle" is used to mean a fundamental physics rule used to describe objects and their interactions, such as conservation of energy or Newton's second law. This category also includes an understanding of the selected concept, such as the independence of perpendicular components of vectors.

The Physics Approach category reflects the expertlike process of selecting relevant physics principles that might be applicable to the situation before applying them to the specific context of a problem and planning a solution [51,53,55,56] and is similar to the "Evidence of conceptual understanding" category [11] and the "General approach" category [44,45] in other rubrics.

Specific Application of Physics assesses the solver's process of applying physics concepts and principles to the specific conditions in a problem. Specific application often involves connecting the objects, quantities, and constraints in a problem using specific physics relationships. It can include a statement of definitions, qualitative relationships between quantities, equations, initial conditions, and a consideration of assumptions or constraints in the problem.

This category separates the identification of appropriate principles and concepts in the physics approach from the actual application of those principles to the specific conditions in the problem. This is consistent with other descriptions of expertlike problem-solving models and strategies [53,56] and other assessments of problem solving [44,45].

Mathematical Procedures assesses the solver's process of selecting appropriate mathematical procedures and following mathematical rules to obtain target quantities. Examples of these procedures include algebraic strategies to isolate quantities or to simplify expressions, substitution, integration operations, or "guess and check" for differential equations. The term mathematical "rules" refers to processes from mathematics, such as the Chain Rule in calculus or appropriate use of parentheses, square roots, logarithms, and trigonometric definitions.

This category has analogs in many other problem-solving frameworks and rubrics [11,12,34,45,51–54] but differs from some in that it doesn't require students to solve equations in any particular manner (e.g., symbolically first) to receive a higher score.

Logical Progression assesses the solver's processes of staying focused on a goal while demonstrating internal consistency. The category checks whether the overall problem solution progresses toward an appropriate goal in a consistent manner where the backing for each step is evident, although not necessarily explicitly stated. The process may include revisions, rerouting, or intuitive leaps. This category does not require explicit evidence that the solution was evaluated because students (and experts) often do not explicitly evaluate their solution unless specifically instructed to do so and the rubric is intended to be independent of strategy-modeling instructional techniques.

The term "logical" is meant to convey that the solution has a coherent ordering and that the solver's reasoning can be understood from what is written, is internally consistent, and is externally consistent with a student's knowledge of nature. It emphasizes the ability to provide coherent explanations, so important in science and engineering [56].

The term "logical progression" can also be found in earlier assessments of problem solving [11,44,45] and this category agrees with the problem-solving assessment that includes clear interpretation or specification of the quantities involved, completeness of the answer, internal consistency of the argument, and external consistency of relationships and the magnitude of values [27]. It differs from some rubrics in that it does not evaluate a student's process based on working forwards or working backwards, both of which are found in expert problem solving.

Scores for each category of the rubric range from 0 (worst) to 5 (best) with additional "not applicable" categories for the problem and for the specific solver, NA(problem) and NA(solver). The NA(problem) score means that a particular category was not probed by the problem because those decisions were not required for that problem. For example, if an explicit description was provided in the problem statement or was not necessary to solve the problem, the useful description would be scored as NA(problem). The NA(solver) score means that based on the overall solution, it was judged that this set of decisions might have been made by the solver but not written down. This often occurs for experts who usually do not write down all of their internal processes. A score of NA

You are designing part of a machine to detect carbon monoxide (CO) molecules (28 g/mol) in a sample of air. In this part, ultraviolet light is used to produce singly charged ions (molecules with just one missing electron) from air molecules at one side of a chamber. A uniform electric field then accelerates these ions from rest through a distance of 0.8 m through a hole in the other side of the chamber. Your job is to calculate the direction and magnitude of the electric field needed so that CO^+ ions created at rest at one end will have a speed of 8×10^4 m/s when they exit the other side.

FIG. 1. Problem corresponding to student solutions in Figs. 2 and 3. It can be solved in at least two different ways. One uses Newton's second law to relate the force on a CO molecule to its acceleration, and then kinematics to relate this acceleration to the final speed. A second method uses conservation of energy to relate the work done on a CO molecule by the electric field to its kinetic energy (and thus its speed) when it exits the box. The correct answer is 1160 N/C.

(solver) does not require a correct solution. For example, if the solver did not draw either a picture of the situation or a force diagram, but correctly wrote down the applicable Newton's law equations, Useful Description would be scored NA(solver). However, if the forces used were inappropriate for the situation and a picture or diagram was not used but might have helped, then NA(solver) would not be used. This score is included because the rubric recognizes the possibility that some students begin to develop some of the automated processes characteristic of an expert [57]. NA(solver) is a statement of ignorance on the part of the rater as to whether the problem-solving skill represented by that rubric category is used by the student.

Furthermore, it is important to note that the scoring for all categories is based on consistency with the rest of the solution. For example, if a solver made an error in constructing a useful description, the scoring in the other categories is based on the consistent use of that error going forward. This avoids overcounting a single error. Since a true problem-solving process is dynamic, allowance is made for solvers to change their interpretation as they move through the solution process.

To promote ease of use, the language describing the scoring of each category is kept consistent. A score of 0 means that there is no evidence of the category although it was judged to be necessary for that solver and that problem, 1 means the category was in evidence but was entirely inappropriate, 2 means mostly inappropriate or incomplete, 3 means mostly appropriate but with significant parts that are inappropriate or incomplete, 4 designates complete and appropriate with minor omissions or minor errors, and 5 is complete and appropriate. A numerical score that ranges from 0 to 5 was found from testing to be the minimum range that raters felt provided a sufficient delineation of responses.

C. Examples of use

Figures 2 and 3 show the application of the rubric to two student solutions to the problem in Fig. 1. The first solution (Fig. 2) has many characteristics of expert problem solving. Although the student has not drawn a picture, based on the appropriateness of the conservation of energy approach used, such a description seems to be unnecessary for that student. Thus, the Useful Description category is scored NA(solver). The Physics Approach (using conservation of energy) and the Specific Application of Physics (applying conservation of energy to this situation) are both correct and complete. Mathematical Procedures has been scored as a 4 because of a minor error in the algebra where a "2" migrated from the denominator to the numerator. Logical Progression has been scored as a 4 because the solution neglects to provide any reasoning for the use of the molecular mass in the final calculation. Other than that, the solution process is clear and includes an evaluation of the final answer.

A second student solution is presented in Fig. 3. Although the student does draw a picture, there are numerous errors in the picture, including the direction of the electric field and the presence of a magnetic field. Thus, Useful Description is a 2. Physics Approach has been scored as 3 because, although some of the principles that the student chose to use are appropriate to the situation they describe, they are not consistent. Specific Application of Physics has been scored as 2 because of the large number of errors in applying these principles, such as assuming a constant velocity in applying Newton's laws, and using an inappropriate mass in the kinetic energy calculation. Although there is nothing wrong with the mathematical operations that the student performs, the mathematical treatment of acceleration is incorrect and the procedures can never lead to a productive conclusion, earning a rating of 3 for Mathematical Procedures. Finally, the Logical Progression of the solution is rated a 2 because the solution has several internal inconsistencies, as well as containing inappropriate units (the equation on the bottom right calculates a time in units of m/s).

This second solution displays many of the characteristics of novice problem solving. The equations seem to have been selected because they contained the quantities referred to in the problem. For example, the force equation includes a magnetic field even though none is mentioned in the problem, perhaps because this is the only way the solver could include a velocity term in the equation. This equating of electric and magnetic force might have matched the pattern of a previously practiced solution with similar surface features. After getting stuck, the solver writes down an incorrect equation for the acceleration, as well as the expression for kinetic energy. There is no apparent backing for these attempts other than trying to find an equation that might relate a velocity to an electric field.

Several more examples of the application of the rubric to student solutions can be found in the training materials for the rubric on our website (http://groups.physics.umn.edu/physed/rubric.html).

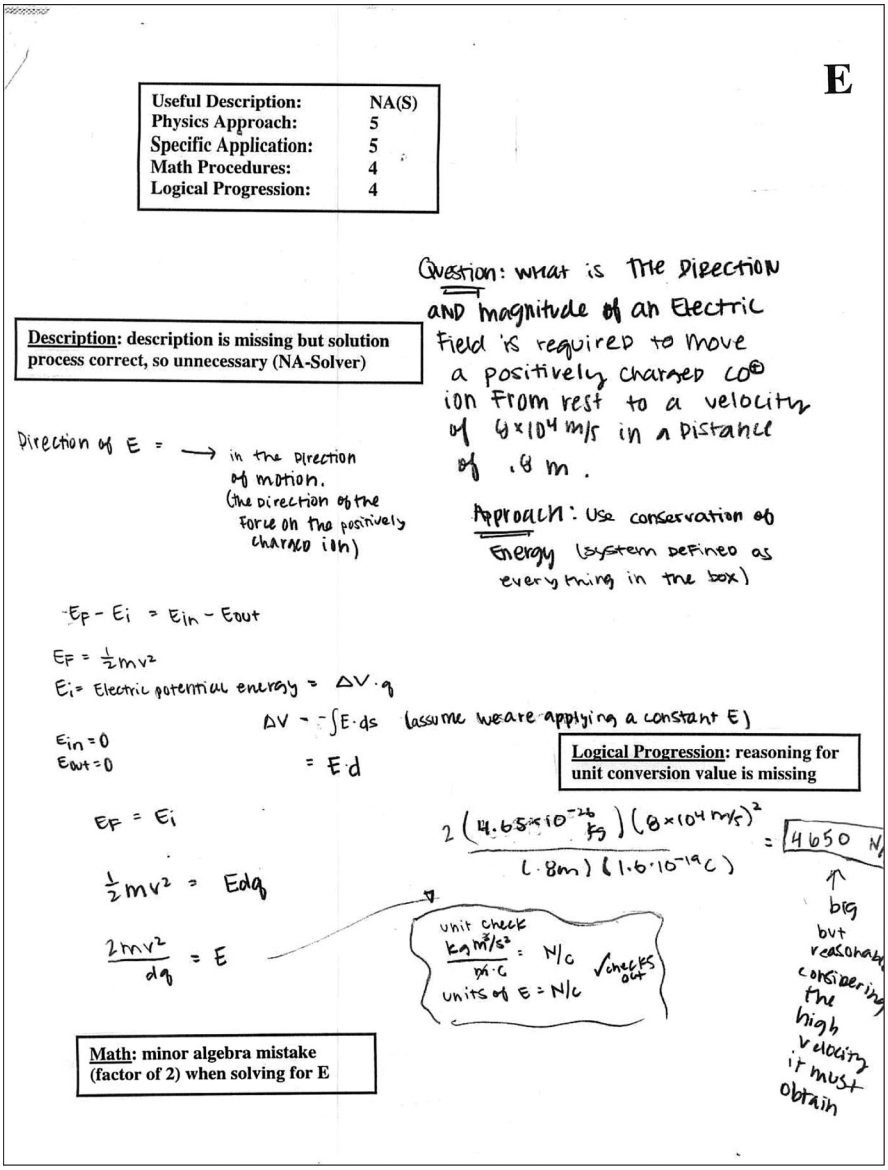

FIG. 2. Example of applying the MAPS rubric to a student solution.

IV. RUBRIC TESTING: VALIDITY AND RELIABILITY

In developing the rubric, we gathered evidence and conducted experiments to determine its validity and reliability. Such tests are necessary before any measurement instrument can be used with confidence.

A. Validity

Validity is established by the evidence, data, and theory that support the interpretation of the measurement [13], along with its appropriateness, meaningfulness, and usefulness [14]. The *Standards for Educational and Psychological Testing* [13] identifies several categories of evidence that can be used to support the validity of an assessment, including content, response processes, relations to other measures, and internal structure. Others [14] also consider the generalizability of scores.

1. Content

Content refers to the wording and formatting of items on an assessment in addition to the procedures for scoring [14]. In the MAPS rubric, the content is the problem-solving categories assessed by the rubric. As discussed above, the rubric categories match the descriptions of critical physics problem-solving processes in the research literature. Furthermore, these categories share many similarities with the work of other researchers who have developed rubrics for assessing problem solving, both at Minnesota [11,43–45] and elsewhere [46–50]. Furthermore, these categories are consistent

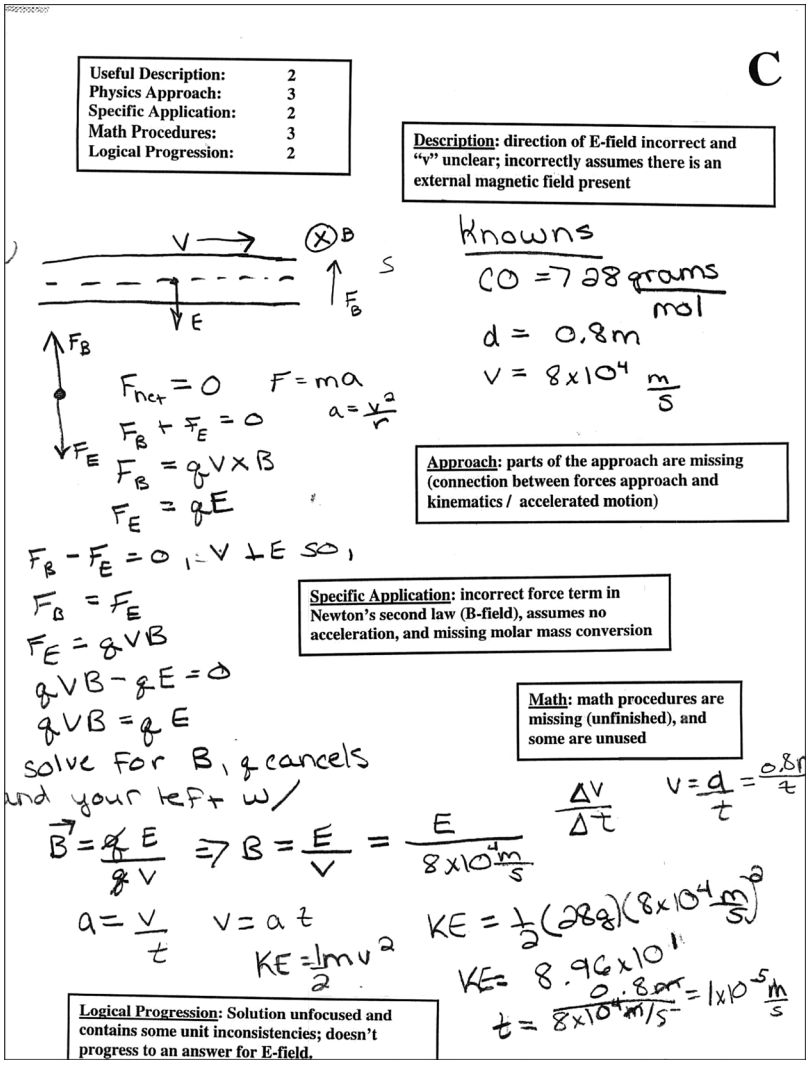

FIG. 3. Example of applying the MAPS rubric to a second student solution.

with a study of faculty beliefs regarding student problem solving [58]. This provides content evidence of validity.

2. Response processes

Response processes refers to evidence that the rubric categories reflect the problem-solving processes actually engaged in by solvers. The problem-solving processes in which expert solvers engage have been well established in the literature and the rubric was constructed to reflect those processes. To investigate if the problem-solving processes of students in transition towards expertise can be characterized by the rubric, we performed a study in which eight students, who were in such a transition, solved physics problems and then were interviewed about their process [15,32]. The relevant parts are summarized below.

All of the interviewed students were volunteers (7 males and 1 female) enrolled in an introductory calculus-based mechanics course for physical science and engineering majors at the University of Minnesota during the Spring 2009 semester. The interviews took place near the end of the semester. The final course grades of these eight participants indicate that they were in the upper half of the class (4 students earned A's, 2 students earned an A−, and the remaining two earned a B+ and a B; the average grade for this class was a B−) and thus that they could be expected to be somewhere between novice and expert in their problem solving. Two of the 8 were non-native English speakers.

During the problem-solving interviews, these students were asked to solve up to 3 physics problems while being video and audio recorded. Participants were asked to talk out loud while working on a problem if that was comfortable, or they could wait and explain their solution at the end. Students were provided the same support that they received on course tests, a copy of the instructor's equation sheet and a calculator.

You are working at a construction site and need to get a 14-N bag of nails to your co-worker standing on the top of the building (9 meters from the ground). You don't want to climb all the way back up and then back down again, so you try to throw the bag of nails up. Unfortunately, you're not strong enough to throw the bag of nails all the way up so you try another method. You tie the bag of nails to the end of a 65-cm string and whirl the string around in a vertical circle. You try this, and after a little while of moving your hand back and forth to get the bag going in a circle you notice that you no longer have to move your hand to keep the bag moving in a circle. You think that if you release the bag of nails when the string is horizontal to the ground that the bag will go up to your coworker. As you whirl the bag of nails around, however, you begin to worry that the string might break, so you stop and attempt to decide before continuing. According to the string manufacturer, the string is designed to hold up to 500 N. You know from experience that the string is most likely to break when the bag of nails is at its lowest point.

FIG. 4. First (main) problem-solving task used in student interviews.

The interview problems were context-rich problems [8] selected to look similar to ones used in their course. Context-rich problems were chosen because they require multiple decisions by the solver and are not easily solved using novice solution procedures. The first, and most involved, problem was adapted from previous research to determine instructors' views on problem solving [58] and is shown in Fig. 4. This problem was made much more ambiguous than most context-rich problems used at this level to allow the students to exhibit the most flexibility in their problem solving. Features that make this problem a difficult context-rich problem include: the target of the problem is not explicitly stated, a combination of at least two principles is necessary for the solution, and the solver must infer or assume some information. This problem also has the characteristic that it is possible to obtain a correct answer with incorrect or incomplete reasoning.

The students spent between 6 and 26 min solving the first problem and demonstrated a large range of behavior in their solution. One student solved the problem correctly and another had a mistake that (s)he corrected during the interview process. Another two students had the correct answer but demonstrated incomplete support for their solution choices. Two students used inappropriate physics and proceeded to an incorrect answer. Two students did not reach a satisfactory answer to the first problem and chose to stop their work to explain their thinking. After working through a problem to their satisfaction, the students were asked to go back and explain their solutions to the researcher, as well as to answer eight questions about their solution process in a semi-structured interview. The audio files for the eight interviews were transcribed and analyzed, with students' statements assigned to one of several prescribed code categories or "nodes." Five of these nodes corresponded to the rubric categories while others corresponded to the specific interview questions or an "Other" node, used for statements not explicitly addressed by the rubric or the questions.

There were 549 total passages coded in the eight interview transcripts. Of these, 276 (50%) were coded as rubric related, 189 as related to an interview question, and 84 as "Other." A more detailed breakdown is shown in Table II. Of the statements in the "Other" coding node, half pertained to monitoring progress, evaluating the answer, and/or checking units. Although these processes are desirable and could contribute to the logical flow of a solution, they are not explicitly scored by the rubric for reasons stated previously. Additional processes in this category included solving equations in symbolic form prior to plugging in numbers and referencing the equation sheet.

The fact that half of the students' statements were directly related to the rubric categories and that each of the five rubric categories had at least 14% of those statements provides evidence that the rubric categories do, in fact, reflect the processes that students actually engaged in while solving physics problems, i.e., response processes evidence for validity.

3. Relationship to other measures of problem solving

A third kind of validity evidence is the extent to which scores on the rubric agree with other measures of students' problem-solving performance. After the interviews in the study just described, the written solutions produced by the eight students were scored twice by one of the authors not involved in the interview process, once based only on the written work and the second time based on both the written work and the recorded explanations they gave during the interviews. The rubric scores were identical in every case with two exceptions: (1) one student's incorrect physics reasoning was not apparent from the written solution, so the rubric score for the Specific Application of Physics category based only on that was one point higher than the score based on both written and verbal evidence, and (2) based on written work only, three of

TABLE II. Total number of transcript passages from all eight students assigned to each coding node.

Coding category (NVivo node)	Total passages
Useful Description	43
Physics Approach	40
Specific Application	82
Math Procedures	39
Logical Progression	72
Eight interview questions	189
Other	84
Total	549

the students had scores of NA(solver) in either the Mathematical Procedures or Physics Approach categories. Based on the interview data they would have had a 5 in these categories. The agreement of rubric results based solely on student written work and those based on both written work and verbal interviews provides this type of evidence for validity.

In addition, we compared rubric scores for student solutions to in-class test problems to grades for those same problems independently assigned by graduate teaching assistants (TAs) in those courses. During one semester, student solutions to problems on in-class midterm exams were gathered from two sections of a calculus-based first-semester introductory physics courses for physical science and engineering majors. One section had 230 students and the other had 250 students. The rubric was applied to four different problems from each of the two sections (eight total problems). The problems addressed topics including kinematics, Newton's laws, and conservation of energy and were similar to traditional end-of-chapter textbook problems. In all cases, the problems were free response, so that the students showed their work. Each of the eight problems was graded independently by a different TA.

As noted previously, there are many ways in which grading could differ from assessing for problem-solving expertise. However, all of the TAs who graded the problems examined in this study had participated in an extensive TA orientation and support process that included instruction in the differences between expert and novice problem solving and the techniques to move students toward more expertlike behavior [59]. It is important to note that the TAs were not familiar with the idea of using a rubric to assess problem solving and were unaware of the MAPS rubric and its categories.

An overall rubric score for each problem solution was calculated by assigning a score between 0 and 5 for each of the five categories of the rubric to a written solution, and then adding all the scores together. This summed score (with a maximum of 25) was then divided by the number of categories with a numeric score [to eliminate the influence of NA(problem) and NA(solver) scores] and the result was then divided by the maximum category score (5) to obtain a percent score for each solution. This gave a single overall score that could be compared with the grade for that problem assigned by the course TAs.

Correlation coefficients between the overall rubric score and the TA score for each of the eight individual problems were high, ranging from 0.75 to 0.96, and are shown in Table III, along with correlations between the TA score and rubric scores for each of the five rubric categories. Because the rubric category scores are ordinal, the Spearman ρ was used for computing correlations. The correlation between the total rubric score and the grader score for all eight problems taken together ($N = 918$, shown in Figure 5) is 0.82 ($p < 0.0001$). In this analysis, we eliminated the 99 solutions that received either a 0% or a 100% on both the TA and rubric scores to control for the effects of ceilinged or floored scores on the correlations. (Including those points raises the correlation coefficient to 0.87.) This very strong correlation provides yet another type of evidence of validity for the problem-solving rubric.

Comparing the correlations between the grade for a problem and its rubric scores in individual categories can provide some insight into the grading process and how certain aspects of problem solving might be emphasized in the course and/or the mind of the graders. As mentioned before, the graders had no knowledge of the rubric. As can be seen in Table III, rubric scores from the Physics Approach, Specific Application of Physics, and Logical Progression categories were more highly correlated with TA score than those from the other categories. This suggests that the graders either didn't pay attention to or gave a lower weight to some aspects of the problem

TABLE III. Correlations between rubric scores (both overall and for each individual category) and TA grades for each of eight midterm problems ($N = 918$ solutions). The overall (sum) score adds together all of the category scores with equal weight as described in the text and then calculates a correlation with the TA grade.

	Problem 1 (kinematics) $N = 48$	Problem 2 (kinematics) $N = 110$	Problem 3 (forces) $N = 92$	Problem 4 (forces & circular motion) $N = 160$	Problem 5 (forces & work) $N = 81$	Problem 6 (forces & work) $N = 156$	Problem 7 (energy) $N = 92$	Problem 8 (energy) $N = 179$	All Problems $N = 918$
Useful Descrip.	0.47	0.60	0.39	0.46	0.41	0.49	0.62	0.73	0.55
Physics Approach	0.90	0.88	0.63	0.79	0.66	0.67	0.71	0.77	0.72
Specific Appl.	0.93	0.81	0.79	0.86	0.77	0.71	0.80	0.81	0.80
Math Procedures	0.69	0.49	0.71	0.70	0.50	0.71	0.69	0.75	0.61
Logical Progress	0.81	0.74	0.72	0.80	0.66	0.70	0.83	0.73	0.73
Overall (sum)	0.96	0.87	0.82	0.90	0.75	0.76	0.89	0.87	0.82

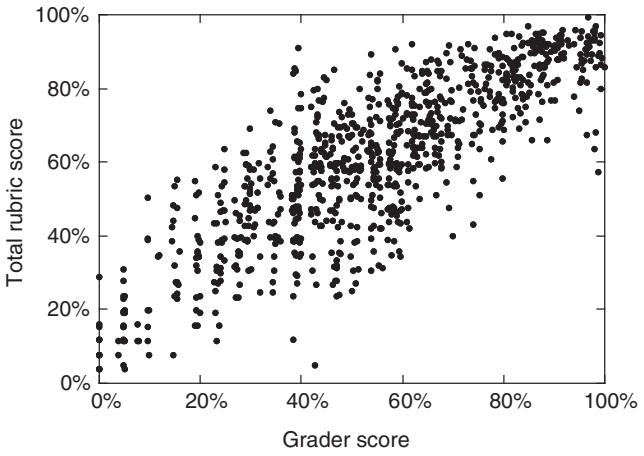

FIG. 5. Scatter plot of total rubric score vs TA grade for all eight midterm problems ($N = 918$). The correlation (ρ) between the two scores is 0.82 ($p < 0.0001$). Points are shifted by a small random number so as not to mask clusters of scores.

solutions corresponding to the other two categories. In all cases, however, the grade given by the grader and the overall rubric score of the expert raters gave a consistent measure of the quality of the solution.

These correlations and the comparison between rubric scores based on written work and interviews described at the beginning of this section provide another type of evidence of validity for the rubric.

4. Internal structure

A fourth type of evidence of validity can be found from studying the internal structure of the MAPS rubric, in particular, the extent to which correlations between scores in the five rubric categories agree with expectations. To investigate this, we computed the partial correlations among the rubric category scores [15].

The intercategory correlation matrix for student solutions from all eight problems is shown in Table IV. Again, all solutions that received either a 0% or a 100% on both rubric and TA scores have been eliminated from the calculation to control for the effects of ceilinged or floored scores. Overall the categories are not highly correlated with each other, showing that they represent, to some extent, independent aspects of problem solving. There are, however, some correlations. The Physics Approach and Specific Application of Physics categories show a significant correlation ($\rho = 0.47$). This correlation has been found in previous research [45] and is not surprising, since it is difficult to correctly apply an incorrect approach. Also, as one might expect, the Logical Progression category, being a measure of the "overall" coherence and consistency of the solution, shows some correlation to 3 of the other categories. These correlations provide a fourth type of evidence of validity of the MAPS rubric.

5. Generalizability

Although not explicitly included in all descriptions of validity evidence, it is important for an assessment to be applicable across different populations and contexts [14]. In other words, the rubric should be able to be applied to written solutions from a variety of solvers from novice to expert, spanning a variety of physics topics, and from a variety of classes taught by different instructors.

We have tested the rubric on a variety of physics problem solutions that span the topics in a typical introductory university physics sequence, including mechanics and electricity and magnetism, from both midterm tests and final exams [15]. We have also tested the rubric with student solutions to different types of problems, including those that are similar to traditional textbook problems and those that are context rich [8], and found that the rubric was applicable to all the problem solutions on which it was tested. It should be noted, however, that certain types of problems do not probe one or more of the problem-solving processes scored in the rubric. For example, some problems provide enough pictures and/or diagrams to make the construction of a useful description unnecessary. Others may provide explicit or implicit prompts to use particular physics principles or draw particular pictures or diagrams. In those cases, the rubric yields an NA score for certain categories.

We also tested an earlier, very similar version of the rubric (identical except it had category scores that ranged from 0 to 4) on instructor solutions. Two problem solutions from each of 38 chapters ($N = 76$) in a popular calculus-based physics textbook [60] from the instructor solution manual were scored using the rubric. These solutions were typically very sparse and did not give much explicit reasoning. In addition, homework solutions handwritten by a physics instructor for an entire introductory physics

TABLE IV. Intercategory correlation coefficients ρ between rubric category scores for eight midterm problems from two introductory physics sections ($N = 918$). Correlations in bold are statistically significant ($p < 0.01$).

	Useful Description	Physics Approach	Specific Application	Math Procedures	Logical Progression
Useful Description	1	**0.15**	**0.20**	−0.05	0.06
Physics Approach		1	**0.47**	0.01	**0.28**
Specific Application			1	**0.19**	**0.24**
Math Procedures				1	**0.50**
Logical Progression					1

course ($N = 83$) were also scored with the rubric. These solutions were more detailed and included steps of the reasoning process. Because these solutions looked qualitatively different from student solutions, it was impossible to use a blind scoring process and in each case, the scorer knew the source of the solutions.

The frequency of rubric scores was very similar for both types of instructor solutions, regardless of the level of written detail. Virtually all (>97%) rubric scores for these instructor solutions received the highest possible value or an NA(problem) or NA(solver) score. By comparison, scores of student solutions (using this same earlier version of the rubric) span the entire range of rubric scores. Figure 6 shows rubric scores in all categories from 160 student solutions (described more completely in the next section) and the 159 textbook and instructor solutions described above. From the differences in score frequencies it is easy to distinguish between instructor and student solutions, independent of their format. This applicability of the rubric to solutions to problems from a range of topics, from a range of sources, written by a range of solvers provides yet another type of evidence of validity.

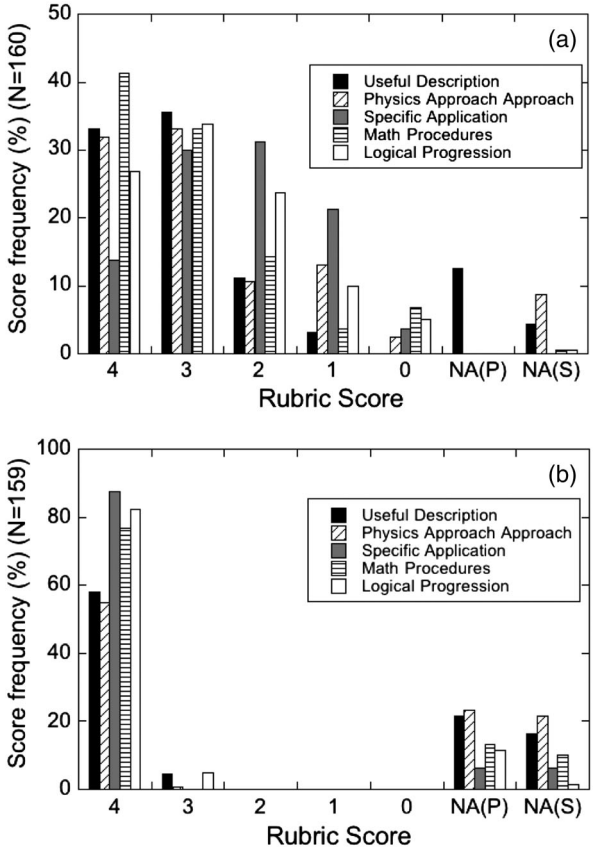

FIG. 6. Rubric scores from (a) 160 student solutions and (b) 159 solutions written by an instructor or taken from a textbook solution manual. This rubric was identical to the final rubric but used a maximum score of 4.

B. Reliability

Another important test of a rubric is its reliability, which includes the agreement of scores between multiple raters, as well as agreement of scores between raters just learning to use the instrument and the rubric developers. In order to obtain good agreement, raters must agree on the meaning of the scoring categories as well as the levels within each category. To assess reliability, we conducted two types of studies. The first type involved raters who had considerable experience teaching physics and/or in physics education research. The second type involved raters who had much less experience teaching physics and little familiarity with physics education research.

1. Studies with expert raters

The first study of this type involved two raters, one of whom was an advanced graduate student in physics education research and a rubric developer. The other rater was an experienced high school physics teacher. A total of eight different free-response final exam problems, five from a calculus-based introductory mechanics course and three from an algebra-based introductory mechanics course, were scored over a time period of one month. Twenty solutions were selected semirandomly for each of the eight problems, with stipulations that the solutions be legible and reflect a range of quality and detail. First, each rater independently scored the 20 solutions to the first problem and recorded their individual scores. It is these scores that were used for the reliability comparison.

Next, in a process to determine how much training was necessary to achieve an interrater reliability sufficient for use of the rubric for research purposes, the raters discussed their interpretation of the rubric scores for that problem, including the problem-specific criteria used to specify the scoring levels for each category, until a final consensus was reached. Then, each rater independently scored 20 solutions to a second problem. The consensus building discussion was then repeated for the second problem. This process was repeated until all problems were scored. At the end of the process for all eight problems, each rater independently rescored the solutions to the first problem from approximately one month earlier. The results of the independent scorings for all eight problems, before discussion, were then analyzed to determine reliability. Because the descriptions of the scoring levels for each of the rubric categories are written generally, such an iterative training process is necessary for building a consistency between the two raters that is sufficient for research purposes.

One method for quantitatively judging reliability is through the use of the weighted kappa [61]. Weighted kappa values range between −1 and 1 with 0 indicating chance agreement and a 1 indicating perfect agreement. The weighted kappa, as a function of the number of solutions scored, is shown in Fig. 7, as well as in

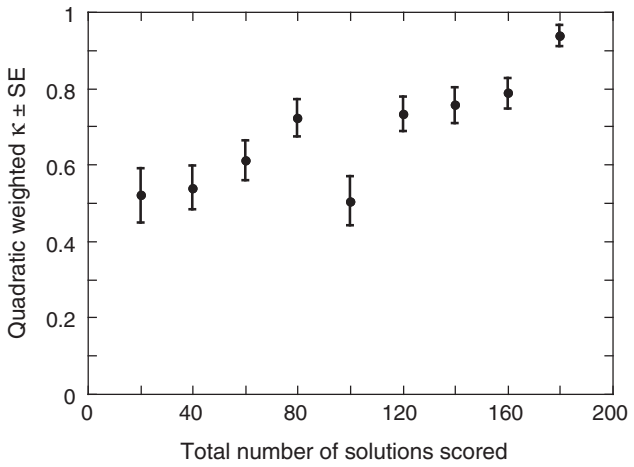

FIG. 7. Graph of score agreement (weighted kappa) for two raters as a function of number of solutions scored for eight problems. The solutions were scored sequentially in time. The ninth data point is a rescoring of the first problem initially scored as the first data point.

Table V. The ninth and final data point is an independent rescoring of the 20 solutions of the first problem after all eight problems had been scored and discussed. Table V also includes the before-discussion perfect score agreement percentage between the two raters, both overall and for each rubric category, for each of the problems. The percentage of score agreements within 1, averaged over all eight problems, was in excess of 97% ± 1% for each of the 5 categories and 99% ± 1% overall. The reason for the dip in agreement on problem 5 is unclear, but a comparison of the scores assigned by each rater in the Physics Approach and Specific Application of Physics categories indicated there were several instances in which the high school teacher scored a "zero" for these categories or NA(solver) when the researcher assigned a numerical score. Since a "zero" represents a missing but necessary solution aspect for the rubric version used, it is possible that the teacher had different and perhaps narrower criteria for evidence of these categories than the researcher.

As can be seen from Fig. 7, the agreement between the raters generally increased with time. A kappa value above 0.60 is considered to indicate "substantial agreement" and a value above 0.80 is considered "almost perfect agreement" [62]. Thus, we believe that the rubric can, with some training, provide a statistically reliable way to assess students' problem-solving performance that is sufficient for research purposes and more than adequate for instructor use.

In addition to this study, we also conducted three other studies with multiple raters in which raters who were experienced physics teachers or physics education researchers used the rubric to assess a common set of student solutions. None of the raters had been involved in the development of the rubric. The results from those studies were similar to the one just described. After scoring and discussing 10 common student solutions to a problem, further rating of student solutions to the same problem produced exact agreement in all five categories between 50% and 85% of the time and agreement within one score in excess of 90% of the time.

2. Studies with less experienced raters

To determine the extent to which an individual instructor, without access to consultation from an expert rater or one of the rubric developers, could use the rubric to score problem solutions in a manner consistent with the intentions of the developers, we performed two studies with less expert raters. The first involved physics graduate students in at least their third year of study who were experienced TAs. Eight such graduate students were randomly assigned to two groups of four people each. Those in the first group used the rubric to score eight student solutions from an introductory mechanics final exam problem and those in the second group used the rubric to score eight student solutions from an introductory electricity and magnetism final exam problem. Both problems were free response. The graduate students were provided with an instruction

TABLE V. Percentage of perfect score agreement between the two raters for each of the eight final exam problems in the reliability study. Scores could range from 0 to 5 in each category. Percent agreement values include NA(solver) scores. Problems 1–5 were taken from a calculus-based mechanics course and problems 6–8 were from an algebra-based course. As described in the text, the agreement within 1 averaged over all categories was in excess of 97% ± 1%.

Category	Problem 1 energy $N = 20$	Problem 2 forces $N = 20$	Problem 3 momentum $N = 20$	Problem 4 angular $N = 20$	Problem 5 oscillations $N = 20$	Problem 6 kinematics $N = 20$	Problem 7 forces $N = 20$	Problem 8 momentum $N = 20$	Problem 1 rescore $N = 20$
Useful Description	70%±10%	70%±10%	85%±8%	65%±11%	NA(problem)	85%±8%	85%±8%	65%±11%	95%±5%
Physics Approach	65%±11%	45%±11%	90%±7%	70%±10%	40% ± 11%	90%±7%	75%±10%	95%±5%	100%±0%
Specific Appl.	60%±11%	55%±11%	40%±11%	75%±10%	45%±11%	50%±11%	75%±10%	90%±7%	90%±7%
Math Procedures	60%±11%	55%±11%	45%±11%	75%±10%	65%±11%	80%±9%	75%±10%	70%±10%	95%±5%
Logical Progress	65%±11%	55%±11%	65%±11%	75%±10%	30%±10%	60%±11%	85%±8%	70%±10%	95%±5%
Overall	64%±5%	56%±5%	65% ± 5%	72%±4%	45%±5%	73%±4%	79%±4%	78%±4%	95%±2%
Weighted kappa	0.52±0.07	0.54±0.06	0.61±0.05	0.77±0.04	0.57±0.06	0.74±0.04	0.76±0.05	0.79±0.04	0.94±0.03

TABLE VI. Percent agreement of third-year graduate student scores with rubric developer scores averaged over all problems before and after training (mechanics problems). The combined agreement within one score is 74% ± 2% before training and 85% ± 2% after training. *N* is the number of problems scored by the participants in the group (4 raters×8 problems before training; 4 raters×10 problems after training).

Mechanics	Before training			After training		
Rubric category	Perfect agreement ($N = 32$)	TAs one above ($N = 32$)	TAs one below ($N = 32$)	Perfect agreement ($N = 40$)	TAs one above ($N = 40$)	TAs one below ($N = 40$)
Useful Description	18% ± 7%	32% ± 8%	7% ± 5%	53% ± 8%	40% ± 8%	5% ± 3%
Physics Approach	31% ± 8%	38% ± 9%	9% ± 5%	25% ± 7%	20% ± 6%	20% ± 6%
Specific Application	38% ± 9%	43% ± 9%	9% ± 5%	50% ± 8%	33% ± 7%	8% ± 4%
Math Procedures	13% ± 6%	56% ± 9%	3% ± 3%	30% ± 7%	40% ± 8%	8% ± 4%
Logical Progression	38% ± 9%	28% ± 8%	3% ± 3%	63% ± 8%	10% ± 5%	18% ± 6%
Overall	28% ± 4%	40% ± 4%	6% ± 2%	44% ± 4%	29% ± 3%	12% ± 2%

sheet, a copy of the rubric, brief definitions of each rubric category, the problem statement, a correct solution to the problem, a blank scoring template table, and eight student solutions. There was no other contact with the researcher and no organized contact among the graduate students.

The graduate students in both groups were asked to use the rubric to score their eight student solutions (32 total solutions per group) without any explicit training or discussion. After submitting their scores and rationale, they received brief self-training materials written by one of the rubric developers consisting of rubric scores and rationales for three of the eight solutions they had previously scored themselves. Figures 2 and 3 are examples drawn from the self-training materials. The graduate students received written instructions to read the materials and compare the given scores and rationales to their own previous work. They were then instructed to rescore the remaining five solutions from before in addition to five new solutions (40 total solutions per group). Although the graduate students were not given an explicit deadline for returning materials, most completed the task within one week.

Tables VI and VII show the percentage agreement of the graduate students' scores with the developers' scores. As can be seen, even before training, the agreement of scores within 1 was already high (74% ± 2% for mechanics problems and 81% ± 2% for electricity and magnetism problems) and the brief training resulted in only modest gains (agreement within 1 of 85% ± 2% for mechanics and 88% ± 2% for electricity and magnetism). Weighted kappa values for the post-training agreements are 0.41 ± 0.04 for the mechanics problems and 0.43 ± 0.04 for the electricity and magnetism problems, often designated as "moderate agreement." These results suggest that consistency between the interpretations of such raters and the rubric developers (though not to the level required for research) can be developed without extensive training.

Based on this study and comments of the graduate student raters, the training materials and methods were revised in a number of ways, including modifying the rubric language and scoring scales, increasing the number of developer-scored example solutions from three to five, and providing a wider range of score examples for each rubric category, including NA(solver) examples, in addition to improving the clarity and readability of the training materials.

A second study was then performed with first-year physics graduate students who were in their first semester as teaching assistants. In contrast with the first study, the entire training and rating process took place within a single session. Nineteen students participated. In the first 20 min, the students read one page of instructions explaining the

TABLE VII. Percent agreement of third-year graduate student scores with rubric developer scores averaged over all problems before and after training (electricity and magnetism problems). The combined agreement within one score is 81% ± 2% before training and 88% ± 2% after training. *N* is the number of problems scored by the participants in the group (4 raters×8 problems before training; 4 raters×10 problems after training).

Electricity and magnetism	Before training			After training		
Rubric Category	Perfect agreement ($N = 32$)	TAs one above ($N = 32$)	TAs one below ($N = 32$)	Perfect agreement ($N = 40$)	TAs one above ($N = 40$)	TAs one below ($N = 40$)
Useful Description	56% ± 9%	31% ± 8%	3% ± 3%	41% ± 8%	25% ± 7%	15% ± 6%
Physics Approach	43% ± 9%	39% ± 9%	4% ± 3%	50% ± 8%	28% ± 7%	18% ± 6%
Specific Application	53% ± 8%	41% ± 9%	6% ± 4%	45% ± 8%	20% ± 6%	30% ± 7%
Math Procedures	29% ± 8%	8% ± 5%	12% ± 6%	50% ± 8%	8% ± 4%	17% ± 6%
Logical Progression	19% ± 7%	41% ± 9%	13% ± 6%	38% ± 8%	35% ± 8%	13% ± 5%
Overall	41% ± 4%	33% ± 4%	7% ± 5%	45% ± 4%	25% ± 3%	18% ± 6%

activity, solved the physics problem whose solutions they were to rate, read an instructor solution along with rubric and category descriptions, and then reviewed 5 example solutions with rubric scores and rationales provided by one of the developers. Following this minimal training, these graduate students took 10 min to score two solutions to that same problem and write comments. In total, these TAs spent 30–35 min participating in the rubric training and scoring task.

Table VIII shows the agreement between these first-year graduate students' scores with the developer's scores. As can be seen, despite the much shorter training time and relative lack of teaching experience among this second group of graduate students, the agreement was similar to that of the more experienced graduate students. The overall agreement within one score was $77\% \pm 3\%$. The weighted kappa averaged over all five rubric categories is 0.32 ± 0.04 (fair agreement) and all of these measures indicate an overall agreement that is significant ($p < 0.001$).

Thus, even without research-grade training, the level of agreement between the graduate student instructors and the rubric developers in both of these studies is likely sufficient for some instructional purposes such as finding areas of student difficulty for a class or unearthing grading inconsistencies. With feedback from an expert rater as described earlier, the precision of use of the rubric can be of research quality for teachers.

V. UTILITY: APPLICATIONS OF THE RUBRIC

Beyond the standard criteria of validity and reliability, a final consideration for an assessment instrument is its utility. It is important that researchers, curriculum developers, and instructors find an assessment applicable to their educational concerns and useful in clarifying those concerns. In particular, an assessment's results should not depend on artifacts of a specific pedagogy.

If used in a research context, such as testing the efficacy of a treatment or probing the relative problem-solving skills of different populations, a high reliability standard is necessary and is achievable using iterative training procedures among several raters as described previously in this paper. For this reliability to have generalized meaning, at least one of the raters must be very familiar with the extensive research literature on problem solving. With such research grade reliability established, the rubric could also be used to calibrate more indirect, but easier to use, assessment tools such as grading for research use, as illustrated in Sec. IV A 3.

In an instructional context, individual instructors might also choose to use the rubric to assess their own pedagogy or the skills of their students. In this case, the previously mentioned self-training materials available on our website can be used to arrive at an interpretation of the rating categories meaningful to a particular instructor. The rubric could be used to compare student populations within the instructor's context, for example, before and after the application of a long-term instructional strategy. One cannot compare student populations from different instructors unless several raters are used whose experience spans the different instructional contexts and whose inter-rater reliability has been established. Although the rubric is most reliable when used with populations in a statistical sense, it is also possible to use the rubric as one of several inputs to help an instructor diagnose an individual student's problem solving difficulties.

In this section, we discuss some possible applications of the rubric, along with evidence for the rubric's utility in those applications.

A. Assessing students' transition toward expertlike behavior

As the validity tests show, the rubric can be used to measure the degree to which a population of students solve problems in an expertlike manner. This will allow researchers and curriculum developers to assess the problem-solving performance of large numbers of students, a critical need when testing educational interventions in authentic situations. The rubric scores can be used as an absolute scale to determine how close a population is to expertlike problem-solving behavior along the different dimensions of the rubric, or as an overall measure by combining the category scores. It can also be used in a relative manner to determine the difference between a baseline and a treatment group. For example, examination of the scores of particular rubric categories can help assess the degree to which an

TABLE VIII. Percent agreement of first-year graduate students scores with rubric developer scores after a brief training activity. Overall agreement within one score is $77\% \pm 3\%$. N is the number of solutions scored by the participants (19 raters×2 solutions).

Category	Perfect agreement ($N = 38$)	TAs one above ($N = 38$)	TAs one below ($N = 38$)	Quadratic weighted kappa ($N = 38$)	Kappa significance ($N = 38$)
Useful Description	$41\% \pm 8\%$	$13\% \pm 5\%$	$18\% \pm 6\%$	0.24 ± 0.12	$p < 0.05$
Physics Approach	$35\% \pm 8\%$	$21\% \pm 7\%$	$24\% \pm 7\%$	0.40 ± 0.08	$p < 0.001$
Specific Application	$47\% \pm 8\%$	$21\% \pm 7\%$	$13\% \pm 5\%$	0.46 ± 0.09	$p < 0.001$
Math Procedures	$32\% \pm 8\%$	$26\% \pm 7\%$	$18\% \pm 6\%$	0.04 ± 0.12	Not sig.
Logical Progression	$32\% \pm 8\%$	$26\% \pm 7\%$	$16\% \pm 6\%$	0.17 ± 0.11	Not sig.
Overall	$37\% \pm 4\%$	$22\% \pm 3\%$	$18\% \pm 3\%$	0.32 ± 0.04	$p < 0.001$

educational intervention affects different aspects of students' problem-solving performance.

B. Calibrating grading

This rubric is not meant to replace normal grading practices. In grading, an instructor may want to reward a particular behavior, such as solving equations algebraically before putting in numbers, to encourage students to practice it. The rubric, on the other hand, measures characteristics of expertlike problem-solving behavior established in the literature that are independent of any specific pedagogy. A comparison of the overall rubric score with routine grading scores, however, can show whether grading practices reinforce the development of expertlike problem-solving behavior. It can also be used to indicate the quality of grading when many people are involved in the grading process. For example, in our studies, there were a few instances where a TA's grading of a problem did not correlate highly with the rubric score. In all of those cases, a detailed examination of the students' papers showed that the problem was graded inconsistently or incorrectly. Finally, a detailed look at the correlations of category scores with grading (e.g., as shown in Table III) might reveal if an instructor's emphases and biases, whether conscious or unconscious, influence their evaluation of how students progress toward expertise.

C. Generating and selecting problems useful for assessment

Although the rubric has been successfully applied to problems in many different formats spanning the full range of topics in most introductory physics courses, there are some characteristics of problems that can impede the assessment of student problem-solving skills. For example, explicit prompts to carry out certain processes such as drawing a free-body diagram or checking one's work mask aspects of a student's natural problem-solving processes, resulting in NA scores. Questions that break a problem into parts to guide student thinking also impede the assessment of problem-solving skills by generating NA scores, especially in Logical Progression category. In addition, if student solutions are found to produce near maximal scores across an entire class, the question may be an exercise, rather than a problem, for that group of students. Using the rubric can provide instructors with information about how to pose problems that require students to demonstrate the decision making that indicates expert problem solving.

D. Probing for specific areas of student difficulty

Because the rubric gives scores in five categories, it provides significantly more information than a single grade. This information can be used both for student feedback and to indicate to the instructor where there is need for coaching. For example, in one of the studies described above, several students in the class received low scores of 1 or 2 for Specific Application of Physics, but received relatively high scores of 4 and 5 for the Physics Approach and Mathematical Procedures. Such a pattern suggests that, although those students could recognize the physics principles needed to solve a problem and had the mathematical skill to do so, they were unable to apply those principles correctly to the specific situation. In that case, additional instruction could be targeted to the decision-making needed for applying physics principles. Examining a given student's scores across several problems can indicate if there are specific underlying difficulties for that particular student linked to a particular category. On the other hand, examining the category scores of many students for a given problem, or a set of problems on a common topic, might indicate that specific additional instruction might be useful for the whole class.

VI. SUMMARY

In summary, we have developed an instrument in the form of a rubric for assessing written solutions to physics problems along five almost independent axes: summarizing problem information into a Useful Description, deciding on an appropriate Physics Approach based on principles, making a Specific Application of Physics to the conditions in a particular situation, following appropriate Mathematical Procedures, and having an organized, goal-oriented Logical Progression that guides the solution process. Affective qualities such as motivation, interest, and beliefs about physics that are not usually evident from written work are not assessed by this rubric.

Our intent was to develop an instrument that could help assess and quantify students' problem-solving performance with less effort than traditional assessments such as think-aloud interviews, but provide more detailed information than grades, time required to solve problems, or number of mistakes made. With its five categories, the MAPS rubric allows the assessment of multiple aspects of expertlike problem-solving performance for large numbers of students with a reasonable amount of effort. Furthermore, this instrument allows authentic assessment for situations that usually occur in classes such as free-response tests or homework.

Multiple studies of the instrument's behavior indicated that the rubric categories were consistent with both the research literature and the processes students engage in while solving problems. These studies also provided evidence that the rubric and its score interpretations provide a valid, reliable, and useful assessment instrument.

The rubric was applicable to a range of physics topics in introductory university physics courses (mechanics and electricity and magnetism) and a variety of problem types, ranging from those commonly found in textbooks to context-rich problems. Scores on the instrument were highly correlated with independent measures such as the grading of TAs who had been introduced to the research literature on the differences between expert and novice

problem-solving behavior. The rubric also provided additional information that could be used to focus coaching or for modifying problems. Training on using the rubric based on only written documentation resulted in an overall reliability that was statistically significant in a variety of situations and adequate to inform instructor pedagogy. An iterative process of discussion among raters achieved the level of consistency needed for research purposes.

ACKNOWLEDGMENTS

This work was partially supported by the University of Minnesota and the National Science Foundation under Grants No. DUE-0715615 and No. DUE-1226197. We wish to thank the School of Physics and Astronomy of the University of Minnesota for its cooperation in this study and the physics faculty, graduate students, and undergraduates who participated.

[1] National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future* (National Academies Press, Washington, DC, 2007).

[2] National Research Council, *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century* (National Academies Press, Washington, DC, 2012).

[3] J. P. Smith III, A. A. diSessa, and J. Roschelle, Misconceptions reconceived: A constructivist analysis of knowledge in transition, J. Learn. Sci. **3**, 115 (1994).

[4] *Beyond 2000: Science Education for the Future*, edited by J. Millar and J. F. Osborne (King's College, London, 1998).

[5] J. P. Mestre, Probing adults' conceptual understanding and transfer of learning via problem posing, J. Appl. Dev. Psychol. **23**, 9 (2002).

[6] A. Van Heuvelen and D. P. Maloney, Playing physics jeopardy, Am. J. Phys. **67**, 252 (1999).

[7] D. P. Maloney and A. W. Friedel, Ranking tasks revisited, J. Coll. Sci. Teach. **25**, 205 (1996).

[8] P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups, Am. J. Phys. **60**, 637 (1992).

[9] A. Van Heuvelen, Learning to think like a physicist: A review of research-based strategies, Am. J. Phys. **59**, 891 (1991).

[10] K. VanLehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill, The Andes physics tutoring system: Lessons learned, Int. J. Artif. Intell. Educ. **15**, 147 (2005).

[11] P. Heller, R. Keith, and S. Anderson, Teaching problem solving through cooperative grouping. Part 1: Groups versus individual problem solving, Am. J. Phys. **60**, 627 (1992).

[12] A. Van Heuvelen, Overview, case study physics, Am. J. Phys. **59**, 898 (1991).

[13] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (American Educational Research Association, Washington, DC, 2014).

[14] S. Messick, Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning, Am. Psychol. **50**, 741 (1995).

[15] J. Docktor, Development and Validation of a Physics Problem-Solving Assessment Rubric, Ph.D. thesis, University of Minnesota, Twin Cities, 2009, http://www.compadre.org/per/items/detail.cfm?ID=9443.

[16] A. Newell and H. A. Simon, *Human Problem Solving* (Prentice-Hall, Englewood Cliffs, NJ, 1972).

[17] M. E. Martinez, What is problem solving?, Phi Delta Kappan **79**, 605 (1998).

[18] A. H. Schoenfeld, *Mathematical Problem Solving* (Academic Press, Orlando, FL, 1985).

[19] D. R. Woods, An evidence-based strategy for problem solving. J. Eng. Educ. **89**, 443 (2000).

[20] D. P. Maloney, in *Getting Started in PER, Reviews in PER Vol. 2*, edited by C. Henderson and K. A. Harper (American Association of Physics Teachers, College Park, MD, 2011), http://www.per-central.org/items/detail.cfm?ID=11457.

[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] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).

[23] B.-S. Eylon and F. Reif, Effects of knowledge organization on task performance, Cognit. Instr. **1**, 5 (1984).

[24] J. H. Larkin and F. Reif, Understanding and teaching problem solving in physics, Eur. J. Sci. Educ. **1**, 191 (1979).

[25] M. T. H. Chi, R. Glaser, and E. Rees, Expertise in problem solving, in *Advances in the Psychology of Human Intelligence vol. 1*, edited by R. J. Sternberg (Lawrence Erlbaum Associates, Hillsdale, NJ, 1982), pp. 7–75.

[26] J. H. Larkin, J. McDermott, D. Simon, and H. A. Simon, Expert and novice performance in solving physics problems, Science **208**, 1335 (1980).

[27] F. Reif and J. I. Heller, Knowledge structure and problem solving in physics, Educ. Psychol. **17**, 102 (1982).

[28] E. Bagno and B.-S. Eylon, From problem solving to knowledge structure: An example from the domain of electromagnetism, Am. J. Phys. **65**, 726 (1997).

[29] R. Elio and P. B. Scharf, Modeling novice-to-expert shifts in problem-solving strategy and knowledge organization, Cogn. Sci. **14**, 579 (1990).

[30] K. Cummings, J. Marx, R. Thornton, and D. Kuhl, Evaluating innovation in studio physics, Am. J. Phys. **67**, S38 (1999).

[31] J. L. Docktor, N. E. Strand, J. P. Mestre, and B. H. Ross, Conceptual problem solving in high school physics, Phys. Rev. ST Phys. Educ. Res. **11**, 020106 (2015).

[32] J. L. Docktor and K. Heller, Assessment of student problem solving processes, in *Proceedings of the Physics Education Research Conference 2009*, edited by M. Sabella, C. Henderson, and C. Singh (AIP Conference Proceedings 1179, Melville, NY, 2009), pp. 133–136.

[33] G. Taasoobshirazi and J. Farley, Construct validation of the physics metacognition inventory, Int. J. Sci. Educ. **35**, 447 (2013).

[34] J. P. Mestre, J. L. Docktor, N. E. Strand, and B. H. Ross, Conceptual problem solving in physics, in *Cognition in Education*, edited by J. P. Mestre and B. H. Ross (Academic Press, New York, 2011), pp. 269–298.

[35] W. K. Adams and C. E. Wieman, Problem solving skill evaluation instrument–Validation studies, in *Proceedings of the Physics Education Research Conference 2006*, edited by L. McCullough, L. Hsu, and P. Heron (AIP Conference Proceedings 883, Melville, NY, 2007), pp. 18–21; See also W. K. Adams, Development of a problem solving evaluation instrument; Untangling of specific problem solving skills, Ph.D. thesis, University of Colorado, 2009, http://www.compadre.org/per/items/detail.cfm?ID=9674.

[36] Y.-J. Lee, D. J. Palazzo, R. Warnakulasooriya, and D. E. Pritchard, Measuring student learning with item response theory, Phys. Rev. ST Phys. Educ. Res. **4**, 010102 (2008).

[37] J. Marx and K. Cummings, Development of a Survey Instrument to Gauge Students' Problem-Solving Abilities, in *Proceedings of the Physics Education Research Conference 2010*, edited by C. Singh, M. Sabella, and S. Rebello (AIP Conference Proceedings 1289, Melville, NY, 2010), pp. 221–224.

[38] R. Warnakulasooriya, D. J. Palazzo, and D. E. Pritchard, Time to completion of web-based physics problems with tutoring, J. Exp. Anal. Behav. **88**, 103 (2007).

[39] C. Henderson, E. Yerushalmi, V. Kuo, P. Heller, and K. Heller, Grading student problem solutions: The challenge of sending a consistent message, Am. J. Phys. **72**, 164 (2004).

[40] M. W. van Someren, Y. F. Barnard, and J. A. C. Sandberg, *The Think Aloud Method: A Practical Guide to Modeling Cognitive Processes* (Academic Press, San Diego, CA, 1994).

[41] A. Jonsson and G. Svingby, The use of scoring rubrics: Reliability, validity and educational consequences, Educ. Res. Rev. **2**, 130 (2007).

[42] C. A. Mertler, Designing scoring rubrics for your classroom, Prac. Assess., Res. Eval. **7** (2001).

[43] D. Huffman, Effect of explicit problem solving instruction on high school students' problem-solving performance and conceptual understanding of physics, J. Res. Sci. Teach. **34**, 551 (1997).

[44] J. M. Blue, Sex differences in physics learning and evaluations in an introductory course, Ph.D. thesis, University of Minnesota, Twin Cities, 1997, http://www.compadre.org/per/items/detail.cfm?ID=12324.

[45] T. Foster, The development of students' problem-solving skills from instruction emphasizing qualitative problem-solving, Ph.D. thesis, University of Minnesota, Twin Cities, 2000, http://www.compadre.org/per/items/detail.cfm?ID=4766.

[46] K. A. Harper, Investigating the development of problem solving skills during a freshman physics sequence, Ph.D. thesis, The Ohio State University, 2001.

[47] S. Murthy, Peer-assessment of homework using rubrics, in *Proceedings of the Physics Education Research Conference 2007*, edited by L. Hsu, C. Henderson, and L. McCullough (AIP Conference Proceedings 951, Melville, NY, 2007), pp. 156–159.

[48] C. A. Ogilvie, Moving students from simple to complex problem solving, in *Learning to solve complex scientific problems*, edited by D. H. Jonassen (Lawrence Erlbaum Associates, New York, 2007), pp. 159–185.

[49] A. Mason and C. Singh, Do advanced physics students learn from their mistakes without explicit intervention?, Am. J. Phys. **78**, 760 (2010).

[50] E. Yerushalmi, E. Cohen, A. Mason, and C. Singh, What do students do when asked to diagnose their mistakes? Does it help them? I. An atypical quiz context, Phys. Rev. ST Phys. Educ. Res. **8**, 020109 (2012).

[51] F. Reif, J. H. Larkin, and G. Brackett, Teaching general learning and problem solving skills, Am. J. Phys. **44**, 212 (1976).

[52] J. R. Hayes, *The Complete Problem Solver*, 2nd ed. (Lawrence Erlbaum Associates, Hillsdale, NJ, 1989).

[53] J. H. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Models of competence in solving physics problems, Cogn. Sci. **4**, 317 (1980).

[54] G. Pólya, *How to Solve It* (Princeton University Press, Princeton, NJ, 1945).

[55] M. T. H. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices, Cogn. Sci. **5**, 121 (1981).

[56] W. J. Leonard, R. J. Dufresne, and J. P. Mestre, Using qualitative problem-solving strategies to highlight the role of conceptual knowledge in solving problems, Am. J. Phys. **64**, 1495 (1996).

[57] J. I. Heller and F. Reif, Prescribing effective human problem-solving processes: Problem description in physics, Cognit. Instr. **1**, 177 (1984).

[58] E. Yerushalmi, C. Henderson, K. Heller, P. Heller, and V. Kuo, Physics faculty beliefs and values about the teaching and learning of problem solving. I. Mapping the common core, Phys. Rev. ST Phys. Educ. Res. **3**, 020109 (2007).

[59] F. Lawrenz, P. Heller, R. Keith, and K. Heller, Training the teaching assistant: Matchingta strengths and capabilities to meet specific program goals, J. Coll. Sci. Teach. **22**, 106 (1992).

[60] D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 5th ed. (John Wiley & Sons, New York, 1997).

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

[62] J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, Biometrics **33**, 159 (1977).