Authentic assessment of students' problem solving

Qing Xu[1], Ken Heller[1], Leon Hsu[2], Bijaya Aryal[3]

[1]*School of Physics and Astronomy, University of Minnesota, Minneapolis, MN, 55455*
[2]*Department of Postsecondary Teaching and Learning, University of Minnesota, Minneapolis, MN, 55455*
[3]*Center for Learning Innovation, University of Minnesota, Rochester, MN, 55904*

Abstract. Improving curricular materials and practices aimed at complex cognitive processes such as problem solving requires careful planning and useful tools for assessment. To illustrate the challenges of measuring a change in students' problem solving in physics, we present the results of and a reflection on a pilot assessment of the effectiveness of computer problem-solving coaches[1] in a large (200+ student) section of an introductory physics course.

Keywords: problem solving, rubric, computer coaches, assessment
PACS: 01.40.Fk, 01.40,gf, 01.50.H-, 01.50.Kw

INTRODUCTION

In any area of research, the most important measurements are often the most difficult. One example is investigating a complex cognitive activity such as problem solving in a classroom situation. There are four big challenges in making this measurement: (1) specifying the student behavior that signals progress; (2) designing an instrument with the precision and discrimination to measure that progress; (3) constructing an experiment that either measures, controls, or averages over the confounding factors influencing student performance; and (4) using a classroom situation in which improvement is neither blocked nor masked.

Effective problem solving involves a constellation of cognitive processes that have been assessed in laboratory situations using, for example, think aloud interview techniques or classification tasks [2,3]. However, this complexity makes direct quantitative assessment difficult in an authentic situation. In a classroom setting any signal may well be buried in noise arising from the very nature of the educational process. In this paper, we discuss the four challenges listed above in the context of a pilot study conducted to assess the effectiveness of computerized problem-solving coaches in a large section of an introductory physics course.

Almost since the introduction of computers, there have been numerous attempts to use them to help students improve their problem solving skills. The coaches used in the experiment described below are web-based programs developed at the University of Minnesota designed to provide students with individualized guidance and feedback while giving them practice in the decision-making processes critical to competent problem solving. Each of the coaches, which are described in more detail elsewhere [1], helps students solve a single physics problem.

STUDY

In Fall 2011 we pilot-tested 35 prototype computer coaches delivered over the internet in one section (219 students) of an introductory calculus-based mechanics class at the University of Minnesota to assess their usability and explore their educational impact. To set up a sample of coaching users and non-users, students were allowed to satisfy their homework requirement either by completing the computer coaches for a given topic, by submitting a correct answer to the same problems through WebAssign (www.webassign.net) within three attempts, or by a combination of the two methods. Student use of the coaches was monitored by recording their keystrokes. During the course, students took four written in-class tests, each with two free-response context-rich problems to solve and a final exam with five more standard problems. We collected the students' written solutions to these 13 problems.

We also collected written problem solutions from another section of the same course taught during the same semester by a different professor. This class did not use computer coaches but did use Learning Assistants [4] to facilitate small group discussions during lectures. Both sections used the Cooperative Group Problem Solving pedagogy and had professors who emphasized the use of organized problem solving frameworks [5].

Assessment: Instruments And Strategy

In addition to the written problem solutions and keystroke data, we also collected pre- and post-test scores on the Force Concept Inventory (FCI) [6], a

math diagnostic test, and the Colorado Learning Attitudes about Science Survey (CLASS) [7], as well as scores on the problems as determined by the Teaching Assistants (TAs) as part of students' grades.

Students' problem-solving was measured by applying a rubric to their written problem solutions. The rubric was developed and tested for reliability, validity, and utility at the University of Minnesota [8,9] and includes five categories: (1) representing problem information in a Useful Description (UD), (2) selecting appropriate physics principles (Physics Approach, or PA), (3) applying physics principles to the specific conditions in the problem (Specific Application of Physics or SAP), (4) using appropriate Mathematical Procedures (MP), and (5) the overall communication of an organized reasoning pattern (Logical Progression or LP). Each category is scored on scale of 0-5 (with 5 being the most expert-like), or N/A in cases where the category is not applicable.

Two assessors, a PER graduate student and a faculty member with a PhD in PER each scored half of the students' solutions using the rubric. To ensure inter-rater reliability, they first scored the same 10 student solutions, comparing and discussing their ratings, then repeating the process until their agreement was at least 90% before discussion. This training process was repeated for each problem.

As a test of the validity of the rubric, the simple sum of rubric scores from all five categories was compared to the grades assigned to the problems by the TAs for the course. It is important to note that the grading of the problems by the TAs was completely independent of the rubric scoring and that the criteria for each were not necessarily related. However, as one would hope, the correlation between the TA score and the summed rubric score was very high, ranging from 0.82 to 0.85 for each problem.

In contrast, the correlation between student' FCI scores and the summed rubric scores is much weaker, approximately 0.25, indicating that the FCI and the rubric measure substantially different things.

Two strategies were used to measure the effect of the computer coaches on student problem solving performance. The first was to be a comparison of two groups of students within a single section, those who used the coaches frequently and those who did not. This type of comparison controls for class environment by selecting students within a single class. However, it is sensitive to contamination between the two groups, especially in cases such as the one studied, in which much of the class is based on students working together to solve problems. Students from the two groups share their knowledge, diminishing any difference between them.

The second strategy involved comparing two groups of students from different lecture sections, one using the coaches and the other not. In this case, the contamination problem is much smaller, but the classes have different professors, teaching assistants, environments, and possibly even different types of students due to scheduling constraints.

Within And Between Class Comparisons

Dividing the single section into students who used the coaches and those that did not proved to be difficult. Assigning a subset of students to use the coaches and prohibiting others from using them would generate animosity and is likely unethical. In a previous study, offering students up to $225 to use the coaches did not attract enough students to perform a study. In Fall 2011, where students could use the computer coaches to complete their homework, too many of the students chose to use the coaches to form a true "non-users" group. The average number of coaches completed by a student was 22 and the average number attempted was 28. Only 9% of the students completed fewer than 10 of the coaches. Nevertheless, we created two groups: the most frequent completers (FC) (47 students completing 30-35 coaches) and the least frequent completers (LC) (47 students completing 5-16 coaches). Some students reported that they used the coaches only until they could solve the problem on their own, then quit before completing the coach; these incomplete attempts were not counted for purposes of assigning students to the FC and LC groups.

TABLE I. Differences in background variables between frequent (FC) and less-frequent (LC) completer groups. FCI, Math, and CLASS are pretest scores.

Pre-test	FC			LC		
	Overall	M	F	Overall	M	F
N	47	24	23	47	41	6
FCI	47%	58%	36%	67%	70%	41%
Math	61%	65%	57%	68%	69%	60%
CLASS	61%	62%	59%	67%	68%	61%

As can be seen in Table I, the background measures of students in the FC and LC groups are different, particularly with respect to the FCI. FC students were 49% female while the LC students were only 13% female. The LC students also had higher scores on all pre-tests. We conclude that the students who self-selected into the FC group tended to be female and have less physics preparation.

To try to control for differences in students, we created pretest-matched subgroups from the FC and LC groups. Although we obtained groups with nearly identical pretest scores as well as a closer match with regard to gender (5f, 19m in the FC group completing an average of 33 coaches; 3f, 21m in the LC group

completing an average of 11 coaches), the statistical power of the measurement has been greatly reduced because of the smaller number of students.

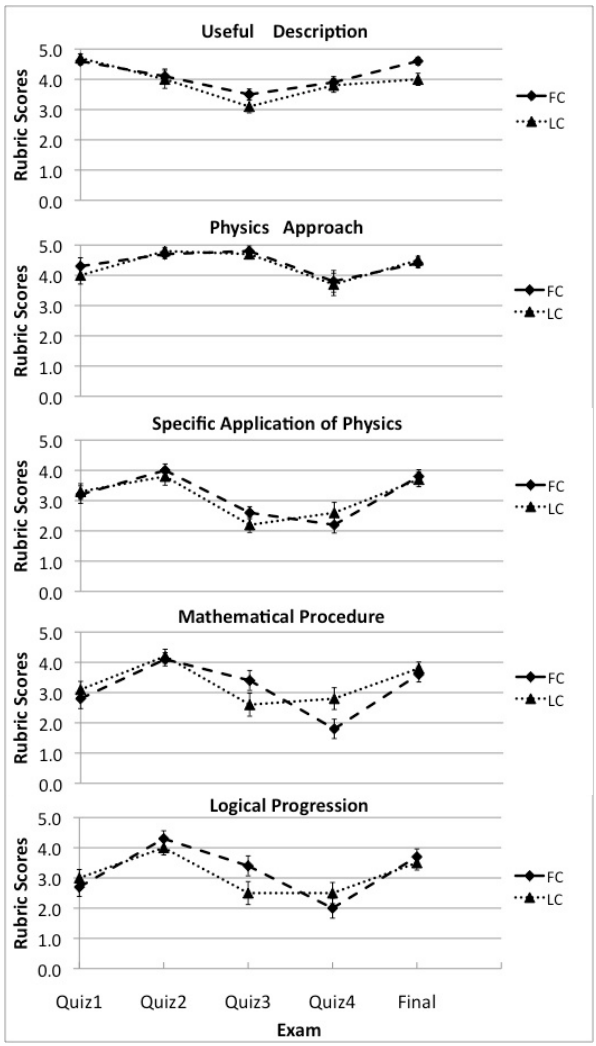

FIGURE 1. Average scores of the matched FC and LC groups on each of the five rubric categories for 4 quiz problems and 2 final exam problems (averaged together). Lines are included only to guide the eye.

Figure 1 shows the average rubric scores for the matched FC and LC subgroups for all of the problems analyzed so far, including 1 problem on each of the 4 in-class quizzes and 2 problems on the final exam.

As shown in the graphs, students' problem solving, to the extent measured by the rubric scores, do not show a monotonic trend over time (nor would one necessarily expect them to [10]) and there is no statistically significant difference between the FC and LC groups. On the first quiz, which occurred near the beginning of the course, the LC group had a higher rubric score in 4 out of 5 of the rubric categories. After the first quiz, the FC group scored higher on a majority of the problems in each category except for Mathematical Procedure, which was not a skill addressed by the coaches. This pattern, while suggestive, is not statistically significant.

To reduce the learning contamination among students, 99 students from each of two different lecture sections of the same course taught during the same semester were selected to form two matched groups based on the FCI, Math and CLASS pre test scores. In the group from the class using the computer coaches, there were about an equal number of FC and LC students (28 FC, 30 LC), as well as 41 students from neither of those two groups in the sample. The gender ratio was nearly equal in two matched groups (27 females in the group from the class using coaches and 28 females in the group from the other class). Because the two classes used different quiz problems but the same final exam, comparisons were made only using the final exam. Based on the two problems analyzed so far, there are no significant differences between the two classes.

REFLECTION

Any useful measurement needs to meet the four challenges set forth at the beginning of this article. Meeting the first challenge, to specify the behaviors that constitute a measurement signal, currently relies on the literature that distinguishes expert from novice problem solving. Based on research in developing expertise from areas as diverse as reading and athletics, this is not expected to be a monotonic progression [10]. It is not even clear that the learning progression is continuous. For example, there might be qualitatively different stages of this development such as that which has been called competent problem solving [11]. In such hypothetical intermediate stages, students might exhibit behavior that, while different from novices, is not necessarily more closely aligned with experts. This challenge can be met by making comparisons over a long enough timescale to be insensitive to such stages. Reviewing previous work, that scale is likely to be longer than a semester [10,12] reducing the sensitivity of our study.

The second challenge is to develop a measuring instrument. Measuring the state of a complex cognitive process in an authentic environment typically uses both qualitative and quantitative techniques combined in a rubric. This instrument needs to be sensitive to the development of a general problem solving process and not to specialized training for specific behaviors such as drawing a certain kind of diagram, doing mathematics in a certain way, or getting the right answer to a specific problem type. The rubric we have developed is intended to be pedagogy independent and relevant to multiple problem types and topics [8,9]. In

small scale studies, the rubric was shown to be valid by scoring various types of expert and student solutions. Furthermore, assessments of students' problem-solving skills from interviews while solving a problem correlated very strongly with the rubric assessment of their written solutions. However, the rubric was developed to distinguish between expert and novice problem solving and its scoring procedures may not be appropriate to track student problem solving progress over the timescale of this study. The rubric may have enough sensitivity if used over a timescale long enough for student behavior to become closer to that of experts, however.

The third challenge is constructing an appropriate experimental study. Such a study might show progress within a single population of students receiving the treatment or it could be a comparison between two populations of students, a treatment group and a non-treatment group. The former must directly address the issues of the fourth challenge. As described previously, samples within a class suffer from limited statistical power and are subject to contamination. Using more than one class addresses both of these issues. However having two classes, even when they are different sections of the same course taught during the same timeframe, brings in confounding parameters such as the effect of the professor and other instructors, the distribution of the student population, or the emphasis and structure of the class. In the case of our study, the two sections also had different midterm tests that made it difficult to compare students' performances as a function of time. The noise introduced by these confounding parameters might be statistically controlled by comparing more than two classes with a correspondingly larger expenditure of analysis effort.

The fourth challenge is to use classes that neither mask nor block the desired effect. For example, in order to measure a change in students' problem solving skills, one must ask appropriate exam questions (traditional problems often don't evoke the problem solving process) and reward students for using a strong problem solving process in their solution (grading cannot simply be based on the final answer nor on the appearance of specific artifacts in the solution such as a particular type of diagram). We believe these two challenges have been met in our experiment. However, when trying to measure the effect of a specific treatment (in our case, the use of computer coaches), any signal might be hidden by the simultaneous use of other problem solving pedagogies. For example, such masking was demonstrated in the case of conceptual learning where the combined use of individually effective pedagogies did not show a cumulative gain. [13]

The study we have outlined above illustrates the statistical and procedural difficulty in achieving appropriate measurement discrimination in an environment with a high noise to signal ratio. Improvements in the assessment techniques are being explored, including increasing the timescale over which the measurements are made, refining the application of the rubric to increase its sensitivity, using different student samples with initial problem solving processes that have a wider range of rubric scores, and making measurements in classes that do not use a cooperative group problem solving pedagogy. It is also possible that this type of measurement may require a more discriminating instrument than the rubric employed. In our particular case, it is also possible that the signal is very small or non-existent.

This work was supported by the National Science Foundation under DUE-0715615.

REFERENCES

1. L. Hsu and K.Heller, "Computer problem solving coaches" in *2004 Physics Education Research Conference,* edited by J. Marx et al., AIP Conference Proceedings 790, American Institute of Physics, Melville, NY, 2004, pp. 197-200.
2. J. H. Larkin and F. Reif, *Eur. J. Sci. Educ.* **1**(2), 191-203 (1979).
3. M. T. H. Chi, P. J. Feltovich and R. Glaser, *Cognitive Science* **5,** 121-152 (1981).
4. V. Otero, N. Finkelstein, S. Pollock and R. McCray, *Science* **313**, 445–446 (2006).
5. K. Heller and P. Heller. *The Competent Problem Solver, calculus version, second edition*. Minneapolis, MN: McGraw-Hill Primis Custom Publishing, 1997.
6. D. Hestenes, M. Wells and G. Swackhamer, *Physics Teacher* **30**, 141-158 (1992).
7. W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein and C. E. Wieman, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010101 (2006).
8. J. Docktor, "Development and validation of a physics problem-solving assessment rubric", Ph.D. Thesis, University of Minnesota, Twin Cities, 2009. Available: http://groups.physics.umn.edu/physed/People/Docktor_dissertation_submitted%20final.pdf
9. J. Docktor and K. Heller, "Assessment of Student Problem Solving Processes" in *2009 Physics Education Research Conference*, edited by M. Sabella et al., AIP Conference Proceedings 1179, American Institute of Physics, Melville, NY, 2009, pp. 133-136.
10. R. S. Siegler, *J. Cogn. and Develop.* **5**, 1-10 (2004).
11. H. L. Dreyfus and S. E. Dreyfus, *Org. Studies* **26**(5), 779-792 (2005).
12. P. Heller, R. Keith, and S. Anderson, *Am. J. Phys.* **60**, 627-636 (1992).
13. K. Cummings, J. Marx, R. Thornton and D. Kuhl, *Am. J. Phys.* **67**, S38-S44 (1999).