”Blinded” Grading Rubrics for Bioengineering Lab Reports (Work in Progress)

Mrs. Yanfen Li, University of Illinois, Urbana-Champaign

Yanfen Li is a Ph.D student in the Department of Bioengineering at the University of Illinois at Urbana-Champaign working under Dr. Kris Kilian. Her research focus is on biomaterials and tissue engineering.

Philip A. Jensen, Rocky Mountain College
Prof. Paul Anthony Jensen, University of Illinois, Urbana-Champaign

Paul Jensen is a Research Assistant Professor at the University of Illinois at Urbana-Champaign in the Department of Bioengineering and the Carl R. Woese Institute for Genomic Biology. He received bachelor degrees in chemical and biomedical engineering at the University of Minnesota and a Ph.D. in biomedical engineering from the University of Virginia. Paul completed postdoctoral training at Boston College before joining the University of Illinois in 2016.

Prof. Karin Jensen, University of Illinois, Urbana-Champaign

Karin Jensen is a Teaching Assistant Professor in bioengineering at the University of Illinois at Urbana-Champaign. At UIUC she teaches undergraduate courses and serves as an academic advisor. Before joining UIUC she completed a post-doctoral fellowship at Sanofi Oncology in Cambridge, MA. She earned a bachelor’s degree in biological engineering from Cornell University and a Ph.D. in biomedical engineering from the University of Virginia.

Work in Progress: “Blinded” Rubrics for Bioengineering Lab Reports

Background

Laboratory courses are typically divided into several sections each led by different teaching assistants, raising questions about consistency of grading across sections. Previous work has demonstrated that TAs can assign a wide range of grades for similar work in engineering courses, resulting in what one set of authors called the “TA lottery” for students [1]. Other authors have noted that writing assignments might be especially prone to inconsistent grading in STEM fields, as many graduate TAs may not have received formal training in writing assessment, while some may even lack confidence in their own writing skills, perhaps because English is not their native language [2, 3]. Given that many instructors are committed to including writing assignments in laboratory courses, in no small part because they may improve proficiency using the scientific method [4], how to best evaluate writing in laboratory courses remains an open question.

Part of the answer to that question involves the use of rubrics, which are widely used to standardize grading in large courses. To be sure, rubric-based grading is imperfect: As examples, consider that use of rubrics does not consistently improve students’ grades [5], and that subjective interpretation of a rubric can vary widely [6]. However, rubric-based grading can increase transparency in the grading process [7] and can demystify learning outcomes in a course generally [8]. Also, rubrics have been shown to decrease students’ complaints throughout a course, probably because of the factors listed above [9]. Rubrics can continue to have a place in large laboratory courses, especially as a means of clearly communicating learning objectives and standardizing grading.

Research Question

This study aims to document to what extent rubric design can influence inter-grader variability in a large laboratory course. Specifically, we test a counterintuitive hypothesis: that much of the variability between graders results not from subjective interpretation of the rubric’s qualitative criteria, as has been suggested, but from the subjective quantification of the evaluation of those qualitative criteria. Said another way, the graders may agree on which objectives were met, but not on how to score them.

Experimental Design

Over three semesters of a bioengineering laboratory course at the University of Illinois at Urbana-Champaign, TAs used rubrics that were either traditional (i.e. qualitative and quantitative) or experimental (i.e. qualitative). Before the TAs graded any assignment, they were provided with a list of expectations for the particular report and common errors observed from previous semesters.

During the first semester (Spring 2016), TAs used traditional rubrics and quantified their evaluations based on qualitative learning objectives. During subsequent semesters (Fall 2016, Spring 2017), TAs evaluated the work using the same qualitative criteria on the rubric, but did not quantify their evaluations. Instead, the course director collected the graded students' work from all sections and determined grades. The TAs' evaluations were quantified using a predetermined algorithm where each statement on the rubric granted a particular number of points. Each rubric statement is weighted differently as predetermined by the course director in accordance to class goals. For the experimental grading system, TAs were "blinded" to the point value of each statement on the rubric. Importantly, in all cases students and TAs were made aware of the relative weights of the components of the rubric (e.g. that interpretation of results was worth 20 points), but the feedback was largely qualitative in the experimental sections (e.g. the student did or did not meet an objective vs. the student earned an 18 of 20). That is, in the experimental sections, only the overall score on the assignment was quantitative—feedback on subsections of the rubric was strictly qualitative.

In all sections we also collected data regarding students' and TAs' perceptions of the course using surveys and the university's course-evaluation system.

Preliminary Results

	Traditional Score					
	Grader A	Grader B	Grader C	Grader D	Grader E	Grader F
Student A	65	45	63	65	73	48
Student B	89	75	87.5	100	94	83
Student C	92	91	95.5	98	98	95
Student D	87	63	95		96	
Student E	87	80	95		95	85
Student F	45	37	62	91	94	68

	Experimental Score					
	Grader A	Grader B	Grader C	Grader D	Grader E	Grader F
Student A	61.5	56	69	68	76.5	51
Student B	88.5	82.5	81	89	88.5	84
Student C	80.5	77	80.5	99	98	97
Student D	87	72.5	91		96	
Student E	88	79	83	93	97.5	91.5
Student F	37	33.5	68.5	90	88	84

Table 1: Grades assigned using traditional grading methods or the experimental system.

As expected, using a pre-determined algorithm for assigning points to qualitative rubric statements slightly decreased variation in TA grade assignments (Table 1). Overall, the variation in the grades assigned to each student decreased using the experimental system, from an average standard deviation of 11.3 (traditional) to 11.0 (experimental). These results were confounded by a seemingly anomalous grades for Student C; these points are far beyond the standard outlier rejection region (Thompson τ of 1.75 vs. rejection region boundary of 1.19). Removing this one set of grades changes the average standard deviations to 10.4 (traditional) vs. 7.4 (experimental), a difference that nears significance ($p = 0.06$, paired t-test).

Student perceptions about the grading systems were evaluated by end-of-semester course evaluations. Student feedback suggested that students viewed the strictly qualitative grading system as more fair (Figure 1). Complaints about points/grades dramatically decreased, with not a single student questioning points scored on any report when the rubrics were strictly qualitative other than the final score. Importantly, assignment grades were comparable between Spring 2016 (traditional) and Fall 2016 (experimental) semesters ($p > 0.3$, paired t-test). "Blinding" the students and TAs to the point values assigned to each statement on the rubric focused discussions solely on the qualitative learning objectives and how these could be improved in future assignments.

Preliminary results (Figure 2) show that TAs who graded both with the traditional and experimental rubrics reported decreased grading time and increased confidence when grading with the experimental rubric. After grading sets of reports using both the traditional and the experimental system, TAs were asked to rate their confidence in the scores they assigned for each grading system. In follow-up discussions, TAs reported that quantitative grading included time spent weighing small differences in point values and revising their qualitative evaluations based on the resultant scores. TAs described a motivation to make rubrics "fit" to the overall score that they anticipated instead of focusing on individual components of the rubric.

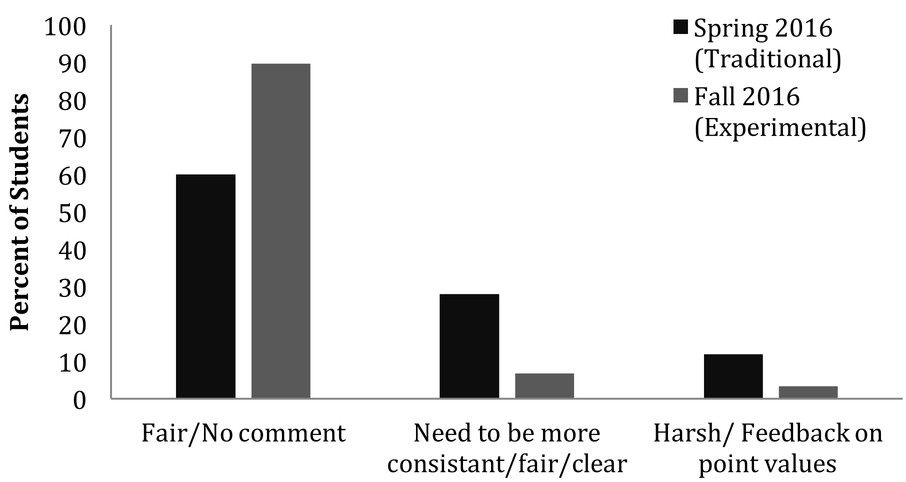

Figure 1. Increased perception of fair grading with experimental grading methods. Student feedback on course grading at the end of the Spring 2016 (traditional scoring), and Fall 2016 (experimental scoring) semesters were separated into categories based on feedback received. Students commenting on several categories were counted in each category. N= 25, 34 for Spring 2016, and Fall 2016 respectively.

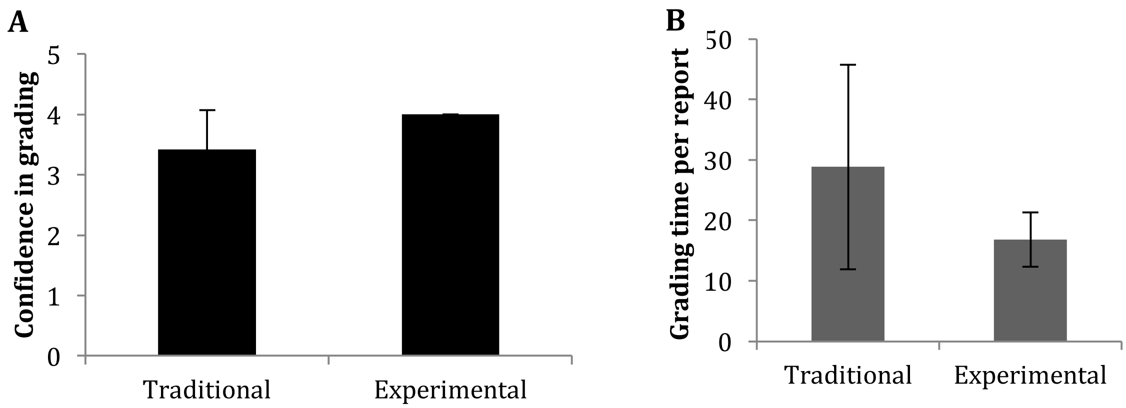

Figure 2. TA feedback on grading indicate a trend in increased confidence (Likert scale 1-5, where 5 is highest confidence) and decreased grading time when using experimental grading methods. Data are shown as mean +/- standard deviation (n=6).

Since TAs did not finalize scores, the "blinded" rubrics added an additional step to the grading process since qualitative rubrics needed to be summed using the predetermined point values. In the present study, the scores were tallied and report by an undegraduate grader who did not evaluate the reports. In the future, we plan to implement an automated system that would sum the student scores after TAs completed grading all the reports.

Conclusions

Our preliminary results suggest that "blinding" TAs from point values on rubrics can decrease both inter-grader variation and time spent grading. Additionally, TAs reported greater confidence in the assigned grades when utilizing the experimental grading method where they evaluated reports only qualitatively, suggesting that TAs are aware of the inter-grader variation introduced by assigning points on a rubric and are not confident in making these assignments when grading. We also found that students whose reports are graded strictly qualitatively (but receive one quantitative grade for the entire report) perceive grading as more fair compared to students whose reports were graded on a qualitative and quantitative rubric. We believe that the experimental system presented, in conjunction with TA training on the use of the rubric will improve inter-grader variation, TA confidence in assignment evaluations, and student perceptions of grading fairness. Ongoing studies will explore the validity of these findings by expanding the cohort of TAs and student reports evaluated.

References

1. Franey, S., A. Gregerson, and M.T. Braun. *Playing the TA Lottery A Study of How Teaching Assistants Impact Grades in Engineering Courses*. in *American Society for Engineering Education*. 2012. American Society for Engineering Education.

2. Powe, A. and J. Moorhead. *Grading lab reports effectively: using rubrics developed collaboratively by ECE and technical writing instructors*. in *Annual Conference of the American Society for Engineering Education, Chicago, IL*. 2006.
3. Brocato, J., J. Harden, and B. Chapman, *Improving the writing-evaluation abilities of graduate teaching assistants in ECE labs.* age, 2005. **10**: p. 1.
4. Libarkin, J. and G. Ording, *The utility of writing assignments in undergraduate bioscience.* CBE-Life Sciences Education, 2012. **11**(1): p. 39-46.
5. Jonsson, A. and G. Svingby, *The use of scoring rubrics: Reliability, validity and educational consequences.* Educational research review, 2007. **2**(2): p. 130-144.
6. Allen, D. and K. Tanner, *Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners.* CBE-Life Sciences Education, 2006. **5**(3): p. 197-203.
7. Frederiksen, J.R. and A. Collins, *A systems approach to educational testing.* Educational researcher, 1989. **18**(9): p. 27-32.
8. Bissell, A.N. and P.P. Lemons, *A new method for assessing critical thinking in the classroom.* BioScience, 2006. **56**(1): p. 66-72.
9. Barney, S., et al., *Improving students with rubric-based self-assessment and oral feedback.* IEEE Transactions on Education, 2012. **55**(3): p. 319-325.

Appendix A: Rubric presented to students and TAs

	Excellent	Satisfactory	Unsatisfactory
Objective/ Purpose (10)	The report is written for the correct audience. (3) The report contains the necessary information in various sections of the report. (2) The report is well-organized. (3) The objectives are clear. (2)	The report is largely written for the correct audience. The report contains most of the necessary information in the appropriate sections of the report. The report is mostly well-organized. The objectives are mostly clear.	The audience of the report is not clear or is not written for the intended audience. Information is scattered throughout the report and/or information is missing. The report would benefit from reorganization. The objectives are largely unclear and need to be more clearly stated.
Data Presentation (35)	Figures and diagrams are well-designed and are the best representation of the data. (5) Figures/diagrams use space in the report effectively (i.e. no large white areas, images are cropped appropriately). (5) There is no extra information, coloring, gridlines or other features on the figures/diagrams. (5) All axes, symbols, legends, etc. are appropriately labeled with correct units. (5) Figure captions contain the appropriate details for the data presented. (10) All figures, diagrams, and tables have descriptive and succinct titles. (5)	Figures and diagrams adequately show the data and are mostly well-designed with a few minor issues. Figures/diagrams mostly use space in effectively. There are a few instances of extra information, coloring, gridlines or other features on the figures/diagrams. Axes, symbols, legends, etc. are appropriately labeled with correct units with one or two minor exceptions. Figure captions contain most of the information needed to interpret the figure but may be missing one or two minor details. All figures, diagrams, and tables have appropriate titles that contain the necessary information to interpret the data.	Figures and diagrams are not well-designed and have several issues and the choice of data presentation is an inaccurate representation of the data collected and/or data is missing. Figures/diagrams do not use space effectively and would benefit from redesign. There are several instances of extra information, coloring, gridlines, etc. on the figures/diagrams. Axes, symbols, legends, etc. are not labeled, have incorrect units, or are missing. Figure captions are lacking key pieces of information or experimental details. Some figures, diagrams, and tables lack appropriate titles.
Analysis (15)	Data analysis is accurate and complete. (10) Calculations and/or models are fully discussed and all units are shown. (5)	Data analysis is mostly accurate with few minor errors. Calculations and/or models are discussed and all units are shown.	Data analysis is largely incorrect and/or incomplete. Discussion of calculations is incomplete and/or incorrect and units are incorrect and/or missing.
Discussion (20)	The report shows the author has a thorough understanding of the experiments performed and data collected. (5) All statements are accurate and appropriate scientific vocabulary is used. (5) Sources of error are identified and thoroughly discussed. (5) Conclusions drawn from the data are reasonable given the data collected and are fully discussed. (5)	The report shows the author has a satisfactory understanding of the experiments performed and data collected. Statements are mostly accurate but may have a few minor errors or misconceptions. Appropriate scientific vocabulary is used. Sources of error are identified but not fully discussed. Appropriate conclusions are drawn from the data but are not fully discussed.	The report shows a lack of understanding of several important concepts regarding the experiments performed and the data collected. Statements are inaccurate and there are several instances where scientific vocabulary is not used or used improperly. Sources of error are not addressed or are not reasonable or are not supported. Conclusions are not drawn from the data or are unreasonable or not supported.
Writing (20)	A strong, formal voice is used throughout the report. (2) The writing is concise and word choices are precise. (5) All figures and tables are appropriately referenced in the text. (2) The report contains no spelling or grammatical errors and is easy to read. (3) There is no ambiguity in the writing. (5) The report looks professional and follows assignment guidelines (page limits, etc.). (2) References are properly formatted. (1)	A formal voice is used throughout the report. The writing is mostly concise and word choices are mostly precise. All figures, tables, and diagrams are referenced in the text but not in the appropriate place. The report contains one or two minor spelling or grammatical errors and is easy to read. There is little ambiguity in the writing. The report looks mostly professional and follows assignment guidelines. References have one or two minor formatting issues.	Some of the writing and vocabulary in the report is informal or inappropriate for a technical document. The writing is wordy and includes unnecessary lead-ins and/or is repetitive. Word choice is largely not precise and is often unclear. Some figures, tables, or diagrams are not referenced in the text and/or are inappropriately referenced. The report contains several spelling or grammatical errors and is difficult to read. There is consistent ambiguity in writing and word choice. The formatting and appearance of the report is unprofessional and distracting and/or assignment guidelines were not followed. References are not formatted properly.

Score:

General comments:

Appendix B: Point values assigned

	Excellent	Satisfactory	Unsatisfactory
Objective/ Purpose (10)	The report is written for the correct audience. (3) The report contains the necessary information in various sections of the report. (2) The report is well-organized. (3) The objectives are clear. (2)	The report is largely written for the correct audience. (2) The report contains most of the necessary information in the appropriate sections of the report. (1) The report is mostly well-organized. (2) The objectives are mostly clear. (1)	The audience of the report is not clear or is not written for the intended audience. (1) Information is scattered throughout the report and/or information is missing. (0) The report would benefit from reorganization. (1) The objectives are largely unclear and need to be more clearly stated. (0)
Data Presentation (35)	Figures and diagrams are well-designed and are the best representation of the data. (5) Figures/diagrams use space in the report effectively (i.e. no large white areas, images are cropped appropriately). (5) There is no extra information, coloring, gridlines or other features on the figures/diagrams. (5) All axes, symbols, legends, etc. are appropriately labeled with correct units. (5) Figure captions contain the appropriate details for the data presented. (10) All figures, diagrams, and tables have descriptive and succinct titles. (5)	Figures and diagrams adequately show the data and are mostly well-designed with a few minor issues. (4) Figures/diagrams mostly use space in effectively. (4) There are a few instances of extra information, coloring, gridlines or other features on the figures/diagrams. (4) Axes, symbols, legends, etc. are appropriately labeled with correct units with one or two minor exceptions. (4) Figure captions contain most of the information needed to interpret the figure but may be missing one or two minor details. (7) All figures, diagrams, and tables have appropriate titles that contain the necessary information to interpret the data. (3)	Figures and diagrams are not well-designed and have several issues and the choice of data presentation is an inaccurate representation of the data collected and/or data is missing. (2) Figures/diagrams do not use space effectively and would benefit from redesign. (2) There are several instances of extra information, coloring, gridlines, etc. on the figures/diagrams. (2) Axes, symbols, legends, etc. are not labeled, have incorrect units, or are missing. (0) Figure captions are lacking key pieces of information or experimental details. (2) Some figures, diagrams, and tables lack appropriate titles. (1)
Analysis (15)	Data analysis is accurate and complete. (10) Calculations and/or models are fully discussed and all units are shown. (5)	Data analysis is mostly accurate with few minor errors. (7) Calculations and/or models are discussed and all units are shown. (3)	Data analysis is largely incorrect and/or incomplete. (2) Discussion of calculations is incomplete and/or incorrect and units are incorrect and/or missing. (2)
Discussion (20)	The report shows the author has a thorough understanding of the experiments performed and data collected. (5) All statements are accurate and appropriate scientific vocabulary is used. (5) Sources of error are identified and thoroughly discussed. (5) Conclusions drawn from the data are reasonable given the data collected and are fully discussed. (5)	The report shows the author has a satisfactory understanding of the experiments performed and data collected. (4) Statements are mostly accurate but may have a few minor errors or misconceptions. Appropriate scientific vocabulary is used. (3) Sources of error are identified but not fully discussed. (3) Appropriate conclusions are drawn from the data but are not fully discussed. (3)	The report shows a lack of understanding of several important concepts regarding the experiments performed and the data collected. (2) Statements are inaccurate and there are several instances where scientific vocabulary is not used or used improperly. (1) Sources of error are not addressed or are not reasonable or are not supported. (2) Conclusions are not drawn from the data or are unreasonable or not supported. (1)
Writing (20)	A strong, formal voice is used throughout the report. (2) The writing is concise and word choices are precise. (5) All figures and tables are appropriately referenced in the text. (2) The report contains no spelling or grammatical errors and is easy to read. (3) There is no ambiguity in the writing. (5) The report looks professional and follows assignment guidelines (page limits, etc.). (2) References are properly formatted. (1)	A formal voice is used throughout the report. (1) The writing is mostly concise and word choices are mostly precise. (3) All figures, tables, and diagrams are referenced in the text but not in the appropriate place. (1) The report contains one or two minor spelling or grammatical errors and is easy to read. (2) There is little ambiguity in the writing. (3) The report looks mostly professional and follows assignment guidelines. (1) References have one or two minor formatting issues. (0.5)	Some of the writing and vocabulary in the report is informal or inappropriate for a technical document. (0) The writing is wordy and includes unnecessary lead-ins and/or is repetitive. Word choice is largely not precise and is often unclear. (1) Some figures, tables, or diagrams are not referenced in the text and/or are inappropriately referenced. (0) The report contains several spelling or grammatical errors and is difficult to read. (0) There is consistent ambiguity in writing and word choice. (1) The formatting and appearance of the report is unprofessional and distracting and/or assignment guidelines were not followed. (0) References are not formatted properly. (0)

Score:

General comments: