ROBUST ASSESSMENT INSTRUMENT FOR STUDENT PROBLEM SOLVING

Problem solving is a complex process that is important for all citizens in our modern world and crucial for learning physics. Although there is a great deal of effort to improve student problem solving throughout the educational system, there is no standard way to evaluate written problem solving that is valid, reliable, and easy to use. Typically complex processes such as problem solving are assessed by using a rubric, which divides a skill into multiple reasonably independent categories and defines criteria to attain a score in each. This paper describes the development and validation of a problem solving rubric for the purpose of assessing written solutions to physics problems.

Jennifer Docktor, University of Minnesota
Kenneth Heller, University of Minnesota

Introduction

Problem solving is one of the primary goals, teaching tools, and evaluation techniques of physics courses. Currently there is no standard way to measure the process of problem solving so that student progress in this domain can be assessed. Most tests of problem-solving performance given in the classroom focus on the correctness of the end result or partial results rather than the quality of the procedures and reasoning leading to the result, which gives an inadequate description of a student's skills (Schoenfeld, 1985). A more detailed and meaningful measure is necessary if different curricular materials or pedagogies are to be compared. This measurement tool could also allow instructors to diagnose student difficulties and focus their coaching. It is important that the instrument be applicable to any problem solving format used by a student and to a range of problem types and topics typically used by instructors.

A version of such an assessment instrument has been developed at the University of Minnesota in the form of a rubric, which subdivides the problem solving process into five approximately independent aspects and assigns a separate score for performance in each category. Those categories are: useful description, physics approach, application of physics, mathematical procedures, and logical progression. The rubric is based on descriptions of problem solving from cognitive psychology, mathematics, and physics (Hayes, 1989; Pólya, 1945; Reif & J. Heller, 1982; Van Heuvelen, 1991) in addition to research studies on the differences between expert and novice problem solving processes (Chi, Feltovich, & Glaser, 1981; Larkin, 1979; Larkin, McDermott, Simon, & Simon, 1980a).

Although scoring rubrics have been used in past problem solving research (Blue, 1997; Foster, 2000; P. Heller, Keith, & Anderson, 1992) these instruments are difficult to use and not extensively tested. This study builds upon those results to develop an easy to use problem solving assessment instrument and establish evidence for reliability, validity, and utility. *Reliability* in this context refers to the agreement of scores from multiple

raters. *Validity* refers to the degree to which score interpretations are supported by empirical evidence and theory (AERA, APA, NCME, 1999). This study will also necessarily develop documentation and training materials for potential users.

The primary questions that our study seeks to answer include: 1) Is it possible to create a general, easy-to-use problem solving assessment that is perceived by instructors as useful for evaluating written solutions to physics problems? 2) To what extent are scores on the problem solving assessment valid - supported by empirical evidence and theoretical arguments? 3) To what extent are scores on the problem solving assessment reliable – consistent across multiple raters?

Problem Solving Processes in Physics

A first step in developing an assessment instrument is to clearly define the construct or content knowledge it is intended to measure. Although there are many descriptions of the important features of problem solving in the literature, there is an agreement that problem solving is a process of decision making. This section briefly reviews common definitions of problem solving from cognitive science, mathematics, and physics. It also summarizes research studies on the processes used by experienced and inexperienced solvers within the domain of physics. These definitions and processes form the basis for the rubric's development, which will be further clarified in the descriptions of its categories.

What is Problem Solving?

Descriptions of problem-solving emphasize that it is a decision-making process that occurs when a solver is presented with a task for which they have no specific set of actions they can use to reach a solution (Newell & Simon, 1972). For example, Hayes (1989) defines the problem solving process in the following way:

> Whenever there is a gap between where you are now and where you want to be, and you don't know how to find a way to cross the gap, you have a problem. Solving a problem means finding an appropriate way to cross a gap. (p. xii)

Similarly, Martinez (1998) describes problem solving as "the process of moving toward a goal when the path to that goal is uncertain" (p. 605). In each of these definitions, problem solving depends on the solver's experience and perception of the task. What is considered a problem for one person may be a routine exercise for another person (Schoenfeld, 1985).

One of the early modern attempts to identify stages involved in the type of quantitative problem solving used in mathematics and science was by the mathematician Pólya (1945). In his first step *Understanding the Problem*, the solver summarizes known and unknown information, introduces suitable notation, and draws a figure. Next, in *Devising a Plan*, the solver uses their knowledge to plan how to connect the given data to the desired goal. Then in *Carrying out the Plan* the solver implements their plan by carrying out the necessary procedures to reach an answer while checking their work along the way. The final step is *Looking Back* or examining the result to check that it makes sense, and if possible using an alternate procedure to achieve the answer. Hayes (1989) expanded these actions to include a first step of recognizing the existence of a problem

and a final step of consolidating gains or explicitly considering what was learned from solving the problem and how it might be useful for solving future problems.

Expert-Novice Research

Information about problem solving processes and knowledge structures have been obtained from research studies comparing experienced or "expert" problem solvers to inexperienced or "novice" problem solvers. Many of these studies focused on the content of physics knowledge and its mental organization as a basis for explaining observed process differences. In most early studies, the experts were physicists and the novices were beginning physics students. Think-aloud protocols (Larkin, 1979; Larkin et al., 1980a) and card-sorting tasks (Chi et al., 1981) usually focused on solving standard textbook problems and were limited to a few topics in mechanics such as motion with constant acceleration, Newton's second law, or conservation of energy.

Process Differences

Several researchers observed that experts engage in a low-detail overview of problem features and expectations, called a qualitative analysis, before writing down quantitative relationships (Chi et al., 1981; Larkin, 1979; Larkin et al., 1980a). Experts use this information to consider possible solution approaches or physics principles that might be useful in solving the problem. Novices tend to skip this step and jump directly to writing down equations or miscellaneous mathematical relationships (Reif & J. Heller, 1982). Since many of the expert's problem solving processes have become automated, they tend to work forward with little explicit planning whereas novices tend to start from the unknown and work backward (Larkin et al., 1980a). Experts also have strong mathematical skills and strategies for monitoring progress and evaluating their answer (Larkin et al., 1980a; Reif & J. Heller, 1982).

Knowledge Organization Differences

From their observations, Larkin (1979) and Chi et al. (1981) also drew conclusions about the content and mental organization of physics knowledge. They found that an expert's memory is structured hierarchically around a small number of fundamental physical principles called "chunks". Such principles are considered fundamental because they can be applied to a wide range of physical situations (Larkin, 1981). Accessing a chunk also cues other useful relations and the procedures or actions to successfully apply those principles (Chi et al., 1981; Larkin et al., 1980a). In contrast, the novice's knowledge structures are disconnected and each relation must be accessed individually. There is no clear link between physics principles and application procedures. This mental organization makes the novice's solution search an inefficient and time-consuming process (Larkin, 1979).

Implications of Definitions and Expert-Novice Research

Several physicists have adapted problem solving definitions and the processes informed by expert-novice research together with observations of student problem solving actions to develop problem-solving strategies or frameworks for physics instruction (K. Heller, 2006; P. Heller et al., 1992; Reif, Larkin, & Brackett, 1976; Van Heuvelen, 1991). These

frameworks use writing to guide the student's use of an organized problem-solving strategy and make explicit the complex processes done implicitly by experts.

These frameworks typically subdivided the first step of understanding the problem (Pólya, 1945) to highlight the importance of multiple representations or problem descriptions in solving physics problems (J. Heller & Reif, 1984; Larkin, 1981; Larkin, McDermott, Simon, & Simon, 1980a, 1980b; Reif & J. Heller, 1982). In particular, J. Heller and Reif (1984) suggest that effective problem solvers first generate a "basic description" that summarizes the relevant information about the situation in symbolic, pictorial, and verbal forms prior to producing a "theoretical description" that contains abstracted diagrams specific to physics concepts and principles.

Although expert-novice research studies provide useful insight into physics problem-solving processes, they also have limitations (J. Heller & Reif, 1984). The physics topics used in the studies were not representative of the entire domain of physics, and the tasks were typically standard textbook-style quantitative problems. The expert-novice dichotomy does not consider intermediate stages in problem solving, such as progressing levels of competency (K. Heller, 2006). In addition, the problems were often "exercises" for the experts and might not reflect the processes engaged in for more difficult problems (Schoenfeld, 1985). Nonetheless, for assessment purposes it is important to consider the expert-like processes of qualitative descriptions, approaches based on fundamental physics principles, procedures for the appropriate application of principles, skilled use of mathematics, and strategies for monitoring progress and evaluating results.

Processes Assessed by the Minnesota Rubric

The process categories for the assessment rubric were based on the research literature in cognitive science, mathematics, and physics. They were developed within the constraints of being easy to interpret, independent of pedagogy, generalizable to multiple problem types and topics, and focused on written work. Many other related rubrics that have been developed to assess student problem solving in physics and other disciplines are available from a general search of the Web. Such rubrics are developed for classroom use to support a specific pedagogy and typically have not been extensively tested for reliability or validity. The rubric under development is based on research on student problem solving at University of Minnesota over many years (Blue, 1997; Foster, 2000; P. Heller et al., 1992). Although there are many similarities in the problem solving processes assessed by instruments in those studies, the current study differs by attempting to simplify the rubric and adding more extensive tests of reliability, validity, and utility. It explicitly considers applicability to a broad range of problem types and topics in physics and the ease of use for both research and instruction.

To make the rubric easy to use, it was constructed with as few dimensions as possible to still span most of the space that distinguishes novice and expert problem solving. The Minnesota rubric considers five problem-solving processes: organizing problem information into a useful description, selecting appropriate physics principles, applying physics to the specific conditions in the problem, using mathematical procedures appropriately, and the overall communication of an organized reasoning pattern.

Useful Description

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. It is similar to Pólya's (1945) stage of understanding the problem or Hayes' (1989) stage of representing the problem.

A problem description could include specifying known and unknown information, assigning appropriate symbols for quantities, stating a goal or target quantity, a sketch or picture of the physical situation, stating qualitative expectations, an abstracted physics diagram, drawing a graph, defining coordinate axes, and/or choosing a system. Unlike other models of problem solving (J. Heller & Reif, 1984; K. Heller, 2006; Van Heuvelen, 1991), this combines both a basic description and a physics-specific description into a single category. The term "description" was chosen to be consistent with other uses of the term (P. Heller et al., 1992; Reif et al., 1976) and avoid the multiple interpretations of the term "representation" (Hayes, 1989; Larkin et al., 1980a, 1980b). The useful description category differs from other instruments (Foster, 2000) by being assessed separately from the general physics approach.

Physics Approach

The *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 the general concept of vector or specific concepts such as momentum and velocity. The term "principle" is used to mean a fundamental physics rule or law used to describe objects and their interactions, such as conservation of energy or Newton's third law.

In addition to assessing the selection of a principle, this category also includes its basic understanding, such as the independent treatment of perpendicular components of vectors. This is similar to the evidence of conceptual understanding category outlined by P. Heller et al. (1992) and the general approach category used by Blue (1997) and Foster (2000).

The *Physics Approach* category reflects the expert-like process of selecting relevant physics principles before applying them to the specific context of the problem (Chi et al., 1981; Larkin et al., 1980b). Although several descriptions of problem-solving emphasize a stage of planning the solution (Hayes, 1989; P. Heller et al., 1992; Pólya, 1945), selecting important relations is a necessary first step in planning the solution (Leonard, Dufresne, & Mestre, 1996; Reif et al., 1976). In addition, the details of planning are difficult to assess because students often do not write down the steps of their solution plan unless explicitly instructed to do so. The planning process is implicitly addressed by this rubric in its other categories.

Specific Application of Physics

Specific Application of Physics assesses the solver's process of applying physics concepts and principles to the specific conditions in the problem. Specific application often involves connecting the objects and quantities in the problem to the appropriate terms in specific physics relationships. It can include a statement of definitions, relationships

between quantities, initial conditions, and 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 problem solving strategies (Leonard et al., 1996) and other assessments of problem solving (Blue, 1997; Foster, 2000). Writing down specific physics relationships, typically in the form of equations, can be seen as another aspect of planning the solution (P. Heller et al., 1992; Reif et al., 1976). This category is similar to the problem-solving model by Larkin et al. (1980b) that designates "connecting symbols in an equation with information in the problem" as a process that follows "selecting relevant physics principles" and "generating the corresponding equation" (p. 323).

Mathematical Procedures

Mathematical Procedures assesses the solver's process of executing the solution with respect to selecting appropriate mathematical procedures and following mathematical rules to obtain target quantities. Examples of these procedures include: isolate and reduce strategies from algebra, substitution, use of the quadratic formula, matrix operations, or "guess and check" from 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 identities.

This category corresponds to carrying out the plan (Hayes, 1989; Pólya, 1945) or the plan implementation process (Reif et al., 1976). It also corresponds to Van Heuvelen's (1991) "math representation" (p. 901) and Larkin et al.'s (1981b) "solving equations" function (p. 323). It is consistent with other assessments of appropriate mathematics (Blue, 1997; Foster, 2000; P. Heller et al., 1992) but differs in that it doesn't require students to solve equations symbolically to receive the highest score.

Logical Progression

Logical Progression assesses the solver's processes of communicating reasoning, staying focused toward a goal, and evaluating the solution for consistency. The category checks whether the overall problem solution is clear, focused, and organized logically. The term "logical" means that the solution is coherent (the solution order and solver's reasoning can be understood from what is written), internally consistent (parts do not contradict), and externally consistent (results agree with qualitative physics expectations).

This category agrees with the problem-solving assessment by Reif and J. Heller (1982) that includes clear interpretation or specification of parameters, completeness of the answer, internal logical consistency of the argument, external consistency of relationships and the magnitude of values, and optimality or the simplicity of the solution. It also emphasizes the importance of "the ability to provide coherent explanations" in science and engineering careers (Leonard et al., 1996, p. 1502). The term logical progression is taken from earlier assessments of problem solving (Blue, 1997; Foster, 2000; P. Heller et al., 1992) but it differs from those measures in that it doesn't score the student's process as working forwards or working backwards.

Several models of problem solving emphasize the final stage as looking back (Pólya, 1945) or evaluating the solution to check that it makes sense (Reif et al., 1976; Van Heuvelen, 1991). The logical progression does not require an explicit evaluation of the solution because students often skip this step unless explicitly instructed to do so, and the rubric is intended to be independent of strategy-modeling instructional techniques. However, steps such as planning and evaluation or checking the result could help a student avoid errors in consistency and coherence, which are scored as part of the logical progression.

Processes Excluded From the Rubric

To make the rubric as independent of specific pedagogy and as easy to use as possible, the metacognitive processes of planning and evaluating the answer are not explicitly assessed by the rubric. Although they are excluded as specific criteria from this rubric, planning and evaluation are implicitly assessed by the several other categories because these processes affect the overall coherence and consistency of the solution. Other aspects of problem-solving not assessed by the rubric include affective qualities such as motivation, interest, and beliefs about physics. These qualities are not usually evident from written work.

Scores on the Assessment Rubric

The current version of the rubric under development is given in the appendix. Scores on the rubric range from 0-5 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 measured by the problem usually because those decisions were not required. For example, if a description was provided in the problem statement or was not really 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 not be necessary for the solver to write down. This occurs for students who were generally successful in solving the problem without writing down all of their internal processes, such as a description or explicitly stating a physics approach. These "not applicable" scores are included because the rubric needs to recognize the possibility that students are beginning to develop some of the automated processes engaged in by experts (J. Heller & Reif, 1984).

To promote ease of use, the language of the score descriptions for each category is consistent. A score of 0 means that there is no evidence of the category and it was necessary for the solver, 1 means the category evidence was entirely inappropriate, 2 means mostly inappropriate or missing, 3 means parts are inappropriate or missing, 4 designates minor omissions or errors, and 5 is complete and appropriate.

When scoring written solutions to physics problems, it is important to consider only what is written and avoid the tendency to assume missing or unclear thought processes are correct (Henderson, Yerushalmi, Kuo, P. Heller, & K. Heller, 2004). Similarly, it is important not to overly emphasize the amount of detail in student explanations.

Methodology

Studies of the rubric's use for measuring written problem solving processes are founded on the concepts of validity, reliability, and utility. Current definitions of validity have shifted from outlining different "types" of validity to a more holistic view with multiple sources of validity evidence (Messick, 1995). This section describes each source and provides examples of validity tests for the rubric scores.

Validity, Reliability, and Utility

Validity is defined as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, NCME, 1999, p. 9). It is concerned with determining the appropriateness, meaningfulness, and usefulness of scores (Messick, 1995). Multiple sources of evidence contribute to a validity argument, including evidence based on content relevance and representativeness, response processes, internal and external measures, generalizability, and consequences of testing (AERA et al., 1999; Messick, 1995). Reliability in this context refers to the consistency or agreement of scores on the assessment, and utility refers to the perceived usefulness of the assessment by instructors.

Validity Evidence Based on Content

Content refers to the wording and formatting of items on an assessment, in addition to the documented procedures for scoring. (Messick, 1995). In this study content is interpreted to mean the process categories being assessed by the rubric and the documentation materials for potential users. Evidence for the relevance and representativeness of content comes from expert judgment, and theoretical descriptions of a domain in the research literature (AERA, et al., 1999).

Validity Evidence Based on Response Processes

An important validity consideration is the extent to which the assessment represents processes actually engaged in by the person(s) being assessed (AERA et al., 1999). It is also important to consider whether the interpretations of scores by judges or raters are consisted with the developer's intentions. In this study, student response processes are explored using both written work (written physics tests) and verbal problem-solving interviews. The responses of raters are compared to problem solving grades from instructors and from their feedback while using the rubric to determine the degree of consistency with the rubric developers' intentions.

Validity Evidence Based on Internal and External Structure

Internal structure refers to the extent to which relationships among parts of the instrument agree with expectations (AERA et al., 1999). External structure refers to the extent to which scores are related to other measures of the same construct or other hypothesized relationships. In this study, the degree of independence of the process categories are determined from statistical measures. For example, past research (Foster, 2000) indicated that the approach and application were correlated. The external structure of the rubric is evaluated by comparing rubric scores for written physics tests to scores assigned by a

grader. It is also evaluated from comparisons of the verbal responses from problem-solving interviews to solutions written on paper during the interview.

Validity Evidence for Generalizability

Although not explicitly included in all descriptions of validity evidence, Messick (1995) highlights the importance of an assessment being general across different populations and contexts. In this study, the rubric is tested on a variety of physics problem solutions that span different topics in standard introductory university physics courses from both mid-term tests and final exams. It is also tested on different types of problems, including those that are similar to traditional textbook problems and those that are context-rich (P. Heller et al., 1992).

Validity Evidence Based on Consequences of Testing

Descriptions of this source of validity highlight the importance of considering both intended and unintended consequences of score interpretations (AERA et al., 1999). In this study, the purposes of the rubric will be clearly outlined in the documentation materials and training. For example, in validity studies the rubric scores might only be meaningful to assess the performance of a class, and might not be meaningful or valid for diagnosing an individual student. A full study of the consequences of using this rubric, once developed, will be the subject of further work.

Reliability Evidence

Reliability refers to the agreement of scores from multiple raters or judges. In this study, reliability is measured from a study with graduate students who undergo a brief written training in use of the rubric. These graduate students are experienced in grading the work of introductory physics students. Their responses are compared to each other and to two expert raters. A quantitative measure of reliability is obtained from percentage of perfect agreement, agreement within one score, and Cohen's Weighted Kappa (Cohen, 1968) which accounts for the degree of difference in scores.

Utility or Usefulness

Evidence for the usefulness of an assessment includes its acceptance by instructors, the extent to which it can distinguish between experts and novices, and the extent to which it can distinguish between different classroom practices.. It is important that researchers, curriculum developers, and instructors are interested in the information obtained from administering an assessment. In this study, interpretations of scores from analyses of written work will be used to propose uses of the assessment from the perspectives of researchers, curriculum developers, and physics instructors.

Studies of Validity, Reliability, and Utility

This section outlines four major stages in testing validity, reliability, and utility of the rubric scores. After developing a draft instrument based on previous research as outlined above, preliminary studies with two raters were used to determine reliability measures and modify categories. Utility was also tested by comparing instructor and student solutions. Next, a study with graduate students involving a brief written training exercise was used to further measure the reliability and validity of the rubric's content including

training materials. An analysis of students' written solutions to physics tests from a semester of introductory physics (mechanics) from a variety of instructors was used to obtain evidence for response processes, generalizability, internal and external structure, and to propose uses of the rubric. Analysis of student interviews (in progress) is used to obtain further evidence of response processes and structural measures.

Preliminary Studies of Reliability and Utility

Following the rubric's initial development, it was used by two raters (one researcher and one high school teacher) to score final exam problem solutions from introductory university physics courses. A total of eight different problems were scored; five were from a calculus-based mechanics course for science and engineering and three problems were from the algebra-based mechanics course. Twenty solutions were randomly selected for each problem (out of approximately 200) that were legible and reflected a range of detail and quality. Interpretation of the rubric was discussed by the raters after independently scoring each problem.

Scores on all 160 solutions were used to determine the agreement of the two raters. Without any explicit training, the percent exact agreement in each of the five categories ranged from 61% to 75% with an average of 67%. Agreement within one score (excluding NA scores) was above 96% in every category. The categories with lowest agreement were Logical Progression and Specific Application of Physics. The category with highest agreement was Useful Description.

In addition, a preliminary study was conducted to determine the rubric's utility for distinguishing instructor or "expert" solutions from student solutions. Two problems were selected randomly from each of 38 chapters in a popular calculus-based physics textbook (N=76), and the solutions printed in the instructor solution manual were scored with the rubric. The solutions were typically very sparse and did not include much reasoning. Then, homework solutions hand-written by a physics instructor for an entire introductory physics course (N=83) were scored with the rubric. These solutions were more detailed and included steps of the reasoning process. The frequency of rubric scores was very similar for the instructor solution manual and the instructor, regardless of the level of detail. Most rubric scores for instructors were the highest possible value or a not applicable score. In comparison, scores of student solutions spanned the entire range of rubric scores. From the differences in score frequencies it was easy to distinguish between the instructor and student solutions.

Study on Training Raters

After the preliminary studies, the rubric was tested with eight physics graduate students who had experience in grading student test solutions. These graduate students were at least in their third year of graduate school. This information was used to check reliability, interpretations of rubric scores, and to obtain feedback on the content of the rubric. The graduate student volunteers were solicited by e-mail and randomly assigned to two groups. Four people scored student solutions from a mechanics final exam problem and four people scored student solutions from an electricity and magnetism (E&M) final exam. Graduate students were provided with the problem statement, a solution to the problem, a copy of the rubric, brief definitions of each category on the rubric, a blank

scoring template table, a set of student solutions, and an instruction sheet. In both groups graduate students were asked to use the rubric to score 8 solutions without any explicit training or discussion. After submitting their scores they received a brief written self-training consisting of example scores and rationale for the first three solutions and were told to rescore the remaining five solutions from before and score five new solutions.

Level of Score Agreement

Reliability was assessed by comparing the graduate students' scores to the consensus scores of two expert raters. Since the reliability values are approximately the same for both the mechanics and E&M problems, the scores for all eight graduate students have been combined into a single analysis. As seen in Table 1, perfect agreement in scores for each category ranged from 20% to 45% before training with an overall average of 34%. After training agreement ranged from 38% to 50% with an average of 44%. Agreement within one score was higher, 77% before and 80% after training. Rater agreement with the expert raters was fair before training (weighted kappa 0.27±0.03) and improved to moderate agreement (weighted kappa 0.42±0.03) after a minimal written training exercise (Cohen, 1968).

Table 1

Percent Agreement of Graduate Student Scores with Expert Raters' Scores Before and after Training

	Before Training		After Training	
Category	Perfect Agreement	Agreement Within One	Perfect Agreement	Agreement Within One
Useful Description	0.38	0.75	0.38	0.80
Physics Approach	0.37	0.82	0.47	0.90
Specific Application	0.45	0.95	0.48	0.93
Math Procedures	0.20	0.63	0.39	0.76
Logical Progression	0.28	0.70	0.50	0.88
Overall	0.34	0.77	0.44	0.85

Scores in the categories Mathematical Procedures and Logical Progression were most affected by the training. These aspects initially had the lowest percent agreement with the expert raters, indicating differences in interpretations of the categories. Viewing the written examples helped to normalize scores resulting in a closer match with the expert rater scores. Useful Description and Specific Application of Physics were not significantly affected by the self-training.

Comments From Graduate Students

The graduate students also responded to questions about the rubric and suggested changes (see the Appendix for a list of questions). Their comments focused on scoring difficulties, difficulties understanding the either the category descriptions or the evidence for a category, and the adequacy of the training materials.

Some graduate students expressed confusion about the "Not Applicable" scores. These scores and the score zero were largely ignored or avoided, even after training. It should be noted that no examples of the NA(Problem) rating were included in the self-training materials. One graduate student (GS) commented, "I am confused by the need for NA(Solver). What is an example of when this would be an appropriate score?" (GS #4). Another graduate student commented that the training "would be more helpful if it covered the score range for each category…No example of NA(P) means I still don't know how/if to apply it" (GS #1).

One graduate student expressed difficulty scoring the mechanics problem, which had multiple parts (a and b) that each required a student to solve for a separate physics quantity. This person expressed difficulty deciding whether to assign separate rubric scores for each part of the problem, or to give one overall score for the solution. The graduate student commented, "Should have scored each part separately - otherwise the score takes a sort of average which does not tell much" (GS #2).

Written comments also indicated the graduate student raters were influenced by their traditional grading experiences. They expressed concerns about scoring math and logical progression when the physics is inappropriate: "I don't think credit should be given for a clear, focused, consistent solution with correct math that uses a totally wrong physics approach" (GS#1); or "When grading math procedure I wondered if it mattered they were trying to solve the wrong problem but did the math right they were trying to do" (GS #5). Some also wanted to weight the categories based on their importance to the problem, as GS #8 indicated:

> [The student] didn't do any math that was wrong, but it seems like too many points for such simple math…I would weigh the points for math depending on how difficult it was. In this problem the math was very simple.

Graduate students also perceived substantial overlap in some categories and had difficulty treating some of the categories independently: "I think description & organization are in some respect very correlated, & could perhaps be combined" (GS #5). GS# 1 remarked, "Specific application of physics was most difficult. I find this difficult to untangle from physics approach. Also, how should I score it when the approach is wrong?"

In response to the training materials, GS #6 commented, "They [example scores] helped me understand what someone else thought was important. They did seem a touch harsh. I also think I was a little lax the first time around. Examples help clarify the details." One graduate student did not perceive the training example score as very helpful, and commented "I did not always agree with them" (GS #2).

Revisions to the Rubric and Training

Based on this data, both the rubric and training were modified. The scores were changed to include the NA(Problem) and NA(Solver) categories more prominently in the rubric, and the 0-4 scale was changed to 0-5. In the previous version, the zero score designated both "all missing" or "all inappropriate", and this score was split into two scores due to the graduate students' tendency to give a score of 1 for showing some work, even if it was all inappropriate. The language was also made more parallel in every category and the order of scoring the categories in the rubric was changed with Useful Description placed before Physics Approach. The training materials were revised to include NA score examples, a clearer description of the rubric's purpose, and score examples written directly on the student solution rather than in a separate table.

Analysis of Written Solutions

An analysis of students' written solutions to physics tests from a semester of introductory physics was used to obtain evidence for response processes, generalizability, relationships of categories, and to propose instructor uses of the rubric. The test copies were collected in the first semester of calculus-based physics for science and engineering (mechanics). Four tests during the semester each included two free-response problems. For each problem approximately 300 student solutions were scored using the rubric. This sample represents a third of the total fall course enrollment. The tests represented standard physics topics including motion with constant acceleration in one and two dimensions, Newton's second and third laws, rotational motion, conservation of energy, conservation of momentum, and conservation of angular momentum.

Example Student Responses

The first problem on the third test could be solved using either the principle of Conservation of Energy or with forces using Newton's Second Law. As seen below, the problem statement cued on a particular object in the problem (the middle block M_3) which affected the response processes for some students. For brevity, only part A) of the four-part problem is written below:

> Problem 1: The system of three blocks shown is released from rest. The connecting strings are massless, the pulleys ideal and massless, and there is no friction between the 3kg block and the table. (A) At the instant M3 is moving at speed v, how far *d* has it moved from the point where it was released from rest? (answer in terms of M_1, M_2, M_3, g and v.)

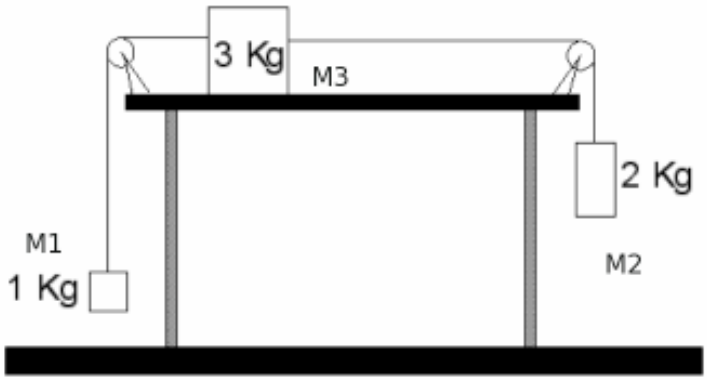

Figure 1. Diagram for Test 3 Problem 1

Approximately 17% of students answered this problem correctly, with most selecting to use the principle of Conservation of Energy. The most common application error (25% of students) was to only consider the kinetic energy of block 3, rather than the kinetic energy of all three blocks. An example of this application error is shown in Figure 2.

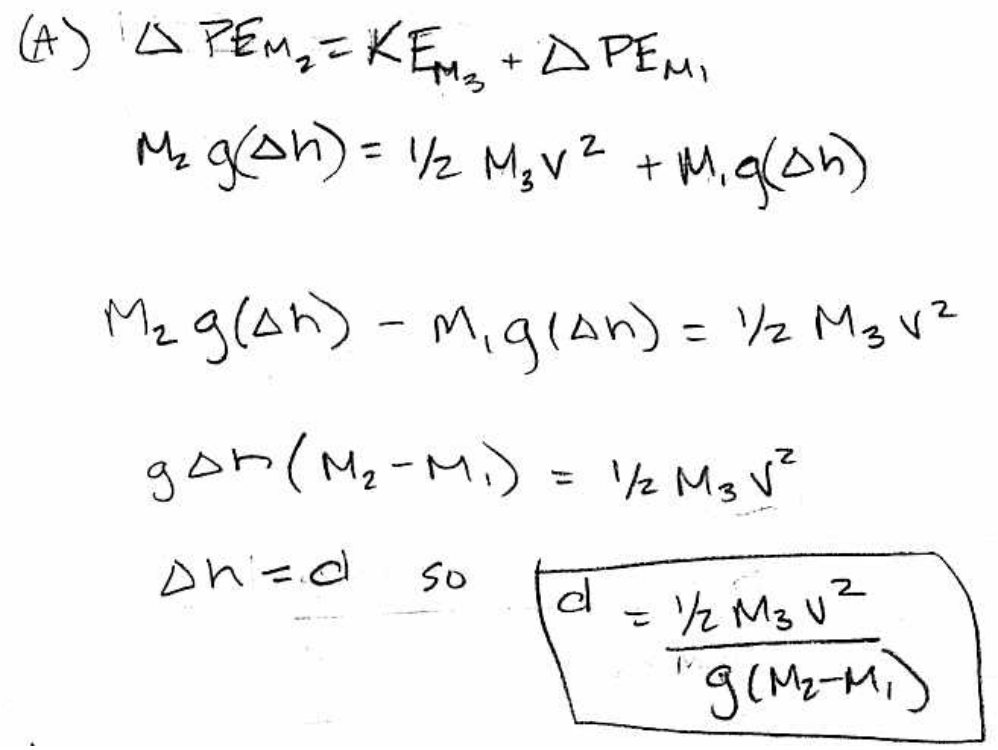

Figure 2. Example Student Solution with Conservation of Energy Application Error.

Another common error was to apply Newton's Second Law with incorrect reasoning that the tension in each string was equal to the weight of the hanging masses. At least 15% of students misapplied Newton's Second Law with this reasoning. An example is provided in Figure 3.

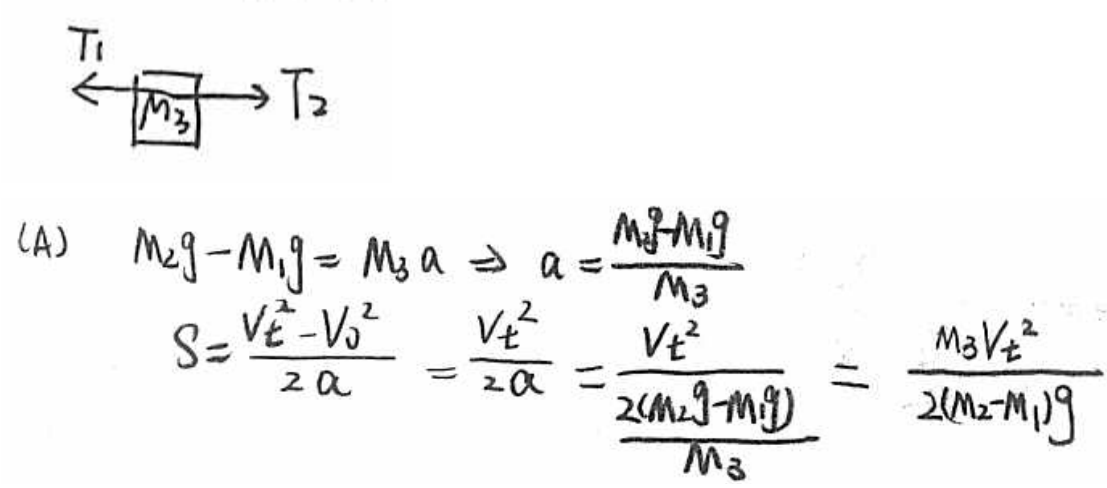

Figure 3. Example Student Solution. Newton's Second Law with T=Mg Reasoning.

For some student solutions, the final answer is correct but the reasoning is unclear. An example is shown in Figure 4. For this student, it is possible that the answer was obtained using correct reasoning (F represents net external forces) but it is also possible that the student used false reasoning, such as the T=Mg error from Figure 3.

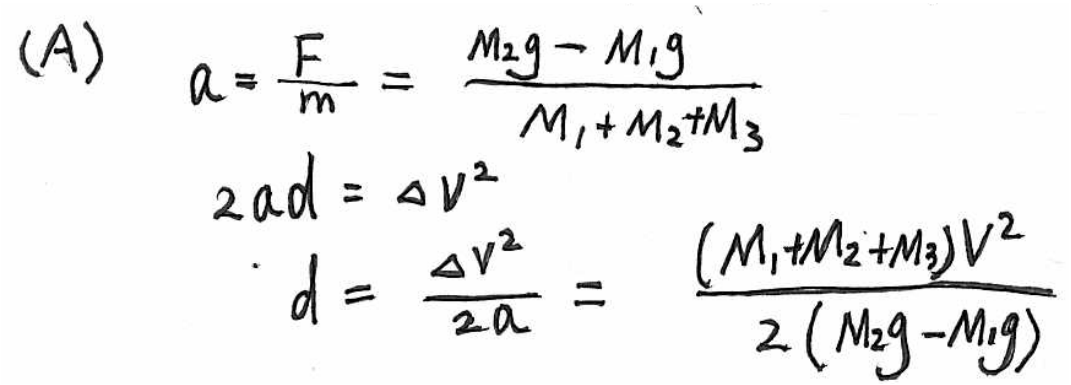

Figure 4. Example Student Solution. Correct Answer with Unclear Reasoning.

Rubric Usefulness

The rubric can be used to indicate areas of student difficulty for a given problem. For example, rubric scores on this test problem indicated 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. Logical Progression scores were generally in the middle, around a score of 3. For students who appropriately applied a Conservation of Energy approach without an explicit description, the Useful Description was scored NA(Solver).

When compared to the standard grading procedure of assigning a single numerical score to a test problem, the rubric provides significantly more information that can be used for coaching students. For example, frequent low scores in a category (such as the low scores in Specific Application) can help focus instruction on modeling this skill and providing guided practice. The rubric only indicates an area of difficulty, however, and a more detailed analysis is required to determine specific difficulties or common responses.

The rubric also provides instructors information about how the problem statement affects students' problem solving performance, which could be used to modify problems. In the test example, the problem statement cued on the middle block and student solutions reflected this focus. Additionally, visualization skills were not measured in this problem and the rubric responded with a high frequency of NA(Solver) scores in the description category.

Problem Characteristics

The rubric was applied to a range of physics topics tested throughout the semester without difficulty. However, there were some characteristics of problems that did seem to affect the generalizability and meaningfulness of the rubric scores. On the first mechanics test, the questions were much too easy for students and the class average was over 80%

correct on each problem. The rubric reflected this, producing high scores in every category. This probably indicates that these questions were not problems for the students. It may also indicate that student problem solving performance depends on the complexity of the problem. Similarly, when processes are not measured for a problem (such as when the description or physics principle is provided), the rubric produces the appropriate Not Applicable scores which shows that this instructional practice does not probe that dimension of student learning.

The analysis of written work also indicated some characteristics of problems can mask the nature of a student's problem solving processes, such as explicit prompts for procedures or physics cues. For example, a question on the second test explicitly prompted students to draw a free-body diagram in the problem statement. It is unclear whether students would have engaged in this procedure if it had not been prompted and many did not use the diagram once it was written.There is also some indication that symbolic problem statements make it more difficult for a student to construct a logical path to a solution. In summary, when interpreting rubric scores it is important to consider the complexity of the problem and possible bias in problem characteristics.

Problem-Solving Interviews

Another source of evidence for validity based response processes is student problem-solving interviews (in progress). In the interview, students are asked to solve physics problems while their actions and voice are recorded. After completing the problem, they must explain their reasoning to an interviewer. The written work is scored using the rubric and compared to the verbal protocols. This will give an indication of the processes engaged in by students during problem solving, and extent to which written solutions are indicative of problem-solving process.

Summary

The goal of this study is to design a simple problem-solving measure for written solutions to physics problems and establish evidence for validity, reliability, and utility. A measure is in the process of development based on the research literature in the form of a rubric, which assigns a separate score for five expert-like problem-solving processes (useful description, physics approach, specific application of physics, mathematical procedures, and logical progression). The initial studies of the rubric indicate that it seems to be possible to design such a measure that is easy-to-use, provides meaningful information, and produces reasonably valid and reliable scores.

The tests with graduate student raters indicated a reasonable level of score agreement, and suggested several changes to the rubric and training materials. Reliability is expected to improve with these revisions. The analysis of test solutions from a semester-long introductory physics course indicated that the rubric is applicable to different physics topics in mechanics. It also indicates that some problem characteristics mask student problem solving processes, such as overly explicit procedural prompts and physics cues. The rubric provides more meaningful information than standard grading by indicating areas of student difficulty that can be used to focus coaching and improve problem writing.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Blue, J. M. (1997). Sex differences in physics learning and evaluations in an introductory course. (Unpublished doctoral dissertation, University of Minnesota, Twin Cities.)

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213-220.

Foster, T. (2000). The development of students' problem-solving skills from instruction emphasizing qualitative problem-solving. (Unpublished doctoral dissertation, University of Minnesota, Twin Cities.)

Hayes, J.R. (1989). *The complete problem solver* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Heller, J.I., & Reif, F. (1984). Prescribing effective human problem-solving processes: Problem description in physics. *Cognition and Instruction*, 1(2), 177-216.

Heller, K. (2006). *Competent Problem Solver – Calculus Version*. Mason, OH: Thomson.

Heller, P., Keith, R., & Anderson, S. (1992). Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving. *American Journal of Physics,* 60(7), 627-636.

Henderson, C., Yerushalmi, E., Kuo, V., Heller, P., Heller, K. (2004). Grading student problem solutions: The challenge of sending a consistent message. *American Journal of Physics*, 72(2), 164-169.

Larkin, J. (1979). Processing information for effective problem solving. *Engineering Education*, 70(3), 285-288.

Larkin, J. (1981). Cognition of learning physics. *American Journal of Physics*, 49(6), 534-541.

Larkin, J.H., McDermott, J., Simon, D.P., & Simon, H.A. (1980a). Expert and novice performance in solving physics problems. *Science*, 208(4450), 1335-1342.

Larkin, J.H., McDermott, J., Simon, D.P., & Simon, H.A. (1980b). Models of competence in solving physics problems. *Cognitive Science*, 4, 317-345.

Leonard, W.J., Dufresne, R.J., & Mestre, J.P. (1996). Using qualitative problem-solving strategies to highlight the role of conceptual knowledge in solving problems. *American Journal of Physics*, 64(12), 1495-1503.

Reif, F., & Heller, J. (1982). Knowledge structure and problem solving in physics. *Educational Psychologist*, 17(2), 102-127.

Martinez, M. E. (1998). What is problem solving? *Phi Delta Kappan*, 79, 605-609.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*(9), 741-749.

Newell, A., & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Pólya, G. (1945). *How to solve it*. Princeton, NJ: Princeton University Press.

Reif, F., Larkin, J.H., & Brackett, G. (1976). Teaching general learning and problem-solving skills. *American Journal of Physics*, 44(3), 212-217.

Schoenfeld, A.H. (1985). *Mathematical problem solving*. Orlando, FL: Academic Press, Inc.

Van Heuvelen, A. (1991). Overview, case study physics. *American Journal of Physics*, 59(10), 898-907.

Appendix – Questions from Rubric Training Study

(After the first scoring of eight solutions)

1. What difficulties did you encounter while using the scoring rubric?
    a. Which of the five categories was most difficult to score and why?
    b. Which student solutions were the most difficult to score and why?
2. What changes, if any, would you recommend making to the rubric? Why?
3. If you were deciding how to grade these student solutions for an introductory physics course exam, how would you assign points? (out of 20 total points)

(After training and second scoring of ten solutions)

4. What difficulties did you encounter while using the scoring rubric?
5. Were the example scores useful? Why or why not?
6. What further changes, if any, would you recommend making to the rubric?

	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.
MATHE-MATICAL 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.