USING CALIBRATED PEER REVIEW™ IN INTRODUCTARY STATISTICS COURSES

Melissa Q. Pittard and William S. Rayens
University of Kentucky, USA
Melissa.pittard@uky.edu, Rayens@uky.edu

The two most common challenges when grading written assignments in large undergraduate statistics courses are finding the time, and getting students meaningfully involved in feedback provided by the effort. Simple, paper-exchange peer grading is one option, but it can be difficult to implement, and more difficult to ensure that students are taking the process seriously. Calibrated Peer Review™ (CPR), is a "web-based, instructional tool" that was designed to electronically facilitate peer grading, while also attempting to address the integrity of the students' effort in the grading process and provide students with an extended learning experience. This article discusses students' and instructors' perceptions of and experiences with CPR in large undergraduate statistics classes at the University of Kentucky.

BACKGROUND

The University of Kentucky department of Statistics offers a general education conceptual-based statistics course, STA 210: Introduction to Statistical Reasoning, which is required by most undergraduates at the university. The sizes of the classes range from about 75 to 135 students. The goal of the course is "to help students develop or refine their statistical literacy skills" (course description). Due to the conceptual nature of the course, several essay-type assignments are assigned throughout the semester. These assignments are difficult to grade in an efficient manner due to the number of students and the number of assignments.

Besides, over many years the authors have noticed that traditional feedback on assignments is often not looked at by the student at all, or at least not in a timely way. Ideally, in a smaller class, time would be allocated to go over a writing assignment or meet individually with students to discuss their performances. But that is just not the environment that most of us work in, nor is it the environment facing higher education today. As the value of statistical thinking and statistical reasoning has become more widely recognized, enrollments in statistics courses at the college level have begun to grow (Scheaffer & Stasney, 2004), and an ever-increasing number of students are taking courses in statistics to satisfy the common quantitative literacy requirement for graduation at their respective undergraduate institutions. The latest national estimate is that 260,000 undergraduate students in the United States were enrolled in a statistics course, an increase of over 40,000 students in a ten-year period (Lutzer, Rodi, Kirkman, & Maxwell, 2007). This number is likely an underestimate since it is based on enrollments in courses offered by mathematics and statistics departments and does not count the many students who take statistics courses in other departments (Dupuis, Medhanie, Harwell, Lebeau, Monson, & Post, 2012).

Simple paper-exchange peer review of such assignments is an option to address the large amount of grading, but quality control of the peer grading is difficult. Students do not always take the process seriously, or may not understand the assignment or the grading instructions well enough to grade the assignment properly. Calibrated Peer Review™ is "web based writing and peer review program" (http://cpr.molsci.ucla.edu/Home.aspx) designed to implement multiple writing assignments into courses. Several of the statistics instructors at the University of Kentucky have been using CPR for these writing assignments over the past three semesters. The general process involves the following steps (http://cpr.molsci.ucla.edu/Home.aspx):

1. Students first write and submit an essay on a topic, in a format specified by the instructor.
2. Students then assess three 'calibration' submissions against a detailed set of questions that address the criteria on which the assignment is based. Students individually evaluate each of these calibration submissions according to the questions specified by the rubric and then assign a holistic rating out of 10.
3. The system automatically gives each student anonymous submissions by three other students. They use the same rubric to evaluate their peers' work as they did for the calibrations.
4. Finally, students are required to evaluate their own submission, an evaluation that will be compared electronically with peer evaluation of that work.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2014 ISI/IASE]

The designing of the calibrations and rubrics can be extremely flexible as to type and number of questions as well as number of responses. The authors have chosen for all rubrics used for their courses to be yes/no questions, usually 10 of these, when the material is rich enough to warrant that many questions. Each portion of the peer review is weighted according to the instructor's design. For example, the text submission of the CPR assignment may have a weight of 30% (that is graded by three peers), the assignments that the students are "calibrated" on may have a weight of 20%, the peer reviews that the student performs for three other students may have a weight of 30% and the self-assessment of that student's work may have a weight of 20%.

The instructor also has flexibility with respect to the "tolerance limits" that a student must score the calibrations, the peer reviews, and the self-assessment. One "tolerance plan" that the authors used was:

> In the calibrations, students must get 7 of the 10 questions correct and cannot deviate from the correct number of question by more than 3 points. In peer reviews, students cannot deviate by more than 3 points from the average of the other peer reviews of that student. Finally, in the self-assessment, students will only get partial credit if they deviate by 4 points from the average peer review grade for their assignment and no credit for the self-assessment if they deviate by 5 or more points from the average peer review grade.

CPR has been implemented successfully in the sciences for several years (see http://cpr.molsci.ucla.edu/Home.aspx). It was developed by the Division of Molecular Sciences across six California institutions in 1998 and to date there are nearly 1600 institutions that utilize CPR. Several articles have been written in the sciences about utilizing CPR, but, there are no articles pertaining to its use in statistical science. Assignments are built and placed in the electronic CPR library where they can be called when needed. These assignments are then available for anyone at an institution with a CPR license. Licensing in not free, and is based on the size and type of the institution, the number of departments utilizing CPR, and the number of students (see http://cpr.molsci.ucla.edu/Home.aspx).

DATA-BASED STUDIES

Student Perception

Students from the fall 2012 and the spring 2013 semesters were surveyed about their perceptions regarding CPR. The survey questions shown in the table below were used, taken from (Walvoord et al, 2008). The two semesters were analyzed separately, since CPR was being piloted by only the authors in the fall of 2012 and utilized by three additional instructors in the spring of 2013. There is some voluntary response bias, and a small amount of extra credit was offered to the 2013 group as incentive to respond. Students who responded seemed to agree that the process was simple and the grades were easy to understand and they even agreed with their grade.

Question	Fall 2012(n=100)		Spring 2013(n=524)	
	Agree or strongly agree	Disagree or strongly disagree	Agree or strongly agree	Disagree or strongly disagree
The process of completing an assignment in CPR was simple.	74%	20%	69.7%	18.9%
The processes of calibration and peer review were simple.	65.7%	28.3%	61.9%	24.8%
The assignment results (CPR explanation of how your grade was determined) were easy to understand.	63.3%	26.5%	65.8%	21.0%
I agree with the grade I received on my CPR assignments.	62%	26%	56.9%	25.4%
The assignments used in CPR helped me to better understand the related course material.	48%	40%	47.2%	31.3%
The calibration and review process helped me better understand the assignments and the related course material.	45%	43%	46.7%	31.4%
CPR helped me improve my critical reading.	48%	40.8%	48.9%	27.5%
Considering this method as a whole, I prefer CPR over turning in a "regular" paper to my instructor.	53.5%	33.3%	34.8%	47.0%

They seemed to split on whether the assignments in CPR seemed to help understand the related course material and whether CPR helped to improve critical thinking. The last question displays some disagreement between the two groups: the 2012 group preferred turning in a “regular” paper, while the 2013 seemed to slightly prefer CPR. This disagreement could be that in that first semester, there was a learning curve for both the instructors and students. The second semester of using CPR went much more smoothly than the first.

Grading Accuracy

One assignment was selected to compare peer-generated CPR scores to instructor-generated scores on the same assignment. This was the final assignment given in the course, and was selected for two reasons: the students were familiar with CPR at that point, and this assignment nicely addressed one of the primary overall goals of the course. In the assignment, students had to locate and choose an article from a news source containing the phrase “statistically significant”. In their responses they had to summarize what the article was claiming to show, identify the null and alternative hypothesis, relate the key phrase to the concept of a p-value, and discuss the issue of practical significance versus statistical significance, all in context.

Twenty-five students were randomly selected from Dr. Pittard’s two classes from the spring 2013 semester (n=270) and their assignment was graded according to the same grading rubric as was used by the students’ peers in CPR. Four STA 210 instructors graded these assignments, all unaware of what the peer-graded results had been, and the text was graded out of ten points, with the instructors scoring a point for each “yes” response to the ten rubric questions.

	Mean of text	Std. dev. Of text	Mean of overall CPR assignment	Std. dev. Of overall CPR assignment
Peer graded in CPR	8.64	1.211	92.08	10.904
Instructor graded	8.14	0.635	*	*

A paired t-test was performed comparing the mean student text score and the mean instructor text score to the mean overall CPR score. The p-value for both t-tests was less than .0001, indicating that, on average, the students score the assignments higher than the instructors. Although the results are statistically significant, there is a question of whether the text-to-text comparison of scores differs practically. Considering that the difference in scores between peer-grading and instructor grading are only a half a letter grade on average is encouraging to the authors and given the direction of the difference, should be encouraging to students using CPR.

The instructor text-to-overall CPR grade comparison is reasonable since the students would receive either the CPR overall score, or the instructor’s score if the assignment were not being graded in CPR. The p-value was also less than .001 indicating again that the students on average are receiving a statistically higher grade using CPR versus instructor-graded assignments. This difference of a letter grade is substantial, but the point can be made that the additional learning achieved within the CPR process is substantial.

INSTRUCTOR EXPERIENCES AND REFLECTIONS

From an instructor’s perspective there are many outstanding features associated with CPR. The system dashboard is simple to use, yet powerful, allowing the instructor to customize timing for individual students, enter a peer grading session as a particular student, as well as changing grades on certain parts of a student’s assignment and controlling whether that change will flow out to affect others associated with that student’s work. It is often possible to get a student into a grading cycle late and still have that student complete all parts of the assignment.

A list of benefits includes easing the grading load of instructors, additional interaction with the assignment, and experience with evaluating a peers’ work. The process of peer review in itself is beneficial to students. In practice, students at some point will need to evaluate their peers work, whether in scientific papers or research proposals, or as an aspect of their profession; within the sciences or otherwise (Rudd et al., 2009). These reasons alone justify the use. There are some caveats to keep in mind, however.

First, the way in which CPR manages the instructor’s rubric is inflexibly holistic. For example, the most common question design is a Yes/No format. Most of us want the peer grader to score the assignment based on the number of "Yes" answers that were assigned. Unfortunately CPR does not do this automatically nor provide a toggle option for the instructor to force this to happen automatically. Instead, a final question, automatic to CPR, asks the peer grader to rate the text. We've had multiple situations where 8 out of 10 answers receive a "Yes" ("correct") but the final assessment was recorded as a 4 or 5.

Perhaps the most common error the students make is getting their documents into the system. In CPR there is an option of uploading the document or entering answers in a text box. The language of the CPR module is a bit biased toward the text box method, though this is not always a good solution where formatting, special characters, and graphics are required. The authors have found it much easier to just have students upload documents, but students have to "upload" the document and then "submit" the document in a second step. The "submit" button is isolated from the "attach" button and it is not clear to students that they need to press both. A very large number of students click "attach" and don't click "submit." The result is that the student's paper is not in the system officially so the student is not allowed to continue with the calibrations. There is no system indication of which missing papers are truly missing and which just need this final boost.

As mentioned in the previous section, there may be some grade inflation when using CPR. Granted, there are some built-in protections to address this: students are motivated by the tolerance limits to take the assignment seriously enough to not give everyone perfect scores, or they will lose points themselves. Still, if on average, students are grading the assignments easier then perhaps they should, then the CPR system cannot account for it. This concern could be addressed with a "spot check" on each assignment by the instructor or even by a TA. A random selection of a few students' CPR results to control for possible grade inflation may take care of this issue.

CONCLUSION

Overall, the authors believe that utilizing CPR for these assignments has been successful given the constraints faced with large lectures. For the students to benefit most from the use of CPR for such assignments, the instructor must explain the process fully, provide assistance in written form (the steps of the process, the timing of each step of the assignment, the due dates), and explain the rationale behind the use of CPR for these assignments. Students can easily come away with "they just want us to do their job for them" if the motivation is not explained.

REFERENCES

Scheaffer, R. L., & Stasney, E. A. (2004). The state of undergraduate education in statistics: A report from the CBMS 2000. *The American Statistician, 56*(4), 165-271

Dupuis, D. N., Medhanie, A., Harwell, M., LeBeau, B., Monson, D., & Post, T. R. (2012). A multi-institutional study of the relationship between high school mathematics achievement and performance in introductory COLLEGE statistics. *Statistics Education Research Journal, 11*(1), 4-20.

Lutzer, D. J., Rodi, S. B., Kirkman, E. E., & Maxwell, J. W. (2007). *Statistical abstract of undergraduate programs in the mathematical sciences in the United States: Fall 2005 CBMS survey*. Providence, RI: American Mathematical Society

Walvoord, M. E, Hoefnagels, M. H., Gaffin, D. D., Chumchal, M. M., & Long, D. A. (2008). An analysis of Calibrated Peer Review (CPR) in a science lecture classroom, *Journal of College Science Teaching, 37*(4), 66-73.

Margerum, L.D., Gulsrud, M., Manlapez, R., Rebong, R., & Love, A. (2007). Application of Calibrated Peer Review (CPR) writing assignments to enhance experiments with an environmental chemistry focus. *Journal of Chemistry Education, 84*(2), 292-295.

Carlson, P.A., & Berry, F. C. (2003). Calibrated Peer Review™ and assessing learning outcomes. *Proceedings of the ASEE/IEEE Frontiers in Education Conference* (pp. F3E1-F3E6).

McCarty, T., Parkes, M. V., Anderson, T. T., Mines, J., Skipper, B. J., & Grebosky, J. (2005). Improved patient notes from medical students during web-based teaching using faculty-calibrated peer review and self-assessment. *Academic Medicine, 80*(10), S67-S70.

Prichard, J. R. (2005). Writing to learn: An evaluation of the Calibrated Peer Review™ program in two neuroscience courses. *Journal of Undergraduate Neuroscience Education,4*(1), A34-A39.