AC 2008-1113: USING CALIBRATED PEER REVIEW AS A TEACHING TOOL FOR STRUCTURAL TECHNOLOGY IN ARCHITECTURE

Anne Nichols, Texas A&M University

Dr. Nichols is an Assistant Professor of Architecture at Texas A&M University. She teaches structural analysis, design, and planning at the undergraduate and graduate level. She is a civil engineer with research interests in the structural mechanics and modeling of masonry and cement materials.

Using Calibrated Peer Review as a Teaching Tool for Structural Technology in Architecture

Abstract

Calibrated Peer Review (CPR)™ is a web-based software tool for incorporating writing assignments in course that are not typically writing intensive. The intent is for students to write *and* critique the work of their peers on technical topics by learning to calibrate writing samples and then anonymously reviewing a subset of their classmates writing assignments, freeing the instructor from the time consuming task of grading every student's work.

This learning tool was used for a required graduate course in architectural structural systems in the Master of Architecture program at Texas A&M University. The student learning outcome was to improve the performance of a written term report on an architectural building case study conducted by a team of first year graduates through practice and exposure to varied levels of quality writing, and to reinforce the need for academic integrity with respect to the incorporation of non-original work.

This paper will present the analysis of the scored data and student performance with respect to the CPR assignments, their originality, and term report quality. The student feed back from directly after the assignments and at the conclusion of the semester will be presented, along with an analysis of that feedback and the effectiveness of the learning tool.

Introduction

Within a graduate professional degree program for Architecture, knowledge of environmental systems (mechanical, electrical and plumbing) and structural systems is necessary to ensure good design and to obtain licensure, but is secondary to architectural design which is what attracts students to the profession.

The integration of these subject areas within design through an architectural education has long been discussed and debated, as Comprehensive Design is an important student performance required for accreditation from the National Architecture Accrediting Board (NAAB).[1] Approaches to achieving integration have had varying levels of success, due, in part, to the offering of these subjects as traditional lecture courses.[2]

Within one such traditional lecture course in structural systems and planning, the graduate students were assigned a team project of a case study of an architectural building of their choice to demonstrate through problem-based learning an application of the course material and show a correlation between design and system application. The assignment required a short report documenting the case with examples and illustrations of the structural system(s) or members, computer analysis, and any other supporting evidence showing the application of the subject matter in the qualitative and quantitative analysis. The teams also presented brief slide show presentations to the class. The graphic design of the presentations were of high quality, but the report writing was often painful to read, lacked clear organization, and was of much lower

quality than the graphic design with improper referencing, overuse of quoted material and lack of quotations for previously published material, usually of digital form (text from the internet).

The misuse of non-original work was addressed by providing the students with the academic integrity policy for the university, and requiring the use of the web based tool, Turnitin[3], which generates an originality report against a large database of submitted reports, the internet and commercial databases of journal articles and periodicals.

The quality of the reports was initially addressed by providing handouts to the class for writing well and on organizing content. Importance was also placed on writing for their professional careers as well as in the field exam they were to submit at the start of their second year of graduate studies. But the resistance to writing was verbalized as a complaint that they "already did that as undergraduates!" (An additional student performance requirement for accreditation is the ability to read, write, listen, and speak effectively.)[1]

Calibrated Peer Review

Calibrated Peer Review (CPR)[4], which is in use by over 600 institutions, is a web-based tool that allows peer-to-peer evaluation of writing assignments. The intent of the tool is to expose students to good writing in a technical course while instructing them on how to identify and access it[5]. The tool is available at Texas A&M and used in such courses, among others, as chemistry, math, physics, and psychology.

The basic process of a CPR assignment begins with the writing of an essay based on guiding questions and selected resources. For this course, the essays were submitted and evaluated with Turnitin prior to submission to the CPR web site. Turnitin is a web-based tool for plagiarism prevention and is licensed for use at Texas A&M. Turnitin allows the instructor to see the originality reports on the writing for each student, and the instructor has the option of allowing the students to view their report in order to correct and revise. The report identifies matching text and phrases to specific sources with a web link, and shows the percentage of the report that matches each source, as in Figure 1, for example. A high match rate suggests copied work and that the matches should be evaluated because it does not indicate if the matched material was properly quoted and cited. When the students view their report, they may be rather surprised by their match rate and will be motivate to reduce the chance of a plagiarism "audit" by the instructor.

Upon submitting their writing to CPR, they proceed to a calibration exercise. The students are graded by the program on their ability to evaluate low quality, medium quality, and high quality writing samples based on content and style criteria questions. The questions provide answer choices and an overall rating from 1-10 is required. The questions can also require the students to enter an explanation for their answer choice. There is a corresponding key the students are graded against (see Figure 2). The style criteria for these exercises were very basic. Could they identify an introduction, a conclusion, and if the writing was free of spelling and grammatical errors? Four questions were based on style while six questions were based on required content. Students then received a review competency index from 1-6 (low-high) for their calibration skill.

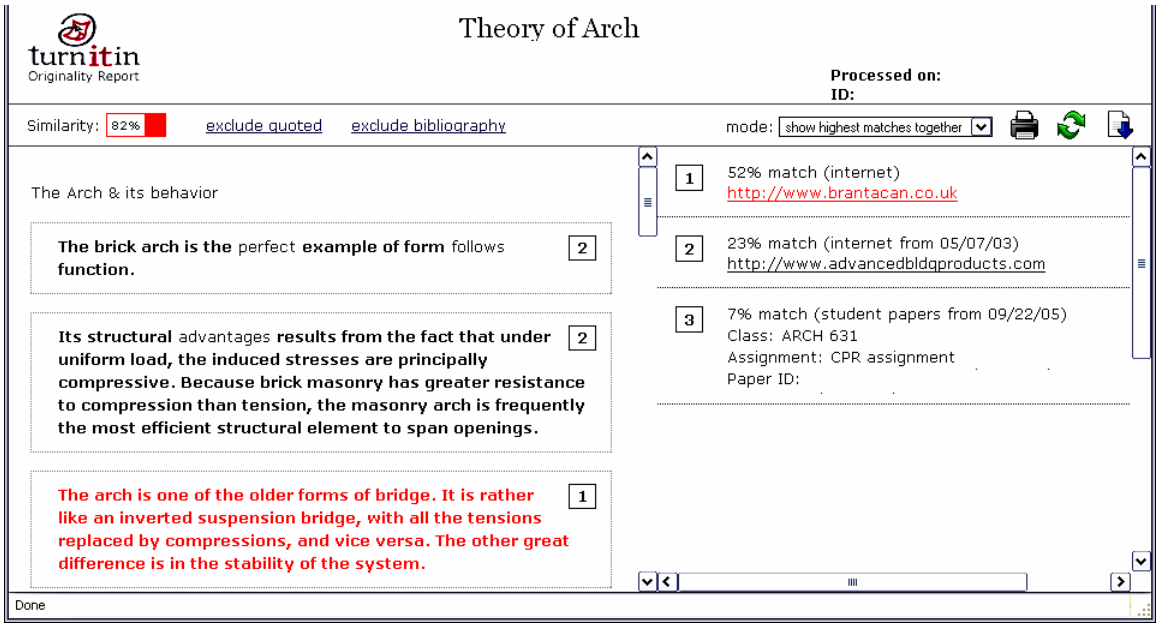


Figure 1. Example Originality Report

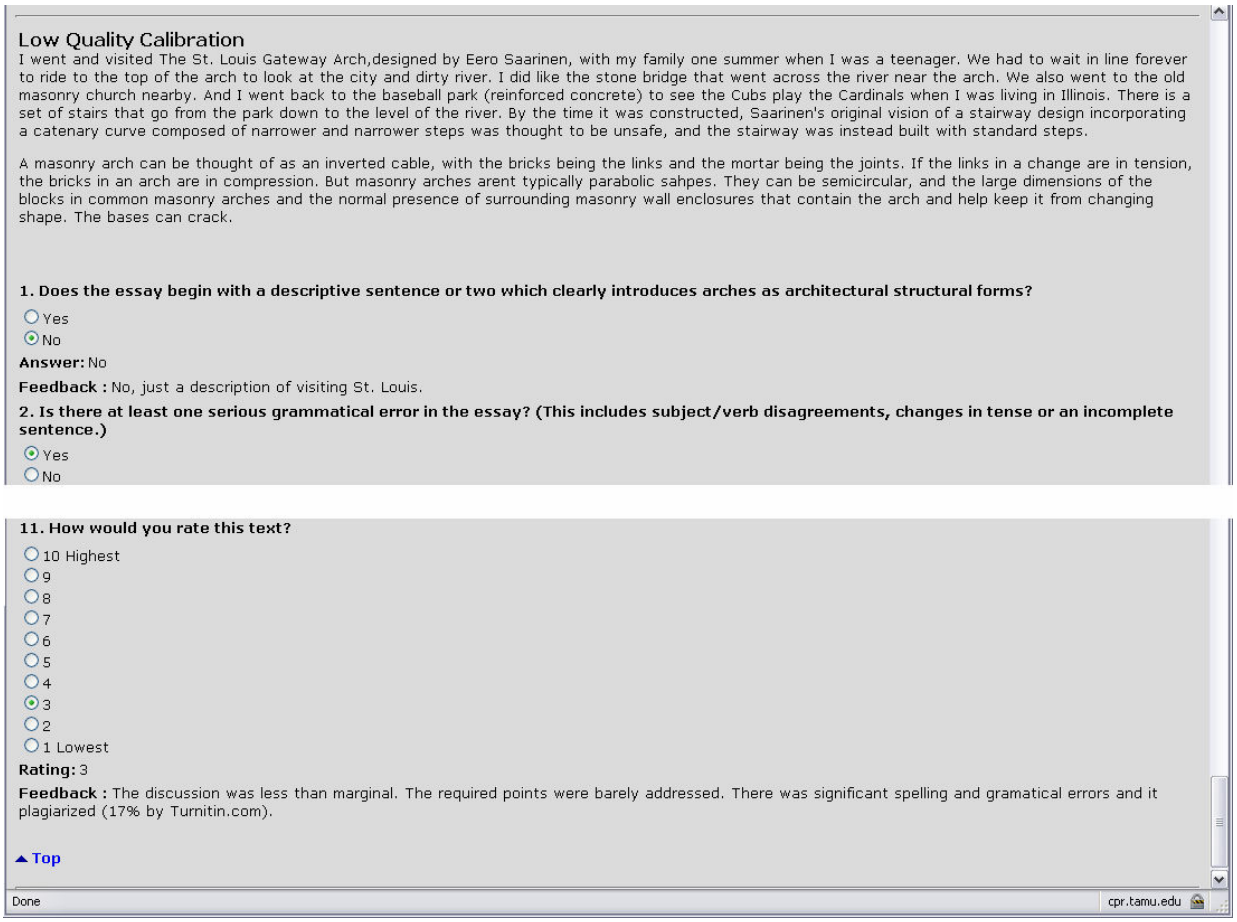
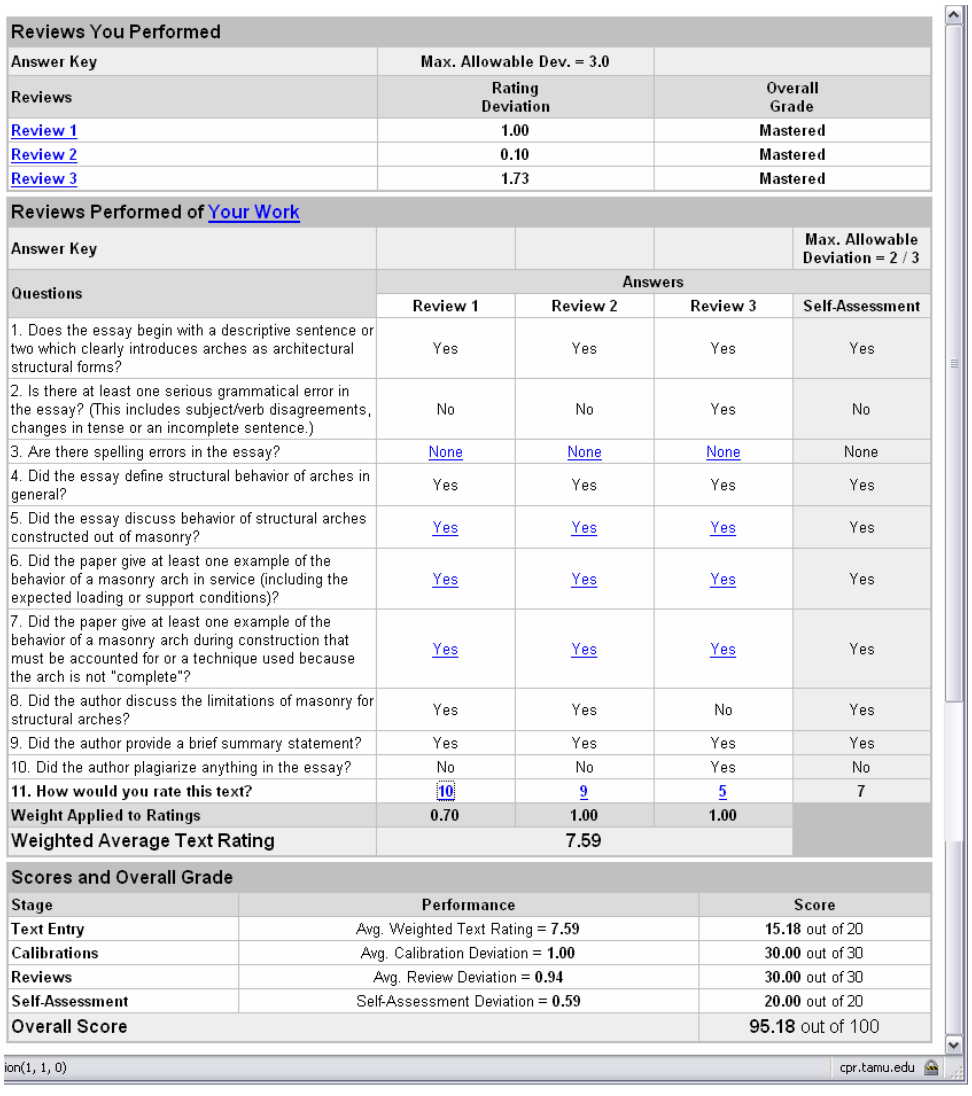


Figure 2. Low Calibration Key

When students successfully pass the calibration exercise, they blindly review written work by three of their peers, and they also reviewed their own writing. Their blind review ratings are compared to a weighted average rating for the work and have to fall within an allowable deviation chosen by the instructor (at low, moderate, or high difficulty) to be considered "mastered". The weighting factors are automatically assigned based on the review competency index of each reviewer. The squared weighting factors are used to compute the weighted average text rating. This prevents a poor reviewer from inflating (or deflating) an average text rating.

The overall scores are based on points for the text entry from the weighted average text rating, calibration mastery, having review rates within range, and a self-assessment rate within range. The proportions of these four areas to the total can be set by the instructor, and were 30, 30, 30, and 10%, respectively, for these assignments. An example student report is presented in Figure 3. The students also have access to the review comments by the three anonymous reviewers.



Reviews You Performed

Answer Key	Max. Allowable Dev. = 3.0	
Reviews	Rating Deviation	Overall Grade
Review 1	1.00	Mastered
Review 2	0.10	Mastered
Review 3	1.73	Mastered

Reviews Performed of Your Work

Answer Key				Max. Allowable Deviation = 2 / 3
Questions	Answers			
	Review 1	Review 2	Review 3	Self-Assessment
1. Does the essay begin with a descriptive sentence or two which clearly introduces arches as architectural structural forms?	Yes	Yes	Yes	Yes
2. Is there at least one serious grammatical error in the essay? (This includes subject/verb disagreements, changes in tense or an incomplete sentence.)	No	No	Yes	No
3. Are there spelling errors in the essay?	None	None	None	None
4. Did the essay define structural behavior of arches in general?	Yes	Yes	Yes	Yes
5. Did the essay discuss behavior of structural arches constructed out of masonry?	Yes	Yes	Yes	Yes
6. Did the paper give at least one example of the behavior of a masonry arch in service (including the expected loading or support conditions)?	Yes	Yes	Yes	Yes
7. Did the paper give at least one example of the behavior of a masonry arch during construction that must be accounted for or a technique used because the arch is not "complete"?	Yes	Yes	Yes	Yes
8. Did the author discuss the limitations of masonry for structural arches?	Yes	Yes	No	Yes
9. Did the author provide a brief summary statement?	Yes	Yes	Yes	Yes
10. Did the author plagiarize anything in the essay?	No	No	Yes	No
11. How would you rate this text?	10	9	5	7
Weight Applied to Ratings	**0.70**	**1.00**	**1.00**	
Weighted Average Text Rating		**7.59**		

Scores and Overall Grade

Stage	Performance	Score
Text Entry	Avg. Weighted Text Rating = **7.59**	**15.18** out of 20
Calibrations	Avg. Calibration Deviation = **1.00**	**30.00** out of 30
Reviews	Avg. Review Deviation = **0.94**	**30.00** out of 30
Self-Assessment	Self-Assessment Deviation = **0.59**	**20.00** out of 20
Overall Score		**95.18** out of 100

ion(1, 1, 0) cpr.tamu.edu

Figure 3. Student Results Example

Implementation

The resistance of students to using this software have been reported as timidity and perceived risk[6]. To overcome the doubt and fear, one strategy is to provide one practice assignment for full credit (little risk) upon completion, and then provided subsequent assignments with increased worth based on calibrated score results.

The graduate students in structural systems and planning were given a trial assignment to become familiar with the program and with masonry arch terminology, to demonstrate knowledge of the behavior and anticipated loading of masonry arches using "structural" language, and to learn to judge whether others were using "structural" language. The sources they could use for their compositions were the text and supplied reading material. They were instructed to include two examples of structural masonry arches when addressing the following questions:

1. What is the structural behavior (strength & serviceability) of an arch and why are they advantageous for long span structures? Be sure to discuss anticipated loads and any required conditions.
2. What are the characteristics of masonry that makes it suitable for structural arches?
3. What are the limitations of masonry for structural arches?
4. What structural behavior must be considered during construction?
5. What structural behavior must be considered in service?

The scoring for this assignment was set to low difficulty, which meant that for each calibration (low, medium and high) the student was required to answer 50% of style questions correctly and 50% of content questions correctly while not deviating by more than 3 points from the calibration text rating to master a calibration and receive credit. The student could not deviate by more than 3 points from the weighted average text rating to master a review of a peer's writing and receive credit. And to receive full credit for assessing their own work (self assessment), the student could not deviated by more than 2 points from the weighted average text rating. They could receive half credit if they deviated more than 2 points but less than 3 points from the weighted average text rating. Figure 3 shows an example of scoring results.

The second CPR assignment, scheduled for three weeks later, was assigned for credit based on the overall score. In this assignment, the students were to become familiar with membrane, net and shell structure terminology, demonstrate knowledge of the behavior and anticipated loading of membrane, net and shell structures using "structural" language, hone their writing skills using the feedback from peers on the introductory assignment, and hone their assessment skills of how knowledge of membrane, net and shell structures is presented. The sources they could use, again, came from the text and supplied reading material. They were instructed to choose between the topics of air inflated membranes, air supported membranes, tensile nets or shells, and include one example of a structural behavior or problem that must be considered in design and a corresponding solution when addressing the following questions:

1. What are the types of loads that are suitable for your selection of an air supported membrane, air inflated membrane, tensile net or shell structure? What are the loads that are unsuitable?
2. What are all the internal reactions and stresses in the membrane or shell?

3. What are the special conditions or problems that must be considered for design? Are they due to the loads or the reactions or the supports or the materials?
4. What are the solutions to the conditions or problems?
5. What are the material properties that are important for resisting stresses and for serviceability (i.e. the important issues of use that aren't related to strength)?

The scoring level for this assignment was set to moderate difficulty, with the expectation that the students would have motivation to do well for the higher risk assignment and had gained confidence and experience with the trial assignment. Moderate difficulty meant that for each calibration (low, medium and high) the student was required to answer 65% of style questions correctly and 65% of content questions correctly while not deviating by more than 2.5 points from the calibration text rating to master a calibration and receive credit. The student could not deviate by more than 2.5 points from the weighted average text rating to master a review of a peer's writing and receive credit. And to receive full credit for assessing their own work (self assessment), the student could not deviate by more than 1.5 points from the weighted average text rating. They could receive half credit if they deviated more than 1.5 points but less than 2.5 point from the weighted average text rating.

Performance Measures

For the low-risk trial assignment (full credit upon completion), the reviewer competency indexes were high, ranging from 4 to 6 with an average of 5.23 and standard deviation of 0.91, while the peer reviewed weighted average text rating ranged from 3.9 to 9, with most ratings above 7 and corresponding well with the self assessments. When self assessments were not within acceptable deviation for the low scoring (2 points), it was usual that the rating the student gave their own work was much higher (3.5 points above on average) than the weighted average text rating from their peers.

For the credit assignment, the reviewer competency indexes were more scattered, ranging from 2 to 6 with an average of 3.83 and a standard deviation of 1.5, while the peer reviewed weighted average text rating ranged from 4.5 to 9.5, with most ratings above 8 and corresponding well with the self assessments. When self assessments were not within acceptable deviation for the moderate scoring (1.5 points), it was usual that the rating the student gave their own work was much higher (2.5 points above on average) than the weighted average text rating from their peers.

The software allows student scoring to be modified or adjusted by the instructor after the assignment is completed. In the credit assignment, over half of the scores were adjusted, mostly by judging that the self assessment values were close to the acceptable range and/or a calibration was close to passing. In rare instances, there was a reviewer with an extremely low reviewer competency index, and their input was removed from the weighted average text calculation. The scores that were re-evaluated, were done so upon request by students who were obviously motivated by the lower results when compared to the trial assignment.

To evaluate the effect of the change from low to medium difficulty in scoring level, the low difficulty level was applied to the credit assignment data. The percentage of students noticing a

substantial reduction from the overall score in the trial assignment to the credit assignment reduces from 43% to 30%. Figure 4 shows the shift in adjusted scores with the moderate difficulty scoring level to the unadjusted low difficulty scoring level. The Student's t-test, which can determine if two sample population means are statistically equivalent, was used to compare the overall scores in the trial assignment to the credit assignment using a 95% confidence level of similarity. The analysis indicates that there is a statistical difference between the trial assignment scores (low difficulty) and the unadjusted credit assignment scores (moderate difficulty), while there is no statistical difference for the credit assignment between the adjusted scores at the moderate difficulty level and at the low difficulty level. It is interesting to note that the weighted average text rating of the credit assignment is not statistically different between the moderate difficulty level and the low difficulty level scorings, but that the review competency index values are. There is a sharp decline in low review competency index values with an increase in high competency index values. The reviewer competency index values are not statistically different when low difficulty level scoring is used for both assignments.

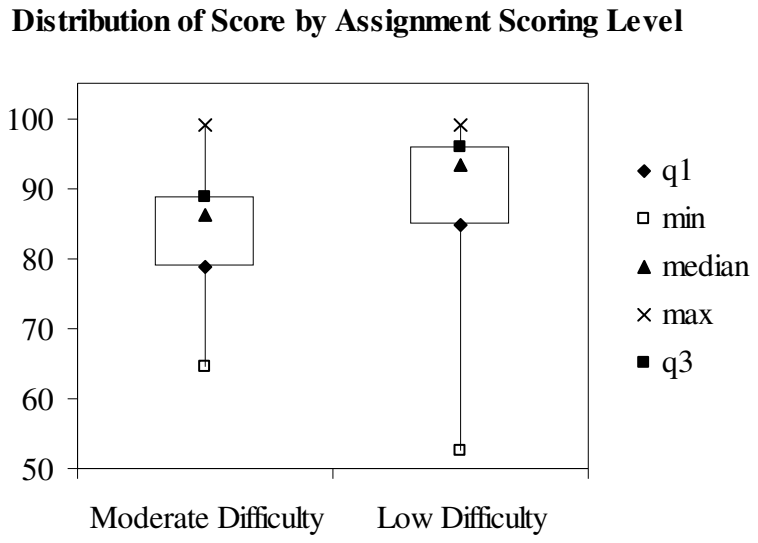


Figure 4.

The weighted average text ratings were statistically different by Student's t-test analysis between the trial and credit assignment (either scoring level) with higher mean and smaller deviation for the credit assignment. This indicates that the reviewers were doing a much better job of coming to a consensus on the quality of the work being reviewed.

The originality report match percentages determined by Turnitin for the essays were also compared for the trial and credit assignments. The average match rating was much lower and less variable in the credit assignment than for the trial, but the originality of the essays were not statistically different by Student's t-test analysis. When the match percentages for the term reports were compared to the credit assignment, the Student's t-test indicated they were statistically different. The reports were team efforts, commonly with one main author and editors for various sections of the report they were responsible for. And although their skills for recognizing and reviewing quality work had shown improvement, there was not an equivalent exercise spent on editing and improvement of their report writing.

Analysis

The term report scores, which are correlated to grading of style as well as content, can be used to evaluate the student learning outcome of improved performance with the use of the CPR software. Five classes of term report scores were evaluated with respect to the class participating in the CPR assignments. The changes to the report requirements made from class to class and the overall quality of each class with respect to the average class grade are also considered.

The average term report scores for the classes which were given specific instructions on content and format for the report, in addition to the academic integrity policy for the university are shown in Figure 5. In class 2, the use of Turnitin became required and it was recommended that the quantitative results come from computer analysis. In class 3, CPR assignments were given in addition to assignments using structural analysis software. The quantitative results in the report were required by computer analysis. In addition, review of specific building case studies in reinforced concrete, steel, and timber were presented during lecture to show examples. In class 4, the only change from class 3 was that no CPR assignments were given.

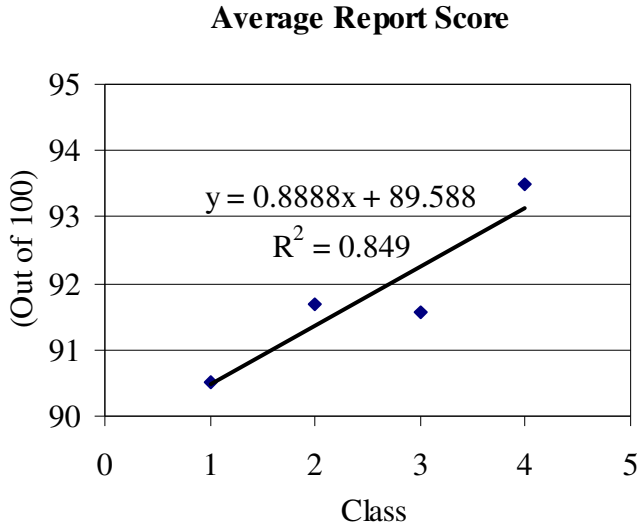


Figure 5.

There is a positive trend in the average term report score by consecutive class, but no direct indication that the CPR assignments contributed to an increase in student performance by report score from class 2 to 3, so grade point average by class was examined. Figure 6 shows the overall performance of the class by term grade relative to the average report score.

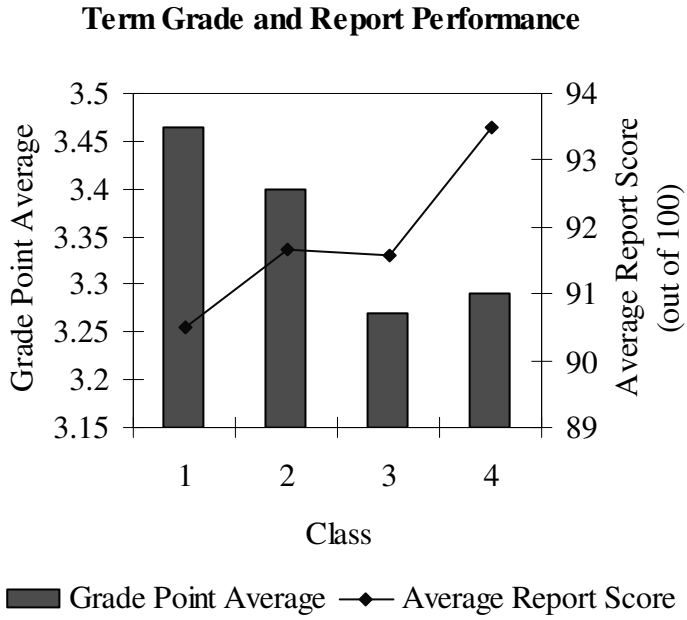


Figure 6.

Although the grade point average was largest for class 1, it should be noted that the contribution of the report to the term grade was 10%, while in subsequent classes it contributed 20%. What is more interesting, is that the grade point averages of the classes that were required to use structural analysis software to investigate their case buildings (classes 3 and 4) were lower as the average report scores continued to rise. There were no noticeable changes in the performance in the other areas contributing to the term grade (assignments and exams), so it may be that the overall quality of the students in the class using CPR could have been lower than the previous class.

What is also of interest is that the report scores were statistically different by the Student's t-test analysis (95% confidence level) between classes 1 and 4, and between classes 2 and 4. No statistical difference was evident when the report scores of the class using CPR (class 3) was compared to any of the other classes.

Conclusions

Although there is no direct correlation between the effectiveness of reviewing skills in the project report quality as reflected by score, the grader of the reports was much more satisfied with the writing quality. This probably allowed for finer scrutiny of the content required in the report; somewhat following the expectation that CPR scoring of low difficulty on the practice assignment could be increased to moderate difficulty on the scored assignment.

It does appear that between the two CPR assignments, the students did not change their calibration effort or inflated opinions of their own work, but were much better at rating their peer's works. There was no feedback from the students about the trial assignment, but a good deal of informal feedback when students requested that their credit assignment scores be re-

evaluated. Quite often the complaint was from a student who thought very highly of their own work and could not see why their peers did not agree. One such student argued that the work should not be graded on style, but on content only. This was in contrast to a student who could write effectively, but complained that the course instructor, not them, should be the one to have to read and grade low quality work.

There were anonymous comments about the CPR assignments in the formal student evaluation of the course. When the assignments were mentioned by a student there was significant indignation directed to the instructor for assigning the writing, with one extreme demand that the instructor be removed from the course. These student evaluations were reviewed by a teaching consultant who concluded that the student perceptions of the course prior to or at the start of the semester were much different than what they had experienced. The recommendation was to make the expectations as clear as possible with detailed, well stated objectives, including the objective of communication, and for the alignment between assignments and goals. It was not suggested that the use of writing tool should be discontinued.

The learning objectives subsequently included in the syllabus that directly address these issues state:

- *...The student will be able to evaluate their own skills, or lack thereof, with respect to reading and comprehension of structural concepts,* ***clarity*** *of written communication, reasonable determination of* ***precision*** *in numerical data, and* ***accuracy*** *of computations.*
- *The student will be able to articulate the physical phenomena, behavior and design criteria which influence structural space and form. (****depth)*** ***...*** *The student will draw upon existing organizational and communication skills to clearly present concepts and personal interpretation of structural knowledge in writing assignments and examinations (****clarity, precision, accuracy, relevance, depth, breadth, logic, significance****).*
- *... In addition, the student will be able to evaluate the comprehension of concepts, clarity of communication of these concepts or calculations, and the precision and accuracy of the data used in the computations in the work of their peers. ...*

Summary

The use of CPR writing assignments over technical material that did not lend itself easily to calculation exercises was an attractive tool to get students in an architectural structures systems course to practice writing and gain exposure to varied levels of quality writing on course topics. The primary motivation for using the tool was the perceived lack of quality in written project reports by students with significantly strong graphical communication skills. As the weight of the project report grade was substantial in the term grade calculation, it was also desirable that the student performance improve for the sake of their own grade point averages.

The two CPR assignments introduced them to the software, required them to demonstrate calibration skills for low, medium and high quality writing, and to blindly review three works by their peers and self-assess their work. Materials on writing and organization, in addition to materials on academic integrity and Turnitin software to check originality were also used. The evaluation of the scoring from the CPR assignments shows that the students were improving at

reviewing peer work and in their own writing by the increase in the weighted average text ratings, but they were not getting better in calibrating.

The influence of the CPR assignments on the project quality was investigated with respect to average report score and then to class grade point average. There was not a direct correlation found, but there was a statistical difference from the accumulated changes in the class following the one using the CPR software. This could be attributed to the change in student expectations, feeling relief by not being forced to write (other than in essay exams and in the team project report), and clarity of goals, assessment objectives, and student outcomes.

The improvement in the perceived quality of the reports by the instructor allowed the content to be more readily interpreted and scrutinized, and there is some satisfaction in this, even though the students were not completely satisfied with the scoring outcomes and convinced of the need to be effective written communicators for their professional success in architecture.

While the motivation of instructor to use CPR assignments was to improve the quality of the group term project reports, the motivation of the students was to individually obtain a high assignment score. The use of the Calibrated Peer Review software tool to motivate individual performance on exam essays should help to bring the student learning outcome and performance satisfaction into alignment.

Bibliography

1. The National Accrediting Board, *NAAB Conditions for Accreditation For Professional Degree Programs in Architecture*, 2004 edition, Washington, DC.
2. Watson, D., "Architecture, Technology, and Environment", *The Journal of Architectural Education* (1997), Vol. 51, No. 2.
3. iParadigms, LLC., Turnitin, http://Turnitin.com (2008).
4. University of California Los Angeles, Calibrated Peer Review, http://cpr.molsci.ucla.edu/ (2001).
5. Robinson, R., "An Application to Increase Student Reading & Writing Skills", *The American Biology Teacher* (2001), Vol. 63, No. 7.
6. Keeney-Kennicutt, W., Bunersel, A.B., Simpson, N, "Overcoming Student Resistance to a Teaching Innovation", *International Journal for the Scholarship of Teaching and Learning* (2008), Vol. 2, No. 1.