Peer Evaluation of Video Lab Reports in a Blended Introductory Physics Course

Scott S. Douglas*, Shih-Yin Lin*, John M. Aiken*, Brian D. Thoms†, Edwin F. Greco*, Marcos D. Caballero** and Michael F. Schatz*

*Department of Physics, Georgia Institute of Technology, Atlanta, GA 30332
†Department of Physics and Astronomy, Georgia State University, Atlanta, GA 30303
**Department of Physics and Astronomy and CREATE for STEM Institute, Michigan State University, East Lansing, MI 48824

Abstract. The Georgia Tech blended introductory calculus-based mechanics course emphasizes scientific communication as one of its learning goals, and to that end, we gave our students a series of four peer-evaluation assignments intended to develop their abilities to present and evaluate scientific arguments. Within these assignments, we also assessed students’ evaluation abilities by comparing their evaluations to a set of expert evaluations. We summarize our development efforts and describe the changes we observed in student evaluation behavior.

Keywords: mechanics, peer evaluation, scientific communication

PACS: 01.40.E-, 01.40.Fk, 01.40.G-

INTRODUCTION

In Spring 2014, Georgia Tech took two sections of its large-enrollment introductory mechanics course using the Matter & Interactions curriculum [1] and ran a “blended” version of that same course (N=355 students). Our blended course featured out-of-classroom laboratory exercises with online video lab reports and peer evaluation. Traditional lectures were largely replaced with online lecture videos [2], but unlike a fully “flipped” course, our blended course still devoted some time to formal instructor lecturing.

We designed the four labs in our course with an eye toward computational modeling and real-world physics practice. We intended these laboratory exercises and the peer evaluation process to address two important learning goals: to help our students develop an understanding of physics as something applicable to their everyday experience, and to develop their practice of scientific communication through presenting the results of their own laboratory exercises and evaluating the quality of their peers’ presentations.

This paper describes our progress toward the latter part of this second learning goal. How does our students’ peer evaluation behavior change over the course of the semester? To address this question, we compared student and instructor ratings of 20 common video lab reports across four labs. Our proximate goal (and the goal of this paper) is to characterize this gain numerically, and to deal with potential sources of systemic bias.

OUR LABS

In a typical lab, students were instructed to acquire observational data by recording video of real moving objects with their own smartphone or laptop cameras. Students then imported these observation videos into the motion-tracking software Tracker [3], which they used to extract position vs. time data. Finally, students used these data (along with other physics concepts) to create computer models of the same motion in VPython [4]. Students had two weeks to perform these laboratory activities, prepare a five-minute video lab report, and upload that lab report to YouTube and submit the link to the course instructors.

In the week following the submission of the lab reports, we conducted the peer evaluation process with the rubric summarized in Table 1 (students were shown this rubric in-class before beginning their first lab assignment). The rubric asks students to evaluate each video lab report in terms of its structure, its physics content, and its production quality. Each of five items on the rubric comprised one rating on a five-point poor-to-excellent scale and one textual comment. Students received significant support and instruction before doing their first evaluations; our corpus of lecture videos contained videos about preparing and evaluating video lab reports, including a step-by-step example evaluation of two actual video lab reports and several videos specifically relevant to each laboratory exercise [2]. In the classroom, we supplemented these instructional videos with in-class small-group practice presentations intended to provide students with helpful feedback on their lab reports in progress.
TABLE 1. Rubric Summary for Spring 2014

| Item | Topic | Features to Evaluate |
|------|-------|----------------------|
| 1    | Organization/Structure | Introduction, conclusion, transitions, overall flow of presentation |
| 2    | Content: Models | Identification of relevant models, discussion of main physics concepts, connection between models & fundamental principles |
| 3    | Content: Prediction Discussion | Comparison between computer-generated data and observational data, explanation of discrepancies between same |
| 4    | Content: Overall | Presence or absence of major physics errors, discussion of specific instructor-posed questions |
| 5    | Production/Delivery | Technical quality of video, vocal quality of narration |

As part of the peer evaluation process, course instructors randomly assigned three peer videos to each student, along with five common instructor-evaluated videos and each student’s own video (for self-evaluation) for a total of nine video evaluations per student per lab. Evaluations were conducted in three phases:

1. **Practice**: Each student evaluated two of the common instructor-evaluated videos for practice (no credit). After evaluating each of the two practice videos (hereafter P1 & P2), each student was shown the detailed instructor evaluation for that video.

2. **Calibration**: Each student evaluated another two of the common instructor-evaluated videos (C1 & C2) for credit, and received a “calibration grade” dependent on how well her evaluation aligned with the expert evaluation for that video. After evaluating each calibration video, each student was shown the detailed instructor evaluation for that video.

3. **Evaluation**: Each student evaluated three peer videos, her own video, and a final instructor-evaluated video in random order. This last instructor-evaluated video (the “hidden” calibration video, HC) was presented to the student as just another peer video, but her evaluation of this video also counted toward her calibration grade. Students were never shown the instructor evaluation for the hidden calibration video.

**RESULTS**

Figure 1 shows the distribution of student ratings for every rubric item for every instructor-rated video in the semester progresses.

**FIGURE 1.** Student ratings for all instructor-rated videos in Spring 2014. Each column represents one rubric item as described in Table 1, each row represents one video, and each cell shows the normalized distribution of student ratings for that item on that video. Green (light) cells indicate instances where the modal student gave the same rating as the instructors; blue (dark) and orange (medium) cells indicate where the modal student rated below or above the instructors, respectively. Instructor ratings are indicated with dark triangles. Reading the figure from top to bottom shows more instructor-student agreement as the semester progresses.
semester. Reading the figure from the top row to the bottom row follows the students’ chronological progress through all the instructor-rated videos in the course; read this way, Fig. 1 shows an apparent overall gain in agreement between student and expert ratings on instructor-rated videos over the semester. The increasing proportion of green cells toward the bottom of Fig. 1 represents an increasing number of instances where a plurality of students agreed with the instructor rating for a rubric item on a video. However, the instructor ratings themselves were not equally distributed among all the labs; the videos in Lab 4, for example, had more “good” instructor ratings (N=9) than did the videos in Lab 1 (N=2). If students were biased toward responding with e.g. “good”, then the apparent gain in agreement may have been an artifact of a systemic bias caused by the chance presence of more “good” instructor ratings toward the end of the course. Likewise, if our students were biased against “very good”, then the lesser number of “very good” instructor ratings in Lab 4 (N=3) relative to Lab 1 (N=6) would also produce an artificial gain in agreement.

Figure 2 addresses this concern by looking at student ratings between labs but within expert ratings, excluding the single “excellent” among all 100 instructor ratings. We excluded the instructor rating “excellent” because of concerns over statistical power and because a single instructor rating in one lab does not allow us to do any cross-lab comparisons within that rating. The differences between all student rating distributions in Lab 1 and subsequent labs were determined to be statistically significant with a two-sample Kolmogorov-Smirnov test, p < 0.05 [5]. Read from left to right, this figure shows at least a small gain in student/instructor agreement within every instructor rating and a large gain within “good”, precluding the possibility that the overall gain in agreement is due solely to a systemic bias caused by a nonuniform distribution of instructor ratings. The mean student-instructor agreement within “poor” (mean 22%, range 15%-34%) and “fair” (mean 23%, range 14%-34%) is substantially less than the mean student-instructor agreement within “very good” (mean 38%, range 32%-44%), and is indeed barely above chance (20% agreement for uniform random guessing). As it turns out, the presence of more “very good” responses in Lab 1 than in Lab 4 actually serves to make the apparent gain artificially low. Student-instructor agreement within “good” (mean 30%, range 23%-44%) improves from Lab 1 to Lab 4 twice as much as within any other rating, has in Lab 4 the same proportion of agreement as does “very good” (44%), and is the only instructor rating which shows a >10% gain in student-instructor agreement.

As illustrated by both Figs. 1 and 2, the “excellents” among the student ratings almost disappear after Lab 2, even though the instructor ratings for the latter two labs are not lower overall. There was only one “excellent” among all 100 instructor ratings in the course. This reflects an instructor norm that holds “excellent” to be a much loftier rating than our students might have initially thought, given that “excellent” constituted 24% of student ratings in Lab 1. If our students had learned to avoid giving “excellent” ratings, then this trend alone may account for some of the overall gain in student-instructor agreement; a hypothetical random guesser would improve his agreement from 20% to 25% simply by eliminating “excellent” from his guesses, achieving a gain of 5%.

We have so far described three major trends. Most of the gain in agreement occurred in the middle of the rating scale (i.e., when the instructors and students said “good”), students almost stopped giving “excellents” toward the end of the semester, and agreement at the high end of the scale (“very good”) began high and stayed high while agreement at the low end (“poor” and “fair”) began low and stayed low. All these trends together suggest that the overall gain in agreement may be partly ex-
plained by a change in student understanding of the rating scale itself (e.g., how much better than “very good” is “excellent”)). Any such change would constitute a change in students’ rating norms, which cannot be fully explored on the basis of ratings alone. To get a clear picture of our students’ rating norms, we will need to examine students’ comments and conduct student interviews.

Finally, while the overall trend shown in Fig. 1 is toward increased agreement, there are still some remarkable instances of disagreement between student and instructor ratings throughout the semester. These disagreements may be helpful in investigating the reasons behind students’ ratings, and illustrate a serious limitation of analyzing the ratings alone. Lab 3 Practice 1 Item 1 (L3P1#1) has an instructor rating of “poor” but shows low student-instructor agreement (few students say “poor”) and low student-student agreement (roughly equal numbers of students say “fair”, “good”, and “very good”). L3P2#1, the same item on the very next video, also has an instructor rating of “poor” but shows a very different distribution of student responses; student-instructor and student-student agreement is very high, since almost all students also rated this item “poor”. L2P2#3 and L2C2#3 show the opposite phenomenon on a different rubric item (one relating to physics content). Here, the distributions of student ratings are roughly similar (both have a single peak at “very good”), but the instructor ratings are very different (“fair” and “excellent”, respectively).

Since the same rubric item on two different videos can yield the same instructor rating but different student ratings (and vice-versa), this suggests a difference in the actual content of those videos and a difference between the video features to which instructors and students attend when evaluating that rubric item (i.e., a difference in video-watching practices). This is corroborated by instructor comments; instructors rated L3P1#1 “poor” because the introduction was “simply [a] reading [of] the problem statement”, but gave L3P2#1 the same “poor” rating for a different reason (L3P2 contained “no intro at all”). When we look within instructor ratings, we are therefore not necessarily looking within specific video features, and the overall agreement gain may still be vulnerable to systemic bias through an uneven distribution of different video features throughout the semester. Once again, we cannot fully explore these video-watching practices or detect any changes in them by analyzing the ratings alone. We will need to examine other sources of data in our future work to understand students’ video-watching practices and the interplay between video content and student rating. Student comments constitute one such data source; these comments span the range from terse declaratory statements to detailed explanations of the reasons behind the ratings. For example, a majority of students wrote some variation of “no introduction” on L3P2#1, while one student commented on L3P2#2 “Pretty good here! I put good instead of very good because my physics professor told me to be mean when grading.” These comments should provide a rich source of information for our ongoing work.

CONCLUDING REMARKS

The gain in student/instructor agreement in these video evaluations is real in that it is not solely a result of a systemic bias introduced by the chance distribution of instructor ratings among videos throughout the four labs. Our students got at least slightly better at agreeing with instructor ratings across the board, but got substantially higher agreement only among the items which instructors rated “good”. The instructor ratings of “poor” and “fair” exhibited lower overall instructor-student agreement than did “very good”. We do not yet have a clear picture of what norms and practices related to peer evaluation our students are adopting to produce these gains, because the identification of the specific norms and practices that inform peer evaluation lie beyond any analysis of the ratings alone.

In our future work, we will attempt to characterize this gain in rating agreement in terms of the norms and practices adopted by our students throughout the semester. We have already begun an investigation of the comments left by students along with each of their ratings, which we expect will shed some light on how and why the reasoning, practices, and norms behind the student ratings evolve during the course of instruction.

This work was supported by the Gates Foundation and the Georgia Governor’s Office of Student Achievement. We gratefully acknowledge the work of Christopher Wang and other members of the Georgia Tech Physics MOOC VIP team who helped develop the peer evaluation software we used in our course. We further acknowledge the work of Dr. David Lawrence and the Georgia Tech Center for the Enhancement of Teaching and Learning in helping to develop our rubric.

REFERENCES

1. R. Chabay, and B. Sherwood, Matter and Interactions, John Wiley and Sons, Inc., 2010.
2. Your world is your lab playlist (June 2014), https://www.youtube.com/playlist?list=PL9HgJKLOnKxedh-yIp7FDzUTwZeTeoR-Y.
3. D. Brown, Tracker (2014), https://www.cabrillo.edu/dbrown/tracker/.
4. D. Scherer, et al., Vpython (2014), https://www.vpython.org.
5. I. T. Young, Journal of Histochemistry & Cytochemistry 25, 935–941 (1977).