# Algorithm for Consistent Grading in an Introduction to Engineering Course

**Prof. Joshua A Enszer, University of Delaware**

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.

**Prof. Jenni M. Buckley, University of Delaware**

Dr. Buckley is an Associate Professor of Mechanical Engineering at University of Delaware. She received her Bachelor's of Engineering (2001) in Mechanical Engineering from the University of Delaware, and her MS (2004) and PhD (2006) in Mechanical Engineering from the University of California, Berkeley, where she worked on computational and experimental methods in spinal biomechanics. Since 2006, her research efforts have focused on the development and mechanical evaluation of medical and rehabilitation devices, particularly orthopaedic, neurosurgical, and pediatric devices. She teaches courses in design, biomechanics, and mechanics at University of Delaware and is heavily involved in K12 engineering education efforts at the local, state, and national levels.

# Algorithm for Consistent Grading in an Introduction to Engineering Course


## Abstract

This Complete Evidence-based Practice paper will describe the design and implementation of rubrics in a 700-student introduction to engineering course.

Timely and meaningful feedback is important to student learning but challenging to deliver in large enrollment classes. The use of rubrics is virtually mandatory to ensure clear communication of expectations and consistency in evaluation. We have implemented a rubric algorithm to address the time-based challenges of both rubric design and implementation.

Rubrics are used to clarify expectations for student work in advance, and also to evaluate submitted student work. The two main elements of a rubric are the criteria and the standards. The criteria (usually the "rows") of a rubric are the characteristics of work that are evaluated, while the standards (usually the "columns") establish levels of quality. The mechanics of rubric construction are explored in detail by Stevens and Levi. Most of their example rubrics have four to six criteria assessed against three standard levels. They suggest constructing these rubrics starting with the "outside" columns and working inward – for each criterion, first establish the highest standard level, then the lowest standard level, and then fill in the middle level(s). This style of rubric can become more cumbersome to construct as the number of standards increases. It has been suggested to design rubrics with an even number of standards to avoid a "middle" option during evaluation.

We have developed the rubrics for our Engineering 101 course by focusing only on two columns within the rubric, describing only the highest quality level (which earns full credit, an A grade) and the minimum acceptable quality level (which earns credit roughly equivalent to a C or C- grade). The other columns in the rubric are effectively left blank, but with a deliberate algorithm that allows the rubric to expand from having two columns to having six – two columns are between A and C-, which represent being closer to the A description than the C- or being closer to the C- description than the A, and two columns are on the other side of the C-, which represents an attempt that is below the minimum standard or no attempt at all. Rubric use follows the same general algorithm: the student work is first compared against the highest quality level, then if necessary the lower level, and finally if necessary the work is determined to be closer to one of these levels or the other.

The final element of this project involves the training of our teaching assistants to obtain consistent evaluation of student work across all students in the class. This consists of a calibration exercise before the start of the semester, and regular spot-checking by lead teaching assistants during the semester.

We describe here our rubric development and implementation process with examples directly from our introductory engineering course (roughly 700 student enrollment in two sections with 15 teaching assistants per section) at the University of Delaware. Through use of a retrospective analysis, we present quantitative evidence that the use of rubrics per our methodology results in

higher grading consistency. In future work we plan to include a comparison of inter-rater reliability for course assignment evaluation.

## Introduction

Rubrics are a tool to communicate expectations of student work. In addition to their use to evaluate student work once submitted, rubrics can be used in advance of student submissions to clarify the desired qualities of their work. Stevens and Levi [1] posit that there are four basic parts to a rubric: a description of the task or assignment, a scale (levels of the achievement, possibly points or grades), a list of dimensions of the task (a more detailed breakdown of requirements and/or skills demonstrated via the task), and a set of descriptions of each level of performance (each combination of possible scale level and task dimension). The language used to describe rubrics is not entirely consistent; elsewhere the task dimensions are called criteria, and the scale refers to standards [2] [3]. Regardless of nomenclature, literature on rubrics consistently specifies either a checklist or a grid in which to communicate the levels of accomplishment that an artifact exhibits when assessed for specific desired qualities. In general, published examples of rubrics include all possible combination of criteria and standards filled in, as shown in Table 1.

**Table 1**. Basic grid format for standard rubrics. Each cell in the table corresponds to a different combination of criterion and standard level. In this example, three criteria are evaluated against three possible levels of standards.

| | Standard level 1 | Standard level 2 | Standard level 3 |
|---|---|---|---|
| Criterion 1 | Description of meeting criterion 1 to the 1st standard level | Description of meeting criterion 1 to the 2nd standard level | Description of meeting criterion 1 to the 3rd standard level |
| Criterion 2 | Description of meeting criterion 2 to the 1st standard level | Description of meeting criterion 2 to the 2nd standard level | Description of meeting criterion 2 to the 3rd standard level |
| Criterion 3 | Description of meeting criterion 3 to the 1st standard level | Description of meeting criterion 3 to the 2nd standard level | Description of meeting criterion 3 to the 3rd standard level |

The criteria of a rubric should match the learning goals associated with the assignment, while the standard levels usually correspond to an evaluation and are listed in sequence (for example, excellent, very good, good… down to the lowest standard level, perhaps corresponding to 100%, 90%, 80%, … the lowest possible percentage possible of the available grade). In constructing this type of rubric, it is recommended to work from the "outside in"; that is, first write the descriptions for how each criterion is met at the highest standard level, then write the descriptions for how each criterion is met at the lowest standard level, then work on descriptions in between. By this method, rubric construction increases in difficulty particularly with the number of standard levels. It is suggested that middle levels should be some combination of the outer levels, though this advice can be difficult to follow from the same references that suggest that sometimes the lowest standard level is simply a negation of the highest standard level. Many example rubrics of this style result in around 4-6 criteria assessed over 3 standard levels [1]. It has been suggested that rubrics should have an even number of standard levels to avoid an exact "middle" option in evaluation [4]. This is consistent with general principles of survey design [5].

A second kind of rubric, called a checklist or a scoring guide, focuses only on the highest standard level associated with each criterion, as shown in Table 2.

**Table 2**. Basic format for a grading checklist or a scoring guide. As opposed to a traditional rubric, only the descriptions of the highest standard level for each criterion is given.

| | Description | Comments |
|---|---|---|
| Criterion 1 | Description of meeting criterion 1 to highest standard | |
| Criterion 2 | Description of meeting criterion 2 to highest standard | |
| Criterion 3 | Description of meeting criterion 3 to highest standard | |

A scoring guide is generally easier to construct but harder to use for evaluation, because the user is effectively delaying how to parse standard levels. While this style of rubric may be better suited to some assignments, especially for more open-ended projects where it can be difficult for a traditional rubric to capture all possible performance evaluations, it is harder for an evaluator to remain objective and/or consistent in the use of this style [1].

Independent of the type of rubric design, keeping rubrics to a length of 4-6 criterion total helps to keep assessment of student work at a more holistic level, which among other things helps to avoid formulaic evaluation of formatting and grammar in favor of assessing the overall effectiveness of student work [6].

**Methods (Rubric Design)**

To address drawbacks inherent in the standard grid and grading checklist styles of rubrics, we developed a rubric design and implementation approach in which only the highest and lowest-acceptable standard levels are described for each criterion. The structure of this rubric results in the format shown in Table 3.

**Table 3**. Basic format for a rubric designed using a "two-column" approach. The remaining cells in the grid are left blank but used according to the algorithm we describe below.

| | Highest Standard Level | Tends toward Highest Standard Level | Tends toward Minimum Standard Level | Minimum Acceptable Standard Level | Attempt | No Evidence |
|---|---|---|---|---|---|---|
| Criterion 1 | Description of meeting criterion 1 to the highest standard | | | Description of meeting criterion 1 to the minimum standard | | |
| Criterion 2 | Description of meeting criterion 2 to the highest standard | | | Description of meeting criterion 2 to the minimum standard | | |
| Criterion 3 | Description of meeting criterion 3 to the highest standard | | | Description of meeting criterion 3 to the minimum standard | | |

By focusing only on what it takes to meet the minimum standard for a given criterion (i.e., the minimum possible performance that would result in assessing a grade of C-) and what it takes to meet the highest standard (the performance that would be assessed a grade of A), we avoid potential pitfalls in describing work in between these standard levels. We introduce two levels between the highest and minimum acceptable levels, to avoid the phenomenon of "fence-sitting" – tending toward the middle possible evaluation option because an artifact does not exactly meet either extreme. We also add columns below the minimum standard to reflect a form of "partial credit" for work, but also an effective "zero" standard for when work completely fails to address a criterion.

Once constructed, the algorithm for assessing student work is as follows, for each criterion:

(1) Decide if the work matches the description of the highest standard. If so, mark this level; If not, move to Step 2.
(2) Decide if the work matches the description of the minimum standard. If so, mark this level; if not, move to Step 3.
(3) If the work is between the two descriptions, decide if it is closer to the highest or the minimum standard and mark the appropriate level; otherwise move to Step 4.
(4) If the work appears to attempt to meet this criterion but hasn't met the minimum standard, mark the attempt level; otherwise mark no attempt.

In our Engineering 101 course at the University of Delaware, we choose to assign percentage scores of roughly 100/90/80/70/50/0 for each criterion. In most cases, a score is computed according to a predetermined weighting of the criteria associated with a given assignment.

Our Engineering 101 course has on average roughly 700 students per year, taught by a team of two faculty members and a set of about 30 undergraduate teaching assistants. The teaching

assistants are trained to use the course rubrics in a session prior to the start of the semester. They are given a sample assignment and rubric to grade independently. Then the faculty facilitate a short discussion of the grading process, one criterion at a time, as a form of calibration. No further training is given, though lead teaching assistants work to "spot-check" grading occasionally through the semester. Between the two years evaluated in this study, we switched from the traditional three-column approach to rubrics (Table 1) to the "two-column" approach (Table 3) and algorithm described above.

**Methods (Data Analysis)**

We conducted a retrospective analysis to determine whether the two-column rubric and algorithm improved grading consistency when used by teaching assistants in conjunction with traditional rubrics. With IRB approval, relevant data were assembled for two non-consecutive semesters (Fall 2017 and Fall 2019) of the aforementioned large enrollment first semester introduction to engineering course. The antecedent semester predated used of the grading algorithm (Pre-Algorithm), while the algorithm was implemented in the most recent iteration of the course (Post-Algorithm). Both versions of the course were co-taught by the same instructors and utilized identical rubrics and assignment instructions for four of eleven weekly summative assignments (Assignments A, B, C, and D). All assignments were subjective in nature and involved activities ranging from designing and administering stakeholder surveys to conducting validation experiments with an early-stage prototype. The Pre- and Post-Algorithm rubrics are shown in Appendix A.

For both Pre and Post-Algorithm versions of the course, undergraduate teaching assistants were randomly assigned as graders for all students in the course. Each student had a unique teaching assistant who graded their work for the entire duration of the course, and every teaching assistant was assigned to a reasonably sized subset of students (around 25-35). In both versions of the course, the teaching assistants attended a half-day orientation at the start of the semester that included a grade normalization exercise to familiarize themselves with the style of grading rubric. The algorithm was introduced to teaching assistants in the Post-Algorithm group at that time. No further training was administered to the teaching assistants throughout the semester.

Given that this study was conducted retrospectively and that there was no overlap in student grading across teaching assistants, typical intra- and inter-rater reliability measures could not be obtained from our historical data. In lieu of these measures, quality and process control statistical methodologies were employed to compare grading consistency for Pre vs. Post-Algorithm. For each of the four common assignments (Assignments A, B, C, and D), Heterogeneity of Variance Tests were performed to detect graders with intra-rater variances that deviated substantially from other graders (REML Method, alpha = 0.05; JMP Pro 14.0). The number of "Grader Outliers" for each assignment was determined and then represented as a percentage of the total number of teaching assistant graders for that course year. The same statistical methodology was used to calculate "Grader Effect," which was defined as the percentage of the overall variance in assignment grades that could be attributed to inconsistencies across graders.

## Results and Discussion

The complete retrospective data set included 577 students and 23 teaching assistants in the Pre-Algorithm year and 698 students and 32 teaching assistants in the Post-Algorithm year. Results are presented in Table 4. Student assignment scores for both years were similarly distributed and were relatively high scoring and tightly clustered (>90% median score with 10-20% IQR; see Table 4).

**Table 4**: Results of Heterogeneity of Variance Tests to compare consistency of grading Pre- and Post-Algorithm. Grader Outliers represents the percentage of all graders in a given course year who demonstrated variance in assignment grading that substantially deviated from other graders. Grader Effect is the total sample variance that can be attributed to inconsistencies across graders.

| | **Student Scores: Median (IQR)** | | **Grader Outliers** | | **Grader Effect** | |
|---|---|---|---|---|---|---|
| **Assign-ment** | **Pre-Algorithm** | **Post-Algorithm** | **Pre-Algorithm** | **Post-Algorithm** | **Pre-Algorithm** | **Post-Algorithm** |
| A | 94% (13%) | 94% (9%) | 26.1% | 3.1% | 30.2% | 17.4% |
| B | 95% (12%) | 95% (10%) | 13.0% | 9.4% | 14.2% | 17.5% |
| C | 91% (16%) | 97% (10%) | 4.3% | 6.3% | 12.1% | 11.9% |
| D | 90% (18%) | 96% (14%) | 17.4% | 9.4% | 11.3% | 18.3% |

The percentage of Grade Outliers was substantially lower in three of the four assignments (Assignments A, B, and D in Table 4) for Post-Algorithm compared to Pre-Algorithm years of the course, and it was roughly equivalent for the remaining assignment (C). On average, Grader Effect accounts for about 15% of the total variance in student grades for both years of the course; however, for one assignment (Assignment A), use of the algorithm substantially reduced grader-attributable variance from 30.2% to 17.4%. Grader Effect is similar for two of the remaining assignments (B and C) and slightly higher Post-Algorithm for one assignment (D).

While this particular study is unable to probe interrater reliability, smaller studies for upper-level undergraduate engineering courses have shown modest gains in this measure when comparing the two-column style to the traditional style of rubric [7]. Because of sample sizes and the sophistication of this analysis, more work is needed to measure interrater reliability via future work.

## Conclusions and Future Work

Rubrics are a common tool to communicate expectations and evaluate details related to student work. Traditional examples of rubrics are either a scoring guide that lists descriptions of work that meets a given standard or a more complex grid that describes different evaluation levels for every criterion associated with student work. We describe here a method that draws from the

strengths of both of these approaches in terms of time needed to design, but that does not increase the time needed to use for evaluation compared to either approach.

The results of our retrospective analysis suggest that the new grading algorithm improves consistency across teaching assistant graders. Applying identical rubrics and assignment instructions, we found that use of the grading algorithm minimizes the frequency of “outlier” graders who are innately more inconsistent than the norm. There is also some evidence that the algorithm reduces fraction of total variation in student grades that can be attributed to differences in grading practices across graders. These findings are especially encouraging given the setting of the study. Specifically, that it is (a) a large enrollment course, (b) involving weekly open-ended writing assignments, (c) that are graded by a non-trivial number of undergraduate teaching assistants, (d) who are trained on the grading rubric only once at the beginning of the semester. Any one of these course characteristics inherently works against grading consistency; thus, the results of this study are promising for further scaling of the algorithm approach. Future work by our team will involve a prospective and well-controlled intra- and inter-rater reliability analysis to definitively establish the internal validity of this grading approach.

**References**


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

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

[3] E. F. Barkley and C. H. Major, Learning Assessment Techniques, San Francisco: Jossey-Bass, 2016.

[4] 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.

[5] M. E. Henerson, L. L. Morris and C. T. Fitz-Gibbon, How to Measure Attitudes, Newbury Park, CA: SAGE Publications, 1987.

[6] M. C. Paretti, L. D. McNair and J. A. Leydens, "Engineering Communication," in *Cambridge Handbook of Engineering Education Research*, New York, Cambridge University Press, 2014, pp. 601-632.

[7] J. A. Enszer, "Developing Reliable Lab Rubrics Using Only Two Columns," in *American Society for Engineering Education Annual Conference & Exposition*, Tampa, FL, 2019.

## Appendix

Below are the rubrics associated with "Assignment A" in our retrospective analysis. Table A.1 shows a traditional rubric (Pre-Algorithm, implemented in 2017) and Table A.2 shows the modified two-column rubric (Post-Algorithm, 2019). The traditional rubric is preceded by the following text: This assignment is worth 50 points total, graded according to rubric below. "Good" receives 90-100% total points for each element; "Average" is 80-89% of points for an element; and "Poor" is 70-79%. Completely missing elements automatically receive 30% of available points. Peer evaluations will be used to calculate individual grades for all group elements.

**Table A.1**. The rubric used to evaluate student work when adopting a set of three standard levels for each criterion.

| **Element** | **Total Points** | **Good** | **Average** | **Poor** |
|---|---|---|---|---|
| 1. Survey Design | 10 | Meets all survey design requirements: (1) logical flow; (2) minimal bias; (3) 5-15 questions; (4) one open-ended question; (5) two different question types | Meets 4/5 design requirements | Meets 3 or fewer design requirements |
| 2. Survey Distribution | 5 | Meets all survey design requirements: (1) logical flow; (2) minimal bias; (3) 5-15 questions; (4) one open-ended question; (5) two different question types | Meets 4/5 design requirements | Meets 3 or fewer design requirements |
| 3. Data Analysis | 10 | - Appropriate choice of graph type for data representation<br>- Use of graphs only when necessary<br>- All graph components are legible | - Inappropriate choices of graphs for some data<br>- Some unnecessary graphs<br>- All graph components are legible | - Inappropriate choices of graphs for most data<br>- Far too many/few graphs<br>- Poor formatting across the board |
| 4. Reporting | 20 (5 pts/ section) | As high quality as template for each section:<br><br>Intro – Study motivation<br>Methods – Describes survey instrument & distribution<br>Results – Highlights key findings<br>Conclusions – States "story" told by data | Missing components in 1-2 sections. Fails to consistently reference figures & tables. | Missing components in 3-4sections. Fails to reference figures & tables. |
| 5. Updated Team Norms | 5 | - 5 or more team norms<br>- All norms relatively important<br>- Norms cover both team philosophy and logistics | - 5 or more team norms<br>- Not all norms are important<br>- Norms cover either team philosophy and logistics | - <5 team norms<br>- Not all norms are important<br>- Norms cover neither eam philosophy nor logistics |

**Table A.2.** The two columns needed to evaluate student work using the method described in the paper. The other columns of the rubric would be left blank, and therefore they are omitted here.

| **Element** | **Total Points** | **Excellent (100% of available points)** | **Minimally Acceptable (70% of available points)** |
|---|---|---|---|
| 1. Survey Design | 10 | Meets all survey design requirements: (1) logical flow; (2) minimal bias; (3) 5-15 questions; (4) one open-ended question; (5) two different question types | Meets at least 3 design requirements |
| 2. Survey Distribution | 5 | Distributed by the deadline | Distributed in time to gather data, but not by the deadline |
| 3. Data Analysis | 10 | - Appropriate choice of graph type for data representation<br>- Use of graphs only when necessary<br>- All graph components are legible | - Inappropriate choices of graphs for most data<br>- Far too many/few graphs<br>- Poor formatting across the board |
| 4. Reporting | 20 (5 pts/section) | - Intro – States the goal of the survey. Summarizes prior UCR. Describes why report is valuable.<br>- Methods – Describes who survey was sent to, how distributed (Qualtrics), and refers to table with survey questions<br>- Results – Number of respondents, Summary statics (percentages, averages) for major questions. Presents 1-2 nicely formatted figures that summarize key points.<br>- Conclusions – States "story" told by data & how data will be used | - Intro – Very cursory description of purpose. Mixes in Methods, Results, or Conclusions.<br>- Methods – No information about survey population. Only partial information about survey questions.<br>- Results – Presents figures without text summary.<br>- Conclusions – Missing or a rehash of the Results. |
| 5. Updated Team Norms | 5 | - 5 or more team norms<br>- All norms relatively important<br>- Norms cover both team philosophy and logistics | - <5 team norms<br>- Not all norms are important<br>- Norms missing either team philosophy nor logistics |

We note that, as each rubric is near its first implementation, there are still places to improve in wording to make things clearer and more accurate – for example, it should not be the "minimum acceptable" to have "three or fewer" qualities in a list of five; otherwise technically zero is fewer than three, and no work would merit a rating near the 70% mark.