Reliability and Validity of an Introductory Physics Problem-Solving Grading Rubric

Kristen Rodenhausen and J. Christopher Moore

Physics Department, University of Nebraska Omaha, 6001 Dodge St., Omaha, NE, 68182, USA

We have developed and validated a rubric for the assessment and scaffolding of problem-solving process in introductory physics courses via an iterative approach. The current version of the rubric consists of eight 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. This means that the rubric can be used within group problem-solving activities as a student support, formative assessment of individual work, and summative assessment, such as exams. For this study, the rubric was used to score \( N = 166 \) student solutions to 6 individually-assigned homework problems covering content in introductory mechanics in a course enrolling 32 students. Inter-rater and re-rater reliability was high for undergraduate Learning Assistant raters receiving only moderate training (approximately 4 hours). Factor analysis identified two factors that have been categorized as: (1) framing & defining, and (2) planning & execution. These factors align with our initial theory of the construct, suggesting evidence for criterion-related validity. Tau-equivalent reliability was found to be 0.76, and an item-total correlations test demonstrated all criteria correlations consistent with averaged behavior.
I. INTRODUCTION

Physics education research teams have identified different problem types, frameworks for instruction, and have designed, implemented, and verified multiple pedagogical components specifically designed to improve problem-solving ability [1–4]. In particular, the cooperative group problem solving (CGPS) pedagogy developed by the Physics Education Research Group at the University of Minnesota contextualizes an explicit, five-step problem-solving strategy [5, 6]. In implementing this pedagogical design, conceptual understanding becomes integrated with general problem-solving process, and students learn how to make good problem-solving decisions.

Several research groups have developed assessment instruments for physics-specific problem-solving. Mason and Singh have developed a survey of student attitudes and approaches to problem solving [7], whereas Cummings and Marx have designed a survey instrument to assess student problem-solving of "textbook" problems in introductory physics courses [8, 9]. For assessment of authentic classwork, the Minnesota Assessment of Problem-Solving (MAPS) rubric can be used to evaluate student performance outside of a survey context, distinguishing between novice and expert problem-solving performance [10]. These instruments are excellent tools for researchers and curriculum developers; however, they are not designed to be used by individual instructors for the assignment of grades.

In this paper, we report on the continued development of a problem-solving rubric aligned with the CGPS pedagogy and appropriate for assignment of course grades. Additionally, the rubric is designed for student use as a model of expert-like processes, a scaffold for group problem-solving, and guidance for how to articulate problem-solving decisions through written solutions. We have previously reported on a 12 criteria grading rubric with the same goals. Version 1 of the rubric displayed varied levels of reliability and face validity, where rater-rater correlation was high but reliability was mixed [11]. For low scores, undergraduate LAs tended to over score compared to faculty member measures. Although internal validity measures were high, interviews revealed that there was some confusion concerning several items from both uninitiated scorers and students. Based on this initial work, a modified version of the rubric was developed consisting of 8 criteria. Here we report on the reliability and validity of the new version.

II. DESIGN CRITERIA

The individual criteria that make up the rubric needed to conform to the specific pedagogy deployed, such that it could serve its purpose as a scaffold to student problem solving in homework and exam problems. The entire rubric 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 problems will be graded [12]. Additionally, 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 [10]. Our intention was to develop a rubric that breaks down problem-solving into individual sub-skills.

Based on this design criteria we developed a scoring rubric that consisted of eight criteria across two factors: framing/defining and planning/execution. The initial design was based on the physics-specific problem-solving strategy outlined in Chapter 2 of Ref. [6]. Table I shows the fundamental problem-solving factors along with their corresponding individual criteria.

| Framing and Defining | Planning and Executing |
|----------------------|------------------------|
| General Principles   | Specific Math Representation |
| Physics Representation| Mathematical Plan       |
| Defined Variables    | Quantitative Execution  |
| General Math Representation | Reflection |

During framing, students suggest a basic qualitative approach and one or more general principles, such as force-acceleration or work-energy ("General Principles"). Only after this framing is a specific physics model including physics formalism created (such as a free-body diagram, called "Physics Representation"). The "Define Variables" criteria includes explicitly stating known values, unknown values, and defining the target quantity. The general equation (called the "General Math Representation") corresponding to the general principle is then stated explicitly. Explicit definition of variables is not necessary to demonstrate expert-like problem solving, but is a valuable artifact to instructors for grading and assignment of partial-credit, and it helps structure student thinking and decision making [10, 13]. Furthermore, requiring explicit articulation of general principles in a mathematical formalism (i.e. $\Sigma F = ma$) is also a pedagogical choice not necessarily required to demonstrate competency [14], but a choice shown to more rapidly improve student performance [15].

Planning and execution includes using the physics formalisms to plan a mathematical solution in a logical manner, which we highlight, does not necessarily mean a "linear" manner. The "Specific Math Representation" is the problem-specific application of the general representation (i.e. $N - mg = ma$). The rubric also requires analytical solutions before numerical values are assigned, referred to as a "Mathematical Plan," where students utilize one or more specific math representations to develop an analytical model. This is a pedagogical decision and, again, is not a necessary component of expert-like problem solving. However, we attempt to assess communication between students’ qualitative...
FIG. 1. (a) Rater 2 scores as a function of Rater 1 scores for $N = 166$ problem solutions. Both raters were undergraduate Learning Assistants. The red line is a linear model with slope equal to 1.0, the blue line is the linear regression, and the shaded region represents the confidence interval. Larger points indicate overlapping data. (b) Inter-rater reliability with the "Reflection" criterion excluded.

III. RELIABILITY

Figure 1(a) shows Rater 2 rubric scores for problem solutions as a function of Rater 1 scores on the same solutions. Both raters 1 and 2 were undergraduate LAs working in CGPS sessions. The red line shown in Figure 1(a) is a linear model with slope equal to 1.0. Ideally, all data would lie along this line, signifying that both Rater 1 and Rater 2 provided the same score. The blue line is the linear regression line, with shaded regions representing the confidence interval. The linear regression is shifted upward from and parallel to the ideal model, indicating that Rater 2 consistently over scored in comparison to Rater 1. On initial analysis, there was strong rater-rater correlation ($R^2 = .98$), but not strong rater-rater reliability due to deviation from the slope-1 model.

Throughout the semester, LAs held weekly meetings to review any needed clarifications and discuss reasoning for particular scores. One of the LAs (Rater 1) had experience using previous versions of the rubric and direct involvement in its development. The other LA (Rater 2) was new to both the LA role and rubric-based evaluation of students’ work.

The explicit training received by the LAs included discussion of the rubric criteria, observing a modeled example of rubric use and model solutions with scores, and whole-group post-grading discussions during the first two weeks. Total training time was approximately 4 hours. Over the course of the semester, LAs held weekly meetings to review any needed clarifications and discuss reasoning for particular scores. One of the LAs (Rater 1) had experience using previous versions of the rubric and direct involvement in its development. The other LA (Rater 2) was new to both the LA role and rubric-based evaluation of students’ work.

Over the course of 14 weeks the rubric was used by undergraduate Learning Assistants (LAs) to score $N = 166$ participant solutions to 6 different problems covering content in motion, force, momentum, and work/energy at a Midwestern research university. Participants were students in a first-semester, calculus-based physics course ($N = 32$), content selection was determined by the institution’s curriculum, and all solutions submitted by students were included in the study. All student submissions were hand-written on structured problem-solving worksheets, digitized, and uploaded through a learning management system (LMS, Canvas). The student population was primarily engineering and science majors in a calculus-based first-semester introductory physics course (64% white, 15% Hispanic, 6% Black, 25% woman).

The explicit training received by the LAs included discussion of the rubric criteria, observing a modeled example of rubric use and model solutions with scores, and whole-group post-grading discussions during the first two weeks. Total training time was approximately 4 hours. Over the course of

Each of the eight criteria is scored from 0-2, with numerical scores corresponding to "Missing," "Inadequate," "Needs Improvement," and "Adequate," respectively. This results in a maximum possible score of 16. The definitions for the levels of mastery for each criteria is not shown due to article length limitations, but is available at Ref. [16].
FIG. 2. Rater 1 scores for $N = 25$ solutions as a function of Rater 1 scores 6 months after initial grading. The red line is a linear model with slope equal to 1.0, the blue line is the linear regression, and the shaded region represents the confidence interval. Both strong correlation and fit to the model are present, indicating high re-rater reliability.

removal of this one criterion resulted in both stronger rater-rater correlation and rater-rater reliability, with the idealized model within the confidence interval.

Figure 2 shows Rater 1 problem scores as a function of Rater 1 re-scores completed 6 months after initial grading. After the semester had ended, Rater 1 regraded each assignment and compared the results to their initial scores. Both strong correlation and fit to the model are present, indicating high re-rater reliability with inclusion (shown) and exclusion (not shown) of the "Reflection" criterion. Similar re-rater results were observed for Rater 2.

TABLE II. Tau-equivalent reliability. $N = 166$.

| $N$ | mean | SD | tau-equivalent reliability, $\rho_T$ |
|-----|------|----|-----------------------------------|
| 166 | 13.89| 2.25| 0.76 |

Tau-equivalent reliability ($\rho_T$) is considered a lower-bound estimate of internal reliability of a measurement. This measure of reliability can be viewed as the expected correlation of two tests that measure the same construct [17]. Table II shows the mean, standard deviation (SD), and tau-equivalent reliability for our rubric. A value of $\rho_T = 0.76$ is on the border between "acceptable" and "good" according to Kline et al. [18].

The mean score across $N = 166$ problem solutions was 13.89 or 87%. For a research instrument an ideal mean is approximately 50%. However, for grading purposes a higher average is typically desired. Thus, a mean of 13.89 corresponds to a "B" grade in typical higher education grading schemes. This may still seem high, however the student population studied is composed predominantly of science, engineering, and math majors. Furthermore, problem solutions evaluated as part of this study came from formative homework assignments with group work encouraged.

An item-total correlation test is used to check if any criteria in a set is inconsistent with the averaged behaviour of the others. A test was performed on the previous 12 criteria version of the rubric, resulting in consolidation and/or discarding of criteria [15]. Table III shows item-total correlations for each of the modified rubric's 8 criteria. All criteria had item-total correlations (>0.4) consistent with averaged behavior, which is an improvement from the previous rubric version.

IV. VALIDITY

By designing a rubric we have taken a construct (pedagogy-specific problem-solving process ability) and turned it into an operation (application for physical action with real products) [18]. Determining the validity of the rubric means answering the following question: how well does the operationalization reflect the construct [19]?

Translational validity is composed of 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 within our pedagogical context. To establish content validity, the criteria must correlate with established research on problem-solving. The rubric criteria were written to explicitly match descriptions of physics problem-solving processes and pedagogical approaches described in the literature, as discussed in more detail elsewhere [11]. Further face validity was determined by survey of 18 high school and college physics instructors, similarly discussed in detail elsewhere [20].

We have also examined one aspect of criterion-related validity by investigating the rubric’s predictive ability: does it predict something it should predict? In particular, the rubric design was based on two factors each having multiple criteria. After deployment of the rubric and using it to score hundreds
TABLE IV. Exploratory factor analysis. Oblimin rotation was used. N = 166 with "Reflection" criterion removed.

| Criteria                        | Planning & Execution | Framing & Defining |
|---------------------------------|----------------------|--------------------|
| What is this Problem?           | 0.74                 |                    |
| Physics Representation          | 0.60                 |                    |
| Defined Variables               | 0.46                 |                    |
| General Mathematical Representation | 0.40                |                    |
| Specific Mathematical Representation | 0.55                |                    |
| Mathematical Plan               | 0.83                 |                    |
| Quantitative Execution          | 0.77                 |                    |

of student solutions to problems across content areas, these factors should emerge from the data using the statistical technique of factor analysis.

A factor analysis of the eight criteria was conducted using varimax and oblimin rotations using the statistical package Jamovi. Two 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 >0.4. Two factors were found to explain the variance: "Framing and Defining" and "Planning and Execution". Table IV shows the factor loading matrix for the rubric with removal of the "Reflection" criterion, as justified previously.

The two factor labels of "Framing & Defining" 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.

Interestingly, both the general mathematical representation and the specific mathematical representation criteria were sorted into the correct factors, where the general representation (i.e. $\Sigma F = ma$) is correlated to framing and the specific representation (i.e. $N - mg = ma$) is correlated to planning and execution. The "Framing & Defining" factor of the rubric forces students to coordinate between stated general principles, pictorial and/or graphical representations of the specific situation, and foundational and generalizable mathematical formulations of the principles. During planning, the rubric forces coordination between these framing components to develop a specific mathematical representation of the problem at hand. Although not required as an explicit, written step for documenting expert-like problem-solving ability, this pedagogical choice was made to scaffold student thinking and promote development and deployment of resources beyond algorithm recall and other novice approaches [14, 21].

V. CONCLUSION

We have continued the development, iteration, and validation of a rubric for the assessment and scaffolding of problem-solving process in introductory physics courses. The first version of the rubric consisted of 12 criteria based on research in expert-like problem-solving practice and aspects of CGPS pedagogy. Through an iterative process, the second version was developed and evaluated. Version 1 displayed varied levels of reliability and face validity, and although internal validity measures were high, interviews revealed that there was some confusion concerning several items. Based on this work, the total number of criteria were reduced and prompts were clarified. The area of concern for Version 2 was the "reflection" criterion, which caused a significant amount of the variance in scores between raters. After removal of this criterion, validity and reliability greatly improved. Adjustments regarding the clarity of the reflection criterion is an ongoing area of work for this team, as reflection on an answer is often the part of problem-solving valued by physics instructors.

It should also be noted that even when reliability between two raters is high within a research context, this still can mean variations in score as large as 15% (as shown in Fig. 1) for an individual student’s work. This is an unavoidable consequence of rubric-based grading, and should be considered by instructors when designing how they ultimately assign grades. Specifically, consistency in who serves as rater for summative work may be desirable, as Fig. 2 shows much lower variation (<6%).

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 across content, (2) pedagogy-specific, and (3) valid and reliable with little training. The reliability and validity of the rubric were proven to be acceptable and advanced from Version 1.

ACKNOWLEDGMENTS

Kristen Rodenhausen is an undergraduate Noyce Scholar funded by NSF DUE #2150788 and administratively supported by the University of Nebraska Omaha STEM TRAIL Center. This work was also supported by an endowment provided by Dr. George F. Haddix.
[1] J.P. Mestre, J. Appl. Dev. Psychol. 23, 9 (2002).
[2] P. Heller and M. Hollabaugh, Am. J. Phys. 60, 637 (1992).
[3] K. VanLehn et al. Int. J. Artif. Intell. Educ. 15, 147 (2005).
[4] P. Heller, R. Keith, and S. Anderson, Am. J. Phys. 60, 627 (1992).
[5] P. Heller et al. in Proceedings of the Physics Education Research Conference, Rochester, NY 2001.
[6] P. Heller and K. Heller, Cooperative Group Problem Solving in Physics (University of Minnesota, 1999).
[7] A. Mason and C. Singh, Phys. Rev. ST Phys. Educ. Res. 6(2), 020124 (2010).
[8] K. Cummings and J. Marx, in Proceedings of the Physics Education Research Conference, Portland, Oregon, 2010.
[9] J. Marx and K. Cummings, in Proceedings of the Physics Education Research Conference, Portland, Oregon, 2010.
[10] J. Docktor et al. Phys. Rev. PER 12, 010130 (2016).
[11] J.C. Moore and T. Crouch, in Proceedings of the Physics Education Research Conference, Washington, DC, 2019.
[12] C. Henderson et al. Am. J. Phys. 72, 164 (2004).
[13] M. Hull, E. Kuo, A. Gupta, and A. Elby, Phys. Rev. ST Phys. Educ. Res. 9, 010105 (2013).
[14] D.M. Hammer, Am. J. Phys. 68(S1) (2000).
[15] T. Crouch and J.C. Moore, in Proceedings of the Physics Education Research Conference, Washington, DC, 2019.
[16] https://www.creatingscientists.com/research/oapps
[17] J.C. Nunnally, in Psychometric Theory (2nd ed.) (McGraw-Hill, New York, 1978).
[18] P. Kline, The handbook of psychological testing (2nd ed.) (Routledge, London, 2000).
[19] W. Trochim, The Research Methods Knowledge Base (Atomic Dog Publishing, Cincinnati, 2000).
[20] A. Kruger, J.C. Moore, Poster, AAPT Winter Meeting, Houston, TX (January 2019).
[21] C.A. Randles and T.L. Overton, Chem. Educ. Res. Pract. 16, 811-823 (2015).