ASEE 2019 Annual Conference CHARGED UP FOR THE NEXT 125 Years
126TH ANNUAL CONFERENCE & EXPOSITION
ASEE

Paper ID #26269

Developing Reliable Lab Rubrics Using Only Two Columns

Prof. Joshua A. Enszer, University of Delaware

Dr. Joshua Enszer is an associate professor in Chemical and Biomolecular Engineering at the University of Delaware. He has taught core and elective courses across the curriculum, from introduction to engineering science and material and energy balances to process control, capstone design, and mathematical modeling of chemical and environmental systems. His research interests include technology and learning in various incarnations: electronic portfolios as a means for assessment and professional development, implementation of computational tools across the chemical engineering curriculum, and game-based learning.

Developing Reliable Lab Rubrics Using Only Two Columns

Abstract

Rubrics have often been touted as an effective tool to communicate expectations and save time grading. One area of study is the improvement of rubrics to increase inter-rater reliability; that is, creating consistent rubrics such that multiple assessors are likely to assign the same score or designation to the same work. The number of columns or "standard levels" within a rubric is often up to debate: for a given criteria, should work be assessed on a three-, four-, or five-point scale, or should another strategy be adopted altogether? One drawback to increasing the number of levels in a rubric is that it may become more likely for multiple assessors to use the rubric to assign different ratings. Another task that comes with increased levels is the writing of descriptions that accurately communicate the kind of work that merits each level. Could it be effective to structure rubrics using only two levels? In this work, we will summarize some of the literature on the development of rubrics, and then describe our process of creating a "two-column" rubric – one that only describes excellent and minimally acceptable performances. We will share examples of how we apply these two-column rubrics in our junior- and senior-level chemical engineering laboratory courses. We explain our algorithm for using the two-column rubrics, including how faculty, teaching assistants, and students are trained to apply the algorithm. Finally, we conducted inter-rater reliability analysis for an example assignment and found modest improvement in agreement between assessors compared to previous evaluation methods. We conclude with our next steps in our development and revision of these rubrics.

Background

The University of Delaware is a medium-sized, mid-Atlantic, public institution whose chemical engineering program graduates on average 80 undergraduates per year. The curriculum includes two semesters of chemical engineering laboratory, though the second semester can be replaced with a research project. Over the past three years, the average enrollment in the first-semester laboratory has been 82 students across about 23 teams, while enrollment in the second semester has averaged about 56 students over 14 teams. Both laboratory courses require an oral presentation at the end of the term. In the first semester, this presentation is delivered live during the final week of class, while in the second semester, the presentation is delivered as a video so that it may be viewed asynchronously.

Over the past few years, the instructors of the laboratory sequence have worked to develop a set of common rubrics. This way, regardless of the technical details of each experiment, a standard set of expectations in terms of organization, format, and presentation is maintained through the semester. The faculty in the courses are each responsible for applying the rubrics for their own projects and reports. Only the oral (or video) presentation rubric is currently used by multiple people to evaluate the same group submission.

Rubrics are used in general to clarify expectations for students, and to help identify specifically where students can improve in their work. There are two core elements of a rubric: criteria and standards [1]. The criteria are the features or characteristics that are evaluated, and the standards are identifiable levels of quality. Stevens and Levi [2] provide considerable detail in rubric construction. Most of their examples result in rubrics with four to six criteria, usually scored across three standard levels. They recommend building these kinds of rubrics from the outside in – that is, for each criterion, describe the highest standard level, then the lowest standard level, and then fill in the middle level(s). They note that this kind of rubric becomes more difficult to generate with the more levels one desires. Stevens and Levi also present what they call a "scoring guide rubric," which focuses more on the criteria and presents only the description of the highest standard level. Exploration of the use of rubrics in chemical engineering has been presented previously. Newell et al. [3] suggest applying four standards levels, rather than three or five, to avoid there being a middle or "neutral" option.

Methods

Development and Implementation of Rubric

The original version of the oral presentation rubric for our laboratory course is shown in Appendix A. This is an example of a scoring guide rubric. There is narrative of expectations of an excellent presentation, but there is no clear rationale for what separates "excellent" from "very good," for example. This presents a clear drawback when it comes to inter-rater reliability, as each evaluator has their own opinion for the different standards.

The original video presentation rubric is shown in Appendix B. Arguably this is not an effective rubric. It could generously be categorized as a scoring guide rubric as well.

One proposal was to move toward more of a check-box style rubric, as described in Stevens and Levi [2]. However, the team quickly found fault with this approach, as it made the rubric appear to take longer to use – instead of evaluating across five criteria as in Appendix A, we would be evaluating in effect some twenty criteria on a yes/no basis. There was also no consensus for how to weight these twenty different criteria, though all agreed that the weighting should not be even across them all.

In consultation with Delaware's Center for Teaching and Assessment of Learning (CTAL), we learned of a middle ground option – describing the highest and lowest standard levels for each of our original five criteria, and simply leaving the middle of the rubric blank. This way, we were clearly describing "A" (excellent, 100% of the points) and "C-" (minimally acceptable, 70% of the points) work, but allowing for evaluating roughly as "B+" (90%) or "B-" (80%) work as well. The algorithm is as follows:

(1) Determine whether the assignment criterion matches the description of the highest standard. If so, mark the "100%" column. If not, continue on.
(2) Determine with the assignment criterion matches the description of the lowest standard. If so, mark the "70%" column. If it seems to meet more than just this minimum standard, continue on.

(3) Determine whether the assignment criterion is closer to the description of the highest or the lowest standard, marking "90%" if the work is closer to the highest standard, and "80%" if the work is closer to the lowest standard.

More examples of this style of rubric can be found at CTAL's website [4].

There is some issue if the assignment is deemed to be rated lower than minimally acceptable. Our team decided to call this an "attempt," worth 50% of the available points, though we have discussed hypothetical situations in which no points at all should be earned. Because we work with upper-level undergraduates, we have not yet encountered a situation that requires anything other than the 100/90/80/70/50% ratings. The currently-used version of the video presentation rubric can be seen in Appendix C.

Measuring Inter-Rater Reliability

There are several ways to measure the effectiveness of a rating system and indeed multiple definitions of inter-rater reliability. Saal et al. [5] summarize roughly a dozen ways to evaluate ratings based on a survey of the literature. In purporting the effectiveness of their rubrics, Newell et al. [3] report the percentage of rubric ratings where all faculty are in agreement or within one standard level of one another. The computation of Cohen's kappa (to compute agreement between two raters) or Fleiss's kappa (to compute agreement among several raters) can be implemented using Minitab [6]. A kappa value of 0 corresponds to an agreement rate as probable as by chance; a kappa of 1 corresponds to perfect agreement among raters. In general a kappa value of 0.75 or more is desired to indicate "good" agreement.

Results and Discussion

The rubric in Appendix B was used in fall 2016 to evaluate video presentations. Three faculty rated each of the eight videos across the four criteria. Of these 32 ratings, all three raters agreed perfectly 7 times (22%) and rated within one standard level of one another 18 times (56%). The overall Fleiss's kappa value is 0.11 and is statistically significantly different than zero ($\alpha<0.01$), but this only really means agreement among the raters is barely better than by chance, far from "good" agreement.

The rubric in Appendix C was used in fall 2018 to evaluate fourteen video presentations, again across four criteria. Of these 56 ratings, all three raters agreed perfectly 8 times (14%) and rated within one standard level 42 times (75%). The overall Fleiss's kappa value is 0.20 and also statistically significantly different than zero ($\alpha<10^{-4}$). Again, this does not correspond to "good" agreement. However, by both measures (percent near agreement and kappa value), the new rubric and algorithm appears to be used more consistently than the old rubric.

One main challenge to continue working on with individual faculty members is the consistent and correct use of this style of rubric. In computing Fleiss's kappa, we allowed for misuse of a rubric standard level to count as a disagreement among raters. If all faculty used the rubric properly, it is likely that the kappa value would be somewhat greater. Even in the old rubrics, when asked to rate students on an integer scale from 1 to 5, some individuals would assign decimal scores, and in one rater's case, down to the nearest hundredth of a point (i.e., a 4.85 out of 5) – which is somewhat ironic in the cases where comments were made on the use of significant digits in student work. Also, as the faculty team teaching our laboratory courses changes on a yearly basis, it takes time to get new members "on board" with this style of evaluation. This said, it is critical to get raters to understand and agree on the algorithm in order to obtain more consistent rubric results.

We have already begun developing and implementing rubrics using this "two-column" approach in other courses and plan to investigate other features of this grading scheme, such as number of student questions about grades before and after implementation. So far, anecdotally, faculty involved in the laboratory courses before and after the development of the new rubric and algorithm report that it takes less time to evaluate student work, and they are able to provide more qualitative feedback than before. Further time may be saved by implementing these rubrics electronically, such as via Canvas SpeedGrader [7]. Using electronic rubrics in a learning management system also makes it more difficult to assign a rating different than the 100/90/80/70/50% allowed by the rubric, and so far, in Canvas our faculty and graduate students have more consistently applied the algorithm described here rather than assigning other scores.

Conclusions

We describe here one approach to rubric design that is consistent with published literature but that takes advantage of the features of a traditional number of standard levels while only requiring two columns to be "filled out" with description. A specific algorithm is applied, using this two-level structure, while still effectively having five distinct levels for assessment. Comparing the implementation of this style of rubric to previously used assessment techniques in our chemical engineering laboratory course, we find that our inter-rater reliability has improved, but there is still room to improve in both the development and communication of these rubrics to get faculty to grade more consistently. Preliminarily, it appears that faculty in our laboratory courses spend less time overall grading while being able to provide more qualitative feedback to students. Because of the perceived time savings, we intend to explore the use of this type of rubric in other courses and assignments.

References

[1] B. E. Walvoord and W. J. Anderson, Effective Grading, San Francisco, CA: Jossey-Bass, 2010.

[2] D. D. Stevens and A. J. Levi, Introduction to Rubrics, Sterling, VA: Stylus, 2013.

[3] J. A. Newell, K. D. Dahm and H. L. Newell, "Rubric Development and Inter-Rater Reliability Issues in Assessing Learning Outcomes," in *American Society for Engineering Education Annual Conference & Exposition*, Montreal, Canada, 2002.

[4] University of Delaware Center for Teaching and Assessment of Learning, "Rubrics," [Online]. Available: https://ctal.udel.edu/resources-2/rubrics/. [Accessed 3 February 2019].

[5] F. E. Saal, R. G. Downey and L. M., "Rating the Ratings: Assessing the Psychometric Quality of Rating Data," *Psychological Bulletin,* vol. 88, no. 2, pp. 413-428, 1980.

[6] Minitab Inc., "Kappa statistics for Attribute Agreement Analysis," 2017. [Online]. Available: https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/how-to/attribute-agreement-analysis/attribute-agreement-analysis/interpret-the-results/all-statistics-and-graphs/kappa-statistics/. [Accessed 3 February 2019].

[7] Instructure, "What is SpeedGrader?," 2019. [Online]. Available: https://community.canvaslms.com/docs/doc-10712. [Accessed 3 February 2019].

Appendix A

Chemical Engineering Laboratory I
Final Oral Presentation Rubric

This rubric will be used by all instructors, TAs, and peers in attendance of your final presentation. Please circle one of the five assessment levels for each rubric item and use the blank space to briefly explain your rationale for assessment against the listed criteria.

Reviewer Name:	
Date:	
Presenting Group Number and Name:	

Overall Mechanics/Presentation	**20 Points**			
Excellent (20)	Very Good (18)	Good (16)	Acceptable (14)	Needs Improvement
Balances appropriate amounts of text and graphics to convey message effectively Selects appropriate font and font sizes, colors, and backgrounds Uses well-sized, professional quality, and appropriately simple equations, graphics, and tables to support content Appropriately times presentation to last 15-20 minutes Exhibits command of body language (eye contact, clarity and speed of speech, composure when speaking and listening) Requires all group members to speak for one segment of presentation for roughly equal amounts of time each Shows evidence of practice / preparation without being overly rehearsed or memorized				
Audience and Motivation	**10 Points**			
Excellent (10)	Very Good (9)	Good (8)	Acceptable (7)	Needs Improvement
Organizes and prioritizes presentation with obvious thought as to the audience of the presentation Clearly motivates interest and importance of proposed project				
Presentation Content	**50 Points**			
Excellent (50)	Very Good (45)	Good (40)	Acceptable (35)	Needs Improvement
Clearly defines key goals of project upfront Provides relevant background, theory, and experimental procedures needed to understand project Identifies appropriate safety procedures and precautions of project Explains statistical and error analysis, including an example of sensitivity analysis Clearly defines key conclusions and recommendations Organizes presentation in an intelligible sequence				
Questions	**10 Points**			
Excellent (10)	Very Good (9)	Good (8)	Acceptable (7)	Needs Improvement
Responds professionally and thoughtfully to audience questions Includes involvement from all members of the group				
Initiative	**10 Points**			
Excellent (10)	Very Good (9)	Good (8)	Acceptable (7)	Needs Improvement
Shows evidence of critical thinking of elements of the project Goes “above and beyond” expectations outlined elsewhere in this rubric, in terms of quality of content and/or style				
Total Score	**100 Points**			

Appendix B

Chemical Engineering Laboratory II
Video Presentation Assignment

Group Number ______ **Total points: /25**

Group Names:

Quality of Video (Easy to follow, clear speaking)	Excellent 5	Very Good 4	Good 3	Acceptable 2	Poor 1
Technical Content (Experiment fully explained, analysis of results)	Excellent 10	Very Good 8	Good 6	Acceptable 4	Poor 2
Creativity (Use of visual effects)	Excellent 5	Very Good 4	Good 3	Acceptable 2	Poor 1
Organization/Questions (Video addresses most important issues and answers to any questions)	Excellent 5	Very Good 4	Good 3	Acceptable 2	Poor 1

Additional Comments:

Appendix C

Chemical Engineering Laboratory II
Video Presentation Assignment

Reviewer Name: ____

Group # being Reviewed: ____

The assessment for this project will be according to the rubric below.

Category	**10 pts**	**9 pts**	**8 pts**	**7 pt**	**<7 pts**
Video Quality	Video is easy to follow; visuals and audio are both clear; video fits within time constraints.			Video runs without any technical glitches.	
	Comments:				
Technical content **(Score in this section is doubled)**	Experimental procedure and analysis of results are fully and correctly explained within the constraints of the video.	Not strong enough to merit a "10" rating, but closer to a "10" than a "7".	Not deficient enough to merit a "7" rating, but closer to a "7" than a "10".	Only viewers with significant familiarity with the experiment can follow the experimental procedure and results.	Not even the "7" criteria are met.
	Comments:				
Audience suitability and creativity	Selection of content and use of visual effects are presented at a level appropriate for the target audience of peers, TAs, and instructors in cleverly communicating information.	Not strong enough to merit a "10" rating, but closer to a "10" than a "7".	Not deficient enough to merit a "7" rating, but closer to a "7" than a "10".	Video includes any content that seems arbitrary or without consideration of entire target audience.	Not even the "7" criteria are met.
	Comments:				
Organization	Video addresses most important issues specific to the experiment in a logical order and addresses questions that are likely to come up. All members contribute to the oral presentation.	Not strong enough to merit a "10" rating, but closer to a "10" than a "7".	Not deficient enough to merit a "7" rating, but closer to a "7" than a "10".	Video does not have a clear structure or sequence and does not include vocal contributions from all members	Not even the "7" criteria are met.
	Comments:				

Total ____________/ 50