Development and validation of a pedagogy-specific problem-solving process rubric

J. Christopher Moore[1] and Taylor Crouch[1]

[1]*Department of Physics, University of Nebraska Omaha, 6001 Dodge St., Omaha, NE, 68182, USA*

We have begun the development and validation of a rubric for the assessment of problem-solving process in introductory physics courses. The initial rubric consisted of 12 criteria based on research in expert-like problem solving practice and aspects of Cooperative Group Problem Solving (CGPS) pedagogy. In contrast to recent work on problem-solving assessment for use in research and curriculum development, this rubric was specifically designed for instructor use in the assignment of grades and for student use as a scaffold. After assessment of seven problems across content in motion and force, exploratory factor analysis identified 3 factors that we have categorized as: (1) framing, (2) physics formalism, and (3) planning and executing. These factors roughly align with our initial theory of the construct, suggesting evidence for criterion-related validity. Tau-equivalent reliability (N = 256) was found to be 0.80, and inter-rater reliability was high.

I. INTRODUCTION

Problem-solving is an important component of the introductory physics course, and instruction in problem-solving produces benefits outside of the classroom [1]. Physics education research teams have identified different problem types, frameworks for instruction, and have designed, implemented, and tested multiple curricular components specifically designed to improve problem-solving ability [2–5].

Heller *et al*. have found that although most faculty believe reflective introspection to be primarily how students learn problem-solving, their instructional goals and methods rarely match these beliefs [6]. This has led to the creation of new effective pedagogies designed to improve student problem-solving ability based on the theoretical framework of cognitive apprenticeship [7, 8]. In particular, the cooperative group problem solving (CGPS) pedagogy developed by the Physics Education Research Group at the University of Minnesota is based on an explicit, five-step problem-solving strategy [9]. Working within groups on context-rich, multiple-stage problems allows for integration of the conceptual aspects of the physics content with general problem-solving process.

This specific pedagogy focuses on modeling expert-like process, coaching and scaffolding of group problem-solving, and student articulation through written solutions. However, one of the key requirements for effective cognitive apprenticeship is introspective reflection [10]. As a metacognitive process, reflection is difficult to "assign" since it is ultimately a mental action that must be done by the student [11]. Furthermore, reflection and general process abilities are difficult to assess compared to content knowledge or "correct" answers, which has led to a dearth of validated, instructor-ready problem-solving process assessment instruments.

Several research groups have been developing assessment instruments for physics-specific problem-solving. Cummings and Marx have designed a survey instrument to assess student problem-solving of "textbook" problems in introductory physics [12, 13]. Mason and Singh have developed a survey of student attitudes and approaches to problem solving [14]. More recently, Docktor has presented the Minnesota Assessment of Problem-Solving (MAPS) rubric that can be used to distinguish novice and expert problem-solving performance in authentic classwork [15].

Although these instruments are excellent tools for researchers and curriculum developers, they are not designed to be used by individual instructors for the assignment of grades on problem-solving tasks. We have begun the development and validation of a problem-solving process rubric that is pedagogy-specific and designed for classroom use. Within the CGPS pedagogical framework, the rubric can be used to scaffold problem-solving, serve as reflective activity for the student, and provide a grading schema for that aspect of the course.

II. DESIGN CRITERIA

Our goal was to develop a rubric that could be used by students as a scaffold for problem solving, and by instructors as a consistent grading framework. The rubric needed to fit the following criteria:

1. It needed to be flexible across content.
2. The individual criteria needed to fit CGPS pedagogy, and therefore serve as a scaffold and coach for student use.
3. It needed to be valid and reliable with little training.

The context in which this rubric was developed is that of a small physics department in a Midwest comprehensive university with no graduate program, and therefore no graduate teaching assistants. Therefore, CGPS sessions are typically led by supervised undergraduate Learning Assistants (LAs). The rubric needed to be easy to use and clear, with minimal training time required to achieve reliability.

The individual criteria that make up the rubric also needed to conform to the specific pedagogy deployed, such that it could serve its purpose as a scaffold to student problem solving in CGPS sessions, homework, and exam problems. It also needed to be a consistent instrument that could be used across introductory courses, sections, and with multiple instructors to send a clear and consistent message to students about how

TABLE I. Problem-solving process factors with associated individual criteria.

Framing	**Physics Formalism**	**Planning & Execution**
Mental Model	Variable Definitions	Physics Descriptions
Determine the Question	Target Quantities	Logical Progression
Qualitative Approach	Quantitative Relationships	Unit Analysis
Physics Model/Diagram(s)		Quantitative Execution
		Reflection

problems will be graded [16]. To be clear, assessment of problem-solving "expertness" should focus only on reasoning. Therefore, our goal was not a general rubric for assessing "expert-like" problem-solving ability, since such a rubric already exists [15]. Our intention was to develop a rubric that breaks down problem-solving into individual sub-skills.

Finally, the rubric needed to be flexible across content within the introductory sequence, such that the student sees the same basic framework from one-dimensional motion through electromagnetism. This design criterion is more difficult to execute than it seems, since problems across content may necessitate very different approaches and epistemological resources [17]. It is primarily this criterion (along with logistical concerns) that has resulted in the delicate balance between forced-methodology through highly structured pedagogy and free, small-group apprenticeship.

Based on this design criteria we developed a scoring rubric that consisted of 12 criteria across 3 factors: framing, physics formalism, and planning and execution. Its initial design was based on the physics-specific problem-solving strategy outlined in Chapter 2 of Ref. [9]. Table I shows the three fundamental problem-solving factors along with their corresponding individual criteria. The full rubric is not shown due to limitations on space, but is available on request [18].

During framing, students create a mental model of the problem (we call this "everyday framing"), which we seek to get them to articulate through simple sketches without physics-specific formalism such as free-body diagrams (FBDs). They then explicitly state the question in informal language and suggest a basic qualitative approach. Only after this everyday framing is a specific physics model (such as a FBD) created. It is important to highlight that a physics model does necessarily include physics formalism.

The criteria for the physics formalism factor includes explicitly stating known values and unknown values, defining the target quantity, physics-specific sketches within physics models (such as vector diagrams), and explicitly defining quantitative relationships. The criteria within this factor is the most pedagogy-specific, and where features within student work will be the most variable across different content. For example, explicit definition of variables is not necessary to demonstrate expert-like problem solving, but is a valuable artifact to our instructors [15, 19].

Planning and execution includes using the formalisms and quantitative relationships to map out an algebraic solution in a logical manner. One aspect of our rubric requires algebraic solutions that are utilized in unit analysis before numerical values are assigned. This is a pedagogical decision and again is not a necessary component of expert-like problem solving. However, within the "Physics Descriptions" criteria we do attempt to assess communication between student's qualitative description and formal mathematical descriptions, which is an aspect of expert-like problem-solving not assessed by most rubrics [19].

Each of the 12 criteria is scored from 0-3, with numerical scores corresponding to "Missing," "Inadequate," "Needs Improvement," and "Adequate," respectively. This results in a maximum possible score of 36.

III. VALIDITY

By designing a rubric we have taken a construct (pedagogy-specific problem-solving process ability) and turned it into an operation (something someone does in the real world with real products) [20]. Determining the validity of the rubric means answering the following question: how well does the operationalization reflect the underlying construct [21]?

Translational validity is composed of two parts: face validity and content validity. For our purposes, face validity means determining whether or not the rubric seems like a measure of problem-solving ability. To establish content validity we must show how the criteria line up with established research on problem-solving ability. As discussed in the previous section, the rubric criteria where written to explicitly match descriptions of physics problem-solving processes and pedagogical approaches described in the research literature.

In the preliminary analysis of our rubric, we also have examined one aspect of criterion-related validity. Specifically, we looked at the rubric's predictive ability: does it predict something it should predict? In particular, the original rubric design was based on three factors each having multiple criteria. After deployment of the rubric and using it to score hundreds of student solutions to problems across content areas, we should be able to see these factors emerge from the data using the statistical technique of principal components factor analysis.

Over the course of eight weeks, the rubric was used to score $N = 256$ student solutions to problems covering content

TABLE II. Exploratory factor analysis. Oblimin rotation was used. N = 256 with no items or factors removed.

Criteria	Framing	Physics Formalism	Planning & Execution
Mental Model	0.624		
Determine Question	0.449		
Qualitative Approach	0.462		
Physics Model	0.403		
Target Quantities	0.326	0.415	
Variable Definitions		0.652	
Quantitative Relationships		0.477	0.341
Physics Descriptions			0.718
Logical Progression			0.776
Unit Analysis			0.466
Quantitative Execution			0.736
Reflection			0.480

in one-dimensional motion, two-dimensional motion, circular motion, forces, energy, and momentum. All student work was hand-written on structured problem-solving worksheets, digitized, and uploaded through a learning management system (LMS, Canvas).

A principal components factor analysis of the 12 items was conducted using varimax and oblimin rotations. Three factors were found to explain 41.8% of the variance. An oblimin rotation provided the best defined factor structure. All items in the analysis had primary loadings over 0.4, while two items had a cross-loading above 0.3 ("Target Quantities," and "Quantitative Relationships"). However, both cross-loading items had a strong primary loading of >0.4 with "Physics Formalism" and "Planning & Execution," respectively. Table II shows the exploratory factor loading matrix for the rubric. No items or factors were removed.

The three factor labels of "Framing," "Physics Formalism," and "Planning & Execution" proposed during the initial development of the rubric suited the extracted factors. Therefore, the rubric demonstrates reasonable predictive validity. Further work needs to be done to establish concurrent, convergent, and discriminant validity. Specifically, we are currently re-evaluating the problem solutions using the MAPS rubric to establish convergent validity and hopefully demonstrate a correlation between our grading schema and a validated measure of ability.

IV. RELIABILITY

Determining the reliability of the rubric means answering the following two questions: is this test consistent at measuring, and can we trust that the score means something? In preliminary analysis of our rubric, we have partially tested external reliability by examining inter-rater correlation. We have also tested internal reliability by examining tau-equivalent reliability and item-total correlation.

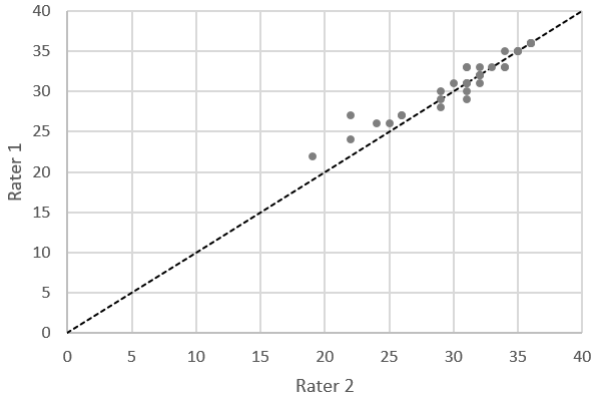

FIG. 1. Inter-rater reliability for N = 36 problem solutions. Rater 1 was an undergraduate Learning Assistant working in the CGPS sections. Rater 2 was a tenured faculty member in a physics department. The dotted line is a linear model with slope 1.0.

Figure 1 shows Rater 1 rubric scores for problem solutions as a function of Rater 2 scores on the same solutions. Rater 1 was an undergraduate LA working in the CGPS sessions and Rater 2 was a tenured faculty member in the same physics department. The problem content area was two-dimensional motion. Pearson's r was found to be 0.96, which indicates high inter-rater correlation but not necessarily inter-rater reliability. The dotted line shown in Figure 1 is a linear model with slope equal to 1.0. The coefficient of determination R^2 was 0.87. For low scores (<25), The undergraduate LA appears to over score compared to the faculty member.

The only explicit training received by the LA before scoring was reading Ref. [9] on the CGPS and attending and observing the same modeling instruction on rubric use provided to students in CGPS sessions. Therefore, the rubric appears to be reasonably reliable with little training. Continued work with more LAs is necessary.

TABLE III. Tau-equivalent reliability. N = 256.

N	mean	SD	tau-equivalent reliability, ρ_T
256	2.54	0.43	0.80

TABLE IV. Item-total correlation. N = 256.

Criteria	Item-total correlation
Mental Model	**0.319**
Determine Question	**0.238**
Qualitative Approach	0.412
Physics Model	0.457
Variable Definitions	**0.342**
Target Quantities	0.456
Quantitative Relationships	0.537
Physics Descriptions	0.607
Logical Progression	0.585
Unit Analysis	0.512
Quantitative Execution	0.480
Reflection	0.431

Tau-equivalent reliability (ρ_T) is a lower-bound estimate of the internal reliability of a measure. This measure of reliability can be viewed as the expected correlation of two tests that measure the same construct [22]. Table III shows the mean, standard deviation (SD), and tau-equivalent reliability for our rubric. A value of ρ_T = 0.80 is on the border between "acceptable" and "good" [20].

Table IV shows item-total correlations for each item. All items showed high item-total correlation (>0.4) except "Mental Model," "Determine Question," and "Variable Definitions," as highlighted (bolded). Preliminarily, it appears that "Mental Model" and "Determine Question" had lower item-total correlation because student's were about to do well on these criteria even if they had no idea how to proceed with the problem due to some content deficiency. Similarly, we have found many students can solve a problem, but fail to satisfy our pedagogical choice to explicitly define variables.

V. CONCLUSION

We have begun the development and validation of a rubric for the assessment of problem-solving process in introductory physics courses. This initial rubric consisted of 12 criteria based on research in expert-like problem solving practice and aspects of Cooperative Group Problem Solving (CGPS) pedagogy. In contrast to recent work on problem-solving assessment for use in research and curriculum development, this rubric was specifically designed for instructor use in the assignment of grades and for student use as a scaffold.

In particular, we had three design criteria: (1) flexibility, (2) pedagogy-specific, and (3) valid and reliable with little training. The reliability and validity of the measure was maintained across multiple raters, rater experience, and content areas, providing limited evidence that the rubric meets design criteria 1 and 3.

Three factors explain a significant amount of the variance in scores: framing, physics formalism, and planning and executing. These factors align with our initial theory of the construct, and also demonstrated the pedagogically-specific features we intended as specified in design criteria 2.

Significant work is still required. Specifically, concurrent, convergent, and discriminant validity need to be established. Furthermore, reliability across multiple raters from various backgrounds needs to be measured.

[1] National Academy of Sciences, *Rising Above the Gathering Storm* (National Academies Press, Washington DC, 2007).
[2] J.P. Mestre, J. Appl. Dev. Psychol. **23**, 9 (2002).
[3] P. Heller and M. Hollabaugh, Am. J. Phys. **60**, 637 (1992).
[4] K. VanLehn *et al.* Int. J. Artif. Intell. Educ. **15**, 147 (2005).
[5] P. Heller, R. Keith, and S. Anderson, Am. J. Phys. **60**, 627 (1992).
[6] P. Heller *et al.* in *Proceedings of the Physics Education Research Conference, Rochester, NY 2001*.
[7] J. Docktor and K. Heller, in *Proceedings of the Physics Education Research Conference, Ann Arbor, Michigan 2009*.
[8] T. Crouch and J.C. Moore, in *Proceedings of the Physics Education Research Conference, Washington, DC, 2018*.
[9] P. Heller and K. Heller, *Cooperative Group Problem Solving in Physics* (University of Minnesota, 1999).
[10] A. Collins, J. Brown, and A. Holum, Am. Educator **6** 38-46 (1991).
[11] A.J. Mason, Ph.D. thesis, University of Pittsburgh, 2009.
[12] K. Cummings and J. Marx, in *Proceedings of the Physics Education Research Conference, Portland, Oregon, 2010*.
[13] J. Marx and K. Cummings, in *Proceedings of the Physics Education Research Conference, Portland, Oregon, 2010*.
[14] A. Mason and C. Singh, Phys. Rev. ST Phys. Educ. Res. **6**(2), 020124 (2010).
[15] J. Docktor *et al.* Phys. Rev. PER **12**, 010130 (2016).
[16] C. Henderson *et al.* Am. J. Phys. 72, 164 (2004).
[17] D.M. Hammer, Am. J. Phys. **68**(S1) (2000).
[18] email: jcmoore@unomaha.edu
[19] M. Hull, E. Kuo, A. Gupta, and A. Elby, Phys. Rev. ST Phys. Educ. Res. **9**, 010105 (2013).
[20] P. Kline, *The handbook of psychological testing (2nd ed.)* (Routledge, London, 2000).
[21] W. Trochim, *The Research Methods Knowledge Base* (Atomic Dog Publishing, Cincinnati, 2000).
[22] J.C. Nunnally, in *Psychometric Theory (2nd ed.)* (McGraw-Hill, New York, 1978).