Grader consistency in using standards-based rubrics

Nathan M. Hicks, Purdue University, West Lafayette (College of Engineering)

Nathan M. Hicks is a Ph.D. student in Engineering Education at Purdue University. He received his B.S. and M.S. degrees in Materials Science and Engineering at the University of Florida and taught high school math and science for three years.

Prof. Heidi A. Diefes-Dux, Purdue University, West Lafayette (College of Engineering)

Heidi A. Diefes-Dux is a Professor in the School of Engineering Education at Purdue University. She received her B.S. and M.S. in Food Science from Cornell University and her Ph.D. in Food Process Engineering from the Department of Agricultural and Biological Engineering at Purdue University. She is a member of Purdue's Teaching Academy. Since 1999, she has been a faculty member within the First-Year Engineering Program, teaching and guiding the design of one of the required first-year engineering courses that engages students in open-ended problem solving and design. Her research focuses on the development, implementation, and assessment of modeling and design activities with authentic engineering contexts.

Grader Consistency Using Standards-Based Rubrics

Introduction

Differences in instructors' grading practices can have a considerable effect on student success[1]. The resulting variation in student success becomes a considerable issue for large-scale courses that require multiple sections and multiple instructional teams to accommodate enrollment demands. At many large universities, these types of courses are common at the introductory level and provide students with the foundational knowledge and skills necessary to succeed in a given field of study. Thus, it is a crucial, albeit challenging, goal to provide a consistent experiences across all sections through, among other things, a high quality grading system that ensures equitable probabilities of success.

Grading quality involves reliability and fairness. These factors can be improved through the standardization of grading with specific learning outcomes. In addition to fairness and reliability, the use of a standards-based approach also improves the meaningfulness of grades[2,3] and informs instructors of their students' mastery of the content. Information derived from grading could be absolutely reliable if all data points were obtained through purely objective, multiple-choice assessments; however, the lack of free-response would drastically limit the range of knowledge and skills that could be authentically assessed[4]. When free-response or open-ended problems are assessed, rubrics can be used to minimize grading variations due to subjective judgment[5]. However, despite the use of a common standards-based rubric, inconsistencies often still persist for individual graders or across teams of graders.

Inconsistencies in grading stem from several factors related to the problem being graded, the individual grader, the grading team (for example, the set of graders for a give section of a course), the time of day, the grader's level of fatigue[4], and the grader's overall experience—novice graders tend to be less consistent[6]. While some variability may be inherently unavoidable[7], the literature suggests that overall consistency may be improved by increasing rubric clarity[8,9] and training graders to use rubrics[10].

At our own university, the members involved with developing and delivering the second course in a two-semester introductory engineering sequence chose four years ago to utilize a standards-based grading approach, in part, to improve fairness and minimize grading variability. After working through a number of logistical issues with the learning management system and initial development of standards-based grading rubrics[11–13], it was time to turn our attention to how effectively our graders were employing those rubrics. As such, in this study, we ask the following research question: How accurately and consistently are graders applying standards-based grading rubrics across multiple sections of a required first-year engineering course at a large university?

Background

Dimensions of grade quality. There are five factors related to the quality of grades and grading systems: fairness, validity, fidelity, integrity, and reliability. For grading to be fair, students should know when and how they are being judged[14]. Further, grades should be based on quality of individual student work (independent of previous achievements or the performance of other students) and comparable across sections, courses, programs, and institutions over time (through consistent levels of grading toughness)[14]. While comparability across programs and institutions may require unrealistic levels of coordination, the former criterion can certainly be attained at the local level through the use of well-crafted standards (e.g., learning objectives) and rubrics.

Grade validity, grade fidelity, and grade integrity are closely related constructs. Grade validity measures the extent to which an instrument measures what it claims to measure, which is often established through expert agreement[15]. Grade fidelity is the converse of grade contamination—high fidelity should represent purely academic achievement and remove non-achievement factors, like attendance[14,16]. Finally, grade integrity is a grade's trustworthiness, or the extent to which a grade represents what it is supposed to represent[14]. Thus, these three constructs represent a hierarchy, where validity is a prerequisite to fidelity, which is a prerequisite for integrity. An individual item on an instrument, such as a rubric, has validity if it measures the dimension it was intended to measure. If an individual grade constructed by that instrument truly represents achievement of the learning outcome and is free from contamination of non-achievement factors, the grade has fidelity. In other words, while all items might validly measure aspects of performance, unless every item collectively represents achievement (by, for instance, excluding items related to timeliness or legibility), the instrument may not be fidelitous. Lastly, an overall grade has integrity if the collection of measures that constitute that grade are valid, fidelitous, and combine in a logical way to represent achievement of the course content.

Grade integrity also demands reliability, which is not only a function of the instruments used, but also the people using the instruments. All grading relies on graders making judgments by integrating information, perception, memory, and training[10]. Even with experienced teachers, grading decisions have been shown to vary considerably, particularly when using provided benchmarks and judging mid-quality work[10]. The measure representing the tendency for multiple graders to assign the same or similar scores is called inter-rater reliability. There are multiple ways to measure inter-rater reliability, with some being more appropriate depending on whether the expectation is for different graders to agree exactly on levels or only need consistent scales[17]. Regardless, reporting multiple estimates is often preferred[18]. While the ideal goal is for consensus interpretation of rubrics for all graders, consistency of scale may be sufficient for achieving fairness, as such a systematic difference can be adjusted[19].

Grading and using rubrics. In some instances, with highly complex learning tasks, traces of subjectivity are unavoidable[7]. A variety of factors contribute to judgments made by graders, including assumed cognitive models of grading, accepted practices or interpretations of peer graders, the tendency to adhere more or less rigidly to standards, and experience with grading and the accompanying presence of mental models or prototypes representing different qualities of work[6]. Graders are also affected by their general values and beliefs about grading, such as values of non-achievement factors, like effort, and perceptions that grades function as rewards or punishments[20]. Again, however, rubrics offer a means to minimize subjectivity, particularly when dealing with complex assignments or tasks[21]. Further, reliable use of rubrics is enhanced with trained graders using analytic rubrics with exemplar cases[5].

Analytic rubrics (as opposed to holistic rubrics, which are not relevant to this study) are two-dimensional matrices consisting of a list of criteria versus gradations of quality[21,22]. To increase likelihood of consistent use, the criteria should be fair, free from bias, strongly aligned with tasks, expressed in terms of observable behaviors, and written at an appropriate level for the students[9]. The levels should be clearly delineated with logical point allocations[9].

Despite their potential for reducing subjectivity, poorly written rubrics are prone to misuse. There are several problems commonly associated with rubric construction: the absence of criteria critical to the assessed task or construct; overly general or overly detailed criteria; the lack of consistency or parallelism between criteria or levels or across different rubrics; sources of ambiguity, such as orphan and widow words or phrases and inconsistent use of qualifiers; missing or redundant descriptors; insufficient, excessive, or unevenly incremented performance levels; and cases where adhering to the rubric produces cognitive dissonance[8,23]. Thus, a delicate balance of specificity is necessary to achieve an effective sets or criteria and levels.

Means to enhance the use of rubrics. The goal of having high quality grades for a standards-based course requires production of valid, high-fidelity rubrics that can be applied reliably to produce fair, high integrity grades. Researchers have proposed a few methods to improve rubric validity, fidelity, and reliability, including rubric assessment and revision, grader training, and score adjustment.

Rubrics should be clearly, comprehensively, and unambiguously worded without contamination from non-achievement components[8,9,14,16]. This can be achieved through construction of new rubrics or revision of old rubrics. Further, anchoring achievement levels with examples of student work helps graders to make decisions and allows developers to simulate and test rubric robustness prior to implementation[24].

The final approach to enhancing the reliability of rubrics is to adjust scores. This could be achieved by showing graders how the grades they assign align with their peer graders (in terms of average and distribution), which tends to influence more extreme graders to become more moderate[25]. Alternatively, calibration rounds can be used to establish complex formulas to adjust for different tendencies[4].

Methods

Context and data collection. This study investigated grading in the second of a two-semester, first-year engineering course sequence that is required for all engineering undergraduates at a large Midwestern university. The course employs standards-based grading using a set of 19 major learning objectives, each with a set of minor learning outcomes, collectively accounting for 88 total learning outcomes.

The course was offered during the spring semester of 2016 and consisted of 15 sections containing, collectively, 1699 students. Each section had its own set of five undergraduate teaching assistants and one graduate teaching assistant for up to 120 students. The undergraduate teaching assistants graded all homework assignments using standards-based rubrics designed for specific learning outcome instances and the graduate teaching assistants supervised the process (while one may question the ethics of undergraduate grading, the improvement in timeliness of feedback to students makes this practice justifiable[15]). Each rubric item had levels of “fully achieved,” “partially achieved,” “underachieve,” and “no evidence of achievement” (although, in some cases, one or more level was disallowed). Throughout the semester, the course’s online learning management system retained all work submitted by every student (for each homework this submitted work included an answer sheet, one or more MATLAB files and/or one or more Excel files), as well as all information regarding the marks assigned and feedback given for each learning objective assessed (between 5 and 10), and the grader responsible. This data was collected and de-identified in May of 2016.

Our investigation consisted of two stages. The purpose of the first stage was to explore the course data to understand the extent of grading inconsistency across assignments and sections and to identify specific learning objectives on specific assignments (which will be referred to as “learning objective instances” or just “instances”) in specific sections that would be the best candidates for more detailed analysis. The purpose of the second stage was to inspect the selected learning objective instances, rubrics, and actual pieces of student work and their assigned scores to measure the reliability of the assigned scores and to understand possible sources of inconsistency.

Exploratory stage. The initial set of data included a total of 126 graded learning objective instances, leading to well over 200,000 individually graded items. Each grader mark of “fully achieved,” “partially achieved,” “underachieved,” and “no evidence of achievement,” was assigned a numerical score of 4, 3, 1.5, or 0 points, respectively, to fit to a traditional Grade Point Average (GPA) scale. These numerical scores were then used to calculate the overall GPA for each instance for each section. Then, for each learning objective instance, we calculated the average and standard deviations of these GPAs across all sections.

Without specific indicators specifying which learning objective instances contained more error than others, we had to identify some way to characterize the scores assigned for each instance, and within each section for each instance. Given the large average section size of over 100 students, we assumed that the average ability level and distribution of ability levels of students in each section would be relatively similar. We did recognize, however, that the quality of explanation for each topic may have varied from section to section, leading to some variability in scores across sections and instances. Taking all of this into consideration, we looked at the distribution of section GPAs with respect to the overall average for each learning objective instance and the distribution of standard deviations of section GPAs with respect to the overall standard deviation for each instance.

This analysis process did not yield any specific learning objective instances that absolutely demanded further attention over others; that is, there was no obvious analytical metric that clearly identified specific instances or sections. However, we established three criteria that we hoped would help us to find examples of variation in rubric interpretation: (1) selected instances should have moderate overall GPAs compared to other instances (i.e., avoiding instances that received mostly high or mostly low grades); (2) selected instances should have some sections with comparatively high and comparatively low section GPAs; and (3) selected instances should have a considerable range of section standard deviations. For the first learning objective instance identified and analyzed, we randomly sampled a few pieces of student work from one section to pilot the process. For that instance and the following two, we then selected one individual section that had a low section GPA, one with a mid-range section GPA, and one with a high section GPAs, making sure that each section had at least a few students receiving each score level.

In-depth analysis stage. First, the primary author analyzed the problems and rubrics qualitatively to identify potential points of confusion for either the students completing the assignment or the graders using the rubrics. Next, we obtained a stratified random sample of student work by randomly selecting approximately five students (if that many existed) who received each level of achievement (i.e., five "fully achieved," five "partially achieved," and so on) for each section, resulting in approximately 60 pieces of student work per learning objective instance and a total of 172 pieces of student work across all three learning objective instances analyzed (the approach for the first problem analyzed was less structured and only had 52 pieces of work across four sections). We then collected a smaller stratified random sample of 10 pieces of student work (two at each level of achievement) for each problem for scoring calibration with both authors. For additional context, it must be noted that this paper's primary author is a graduate research assistant, who has several years of teaching experience at both the high school and undergraduate level, and the second author is a professor and curator for this course who has been responsible for writing many of the learning objectives and rubric items.

For each sub-sample, scoring was nearly unanimous; however, for the few instances of disagreement, we discussed our interpretations until we reached consensus. The primary author then assigned scores to the remaining sample following each calibration round. Once the primary author scored each piece of work, we compared those scores with the scores the students actually received in the course to calculate inter-rater reliability, as well as the distribution of instances in which the student grader assigned higher or lower scores.

During this comparison process, we noted that there were multiple instances of assigned disallowed scores and instances in which students were assigned scores other than "fully achieved" but were not given feedback. We felt these issues warranted further investigation, so we looked across all problems for which there were disallowed score levels to determine the number of times disallowed scores were assigned for each problem. Also, for the three problems and corresponding sections that were analyzed more deeply, we determined the total number of times students should have received feedback—that is, any time they did not receive a grade of "fully achieved"—but did not.

Results

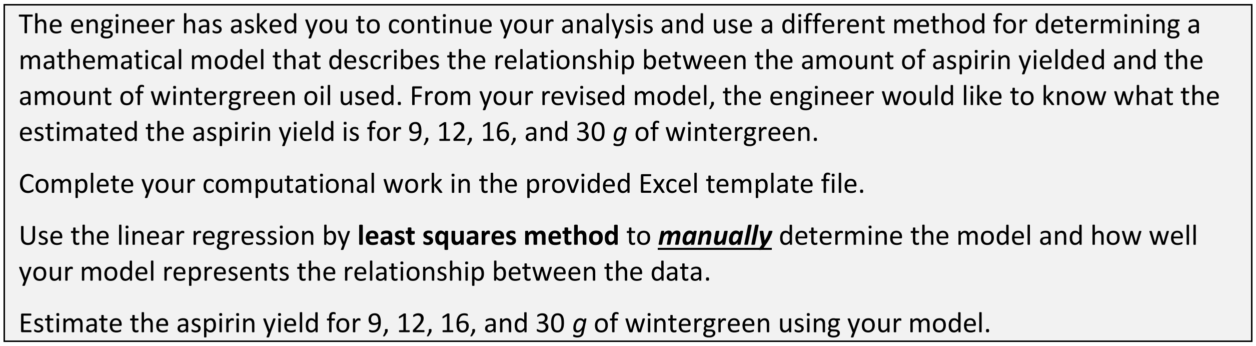
Exploratory analysis. The data resulting from the initial exploratory analysis are a bit overwhelming in quantity and a bit underwhelming in terms of conclusiveness—as stated previously, no single learning objective instances absolutely stood out as being best for analysis, and many other problems easily could have been selected. As such, only a small portion of the data is worth showing for illustrative purposes (see Table 1). The three columns highlighted in Table 1 represent the three learning objective instances we selected based on the guidelines of our selection criteria. The three individual cells with darker highlighting in each column represent the specific sections selected for student work analysis. Again, while no individual learning objective instances absolutely stood out as demanding investigation over others, these three were selected because they fit the criteria described in the Methods section. Further, individual sections were selected based on the criteria previously outlined such that one section had a low section GPA, one had a moderate section GPA, and one had a high section GPA and the selected section needed to have at least a few students scored at each level of proficiency (for instance, for Learning Objective 12.19, some sections had no students assigned some levels of proficiency).

In-depth analysis. While we conducted an in-depth analysis of three different learning objective instances, only one will be shown in detail due to space limitations and similarity in discussion of the results. This section will provide context for and illustrative examples of student work for HW 4, LO 12.19 (homework 4, learning objective 12.19), which related to performing linear regression on a set of points and using the regression equation to make predictions. We will present the overall results of the reliability analysis for all three instances analyzed following illustrative examples.

Table 1. Abbreviated data from exploratory analysis used to identify problems and sections for further analysis. Highlighting illustrates selected learning objective instances and sections.

	Average Section GPAs									Section Standard Deviations								
Learning Objective	1.0	…	12.19	…	13.5	…	14.13	…	19.2	1.0	…	12.19	…	13.5	…	14.13	…	19.2
Assessment	HW 1	…	HW 4	…	HW 5	…	HW 5	…	HW 8	HW 1	…	HW 4	…	HW 5	…	HW 5	…	HW 8
Section A	3.93	…	3.38	…	2.93	…	2.86	…	3.30	0.25	…	1.10	…	1.43	…	1.34	…	1.48
Section B	3.71	…	3.89	…	2.67	…	2.99	…	3.13	0.75	…	0.65	…	1.54	…	1.49	…	1.65
Section C	3.81	…	1.54	…	2.56	…	2.28	…	2.83	0.61	…	0.73	…	1.44	…	1.42	…	1.80
Section D	3.75	…	1.86	…	2.36	…	2.87	…	3.11	0.47	…	0.91	…	1.59	…	1.18	…	1.65
Section E	3.71	…	1.91	…	2.57	…	2.57	…	2.66	0.81	…	0.99	…	1.46	…	1.29	…	1.82
Section F	3.74	…	2.60	…	2.31	…	2.44	…	2.48	0.75	…	1.41	…	1.74	…	1.66	…	1.81
Section G	3.93	…	2.86	…	3.01	…	3.52	…	2.36	0.25	…	1.49	…	1.42	…	0.90	…	1.53
Section H	3.39	…	3.30	…	2.94	…	3.25	…	1.77	1.10	…	1.18	…	1.33	…	1.23	…	1.53
Section I	3.76	…	3.24	…	2.40	…	2.52	…	2.90	0.50	…	1.48	…	1.53	…	1.36	…	1.77
Section J	3.53	…	2.34	…	2.39	…	2.61	…	2.94	0.85	…	0.98	…	1.48	…	1.26	…	1.74
Section K	3.72	…	2.52	…	2.52	…	2.81	…	2.24	0.67	…	1.32	…	1.63	…	1.16	…	1.87
Section L	3.95	…	2.43	…	3.01	…	2.70	…	3.05	0.22	…	1.28	…	1.40	…	1.25	…	1.66
Section M	3.74	…	3.84	…	3.65	…	3.49	…	2.80	0.51	…	0.72	…	0.94	…	1.08	…	1.79
Section N	3.95	…	3.28	…	3.03	…	3.23	…	3.12	0.22	…	1.13	…	1.33	…	1.21	…	1.65
Section O	3.59	…	3.05	…	1.81	…	2.89	…	2.88	0.78	…	1.29	…	1.59	…	1.55	…	1.80
All sections:	3.74	…	2.82	…	2.68	…	2.87	…	2.81	0.66	…	1.33	…	1.52	…	1.35	…	1.75

First, we begin by describing the original question as the students saw it in their assignment. This question followed an initial question asking the students to use a two point method (out of a set of 10 points) to develop a model for the data. Figure 1 shows the specific language for the follow-up question.



The engineer has asked you to continue your analysis and use a different method for determining a mathematical model that describes the relationship between the amount of aspirin yielded and the amount of wintergreen oil used. From your revised model, the engineer would like to know what the estimated the aspirin yield is for 9, 12, 16, and 30 *g* of wintergreen.

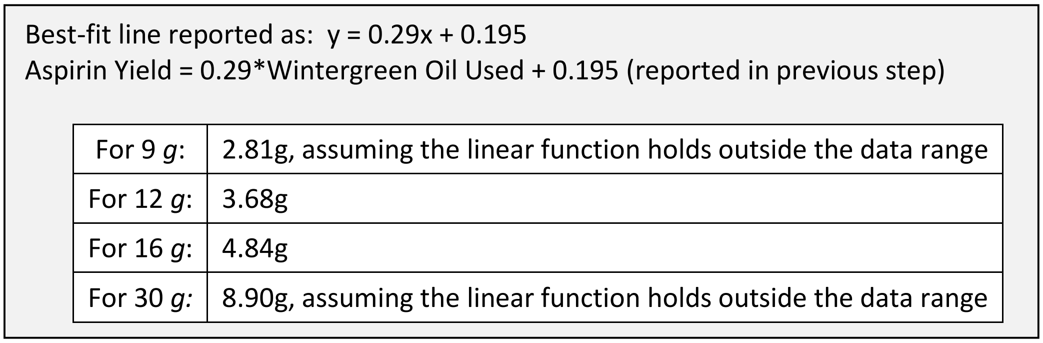
Complete your computational work in the provided Excel template file.

Use the linear regression by **least squares method** to ***manually*** determine the model and how well your model represents the relationship between the data.

Estimate the aspirin yield for 9, 12, 16, and 30 *g* of wintergreen using your model.

Figure 1. The question related to homework 4, learning objective 12.19, as presented on the original assignment.

All graders had access to the following: the original problem set, a solution key for the Excel spreadsheet, a solution key for the Word document answer sheet (see Figure 2), and a grading rubric (see Figure 3). While marking each student, the grader had access to the student's answer sheet and their Excel file.



Best-fit line reported as: y = 0.29x + 0.195
Aspirin Yield = 0.29*Wintergreen Oil Used + 0.195 (reported in previous step)

For 9 *g*:	2.81g, assuming the linear function holds outside the data range
For 12 *g*:	3.68g
For 16 *g*:	4.84g
For 30 *g:*	8.90g, assuming the linear function holds outside the data range

Figure 2. The answer sheet solution key provided to the graders. The equation was determined in a step prior to the prediction calculations.

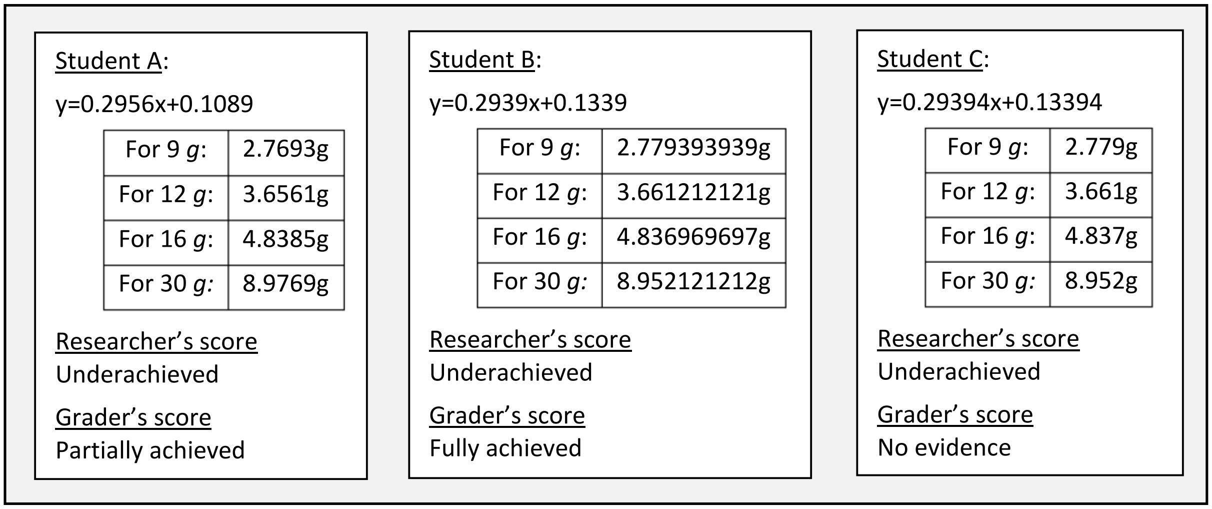
Criteria	**Achievement Levels**			
	No evidence: 0%	**Under-achieved: 50%**	**Partially achieved: 80%**	**Fully achieved: 100%**
Appropriately use the best-fit linear model to make predictions	Anything less than requirements for underachieved	Any 1 elements missing or incorrect from the list for fully achieved	NA	□ 2.4.e Predicts aspirin yield for wintergreen values within range of original data set □ 2.4.e. Acknowledges wintergreen values outside range of data cannot be used for predicting aspirin yield

Figure 3. The scoring rubric provided to the graders for this learning outcome.

It should be noted that the answers provided in the answer sheet solution key actually contained an error. In reality, the best-fit equation for the data given to the students should have been $\text{Aspirin yield} = 0.294 * \text{Wintergreen Oil Used} + 0.134$. As such, the correct values that would have been calculated were 2.78g, 3.66g, 4.84g, and 8.95g, respectively (with the first and last still falling outside the data range). This rubric error immediately shows two possible problems:

(1) if the graders did not pay close enough attention, they would likely have marked nearly all of the student responses as being incorrect; and (2) if the graders did not evaluate student work based on the equation they determined in the previous step, they would likely penalize the students for mistakes that were not directly associated with the learning objective (that is, this learning objective is not associated with determining the best-fit line, but with using that line to make predictions). Finally, another problem is highlighted by the fact that this error in the key was undetected until the first author was analyzing this data after the semester had concluded. Thus, while it may be understandable that an undergraduate assistant might be hesitant to present an error to faculty, none of the 75 graders using this answer key reported a suspected error to the appropriate faculty.

A number of additional issues are best understood by looking at specific examples of student work. Three examples of student work are shown in Figure 4, along with the equations for the lines of best-fit that each student had determined in previous steps and the marks assigned by both the first author and by the official undergraduate grader.



Student A:

y=0.2956x+0.1089

For 9 *g*:	2.7693g
For 12 *g*:	3.6561g
For 16 *g*:	4.8385g
For 30 *g:*	8.9769g

Researcher's score

Underachieved

Grader's score

Partially achieved

Student B:

y=0.2939x+0.1339

For 9 *g*:	2.779393939g
For 12 *g*:	3.661212121g
For 16 *g*:	4.836969697g
For 30 *g:*	8.952121212g

Researcher's score

Underachieved

Grader's score

Fully achieved

Student C:

y=0.29394x+0.13394

For 9 *g*:	2.779g
For 12 *g*:	3.661g
For 16 *g*:	4.837g
For 30 *g:*	8.952g

Researcher's score

Underachieved

Grader's score

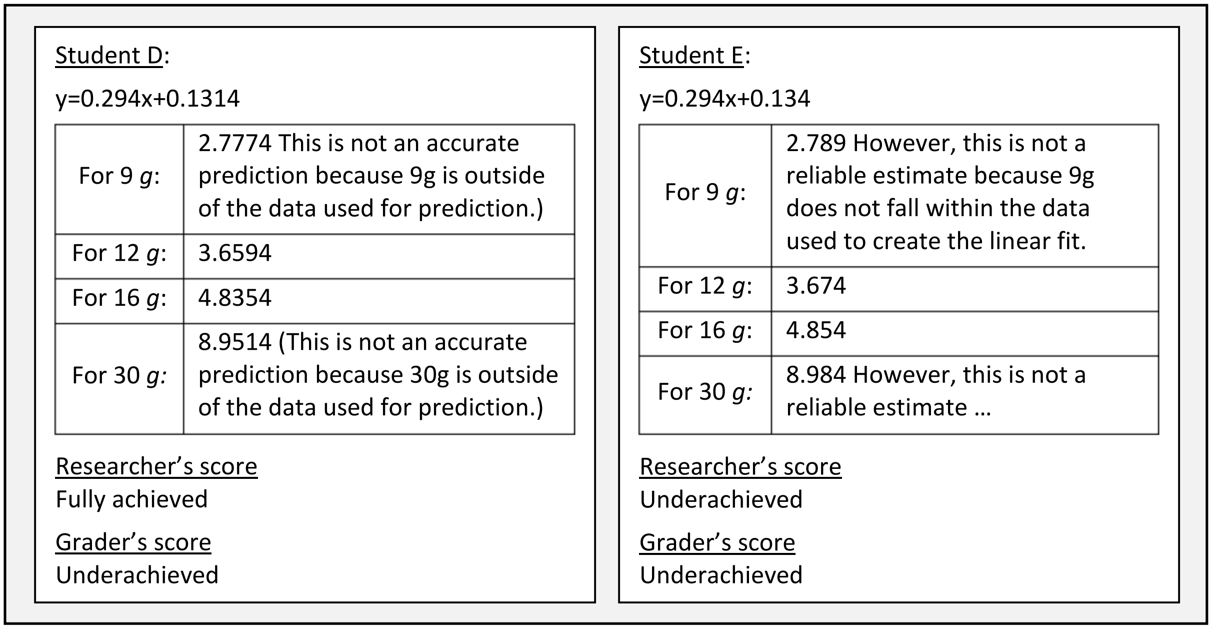
No evidence

Figure 4. Three examples of student work that illustrate grader misuse of rubrics and issues with reporting numbers.

In each of Figure 4's three examples, the student failed to acknowledge that the 9g and 30g predictions fell outside the range of the original data. According to the rubric, this should automatically drop all three students to no higher than "underachieved," as one of the two pieces of evidence for achievement is not demonstrated. However, we can see that for students A and B, the graders assigned marks of "partially achieved" and "fully achieved." This immediately indicates that the grader is either not using the rubric, does not understand the rubric, or does not care to follow the rubric and reasonably indicates insufficient oversight (given that this work clearly should earn "underachieved" and that "partially achieved" is a disallowed level in the rubric and given that this occurred in many other cases).. The fact that students B and C were actually marked by the same grader, and that, in both cases, the predicted numbers approximately correspond with the reported equations, indicate that this particular grader was inconsistent with his or her grading decisions.

The examples in Figure 4 further illustrate another important challenge of grading this question related to the assignment itself: significant figures and rounding. The number of significant figures for the equations and the predicted values vary for each student. For student B, in particular, the predicted values possess several more figures than the equation that was supposedly used to calculate them. Differences in rounding practices can lead to differences in final answers and reporting a rounded equations and unrounded predictions could leave a grader stumped. Of course, the grader could open up the accompanying submitted Excel files and try to determine whether the student's work makes sense, but juggling several files for every student demonstrates the separate problem of inconveniencing the grader. The issue with rounding could be reduced, if not eliminated, by specifying proper rounding practices in the text of the problem (or at least this could be done for the first few instances in the course where this might be an issue) and by suggesting tolerance ranges for the graders (e.g., answers must be $X \pm 1\%$ or between X_1 and X_2).

The two examples of student work shown in Figure 5 demonstrate how different graders might reasonably impose differing levels of strictness based on their interpretations of the student work. In both cases, the students failed to report units. While the authors were not concerned with this detail, one could argue that without units, these predictions are completely meaningless and the students have not achieved one of the two pieces of evidence for achievement. Meanwhile, student D states that the predictions for 9g and 30g are "not accurate," which is not definitively true—the prediction could be accurate, but the data to support such a claim is insufficient. For student E, the numbers reported are slightly off from what the equation would actually produce. Looking at the student's Excel file (again, an issue with juggling documents) shows that the student actually calculated the correct values, but reported different numbers. While it is a mystery as to how this happened, the discrepancy between the work and the answer given present the grader with a dilemma, one for which different graders might reasonably make different grading decisions.



Student D:

y=0.294x+0.1314

For 9 *g*:	2.7774 This is not an accurate prediction because 9g is outside of the data used for prediction.)
For 12 *g*:	3.6594
For 16 *g*:	4.8354
For 30 *g:*	8.9514 (This is not an accurate prediction because 30g is outside of the data used for prediction.)

Researcher's score

Fully achieved

Grader's score

Underachieved

Student E:

y=0.294x+0.134

For 9 *g*:	2.789 However, this is not a reliable estimate because 9g does not fall within the data used to create the linear fit.
For 12 *g*:	3.674
For 16 *g*:	4.854
For 30 *g:*	8.984 However, this is not a reliable estimate …

Researcher's score

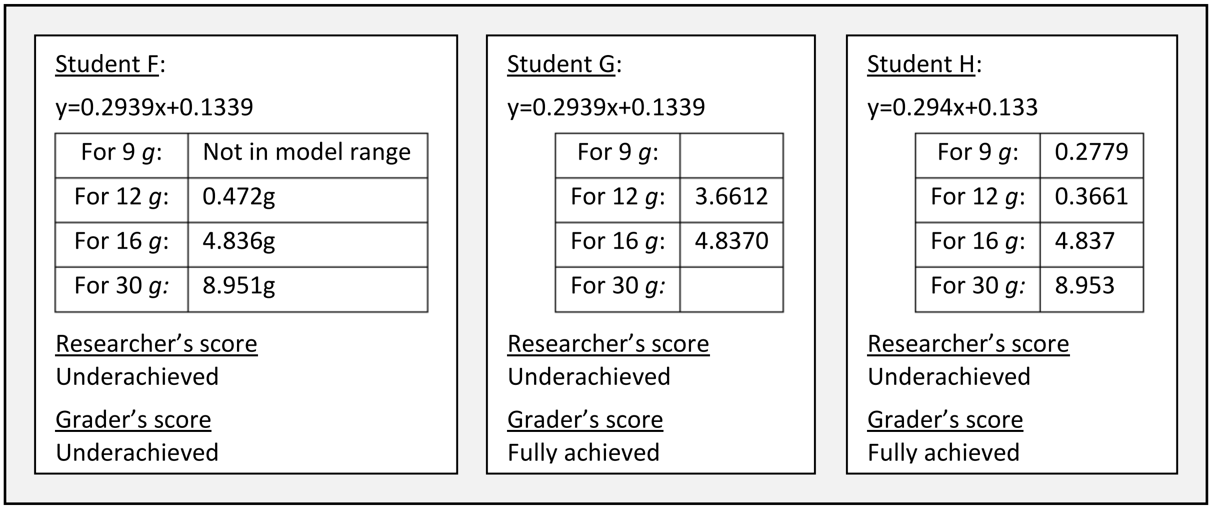
Underachieved

Grader's score

Underachieved

Figure 5. Examples that demonstrate reasonable differences in interpretation of student work.

Figure 6 presents three more examples that might lead to feelings of confusion for a grader. For student F, for instance, they correctly acknowledges that 9g falls outside the data range, but fail to do so for 30g. Similarly, they correctly applied their equation to make predictions for 16g and 30g, but produce a wildly wrong value for 12g. As the grader is supposed to assign one single mark for each of the four sub-problems, do these inconsistencies demonstrate partial evidence for each piece of evidence or does failing to achieve that evidence in every instance indicate failure to achieve that evidence? One could reasonably argue either way. Student G presents a related, but unique dilemma: does not providing an answer for the predictions outside the data range constitute "acknowledging" the limitation? It seems that the grader made that interpretation, but we felt uneasy with making assumptions about student knowledge without explicit demonstrations. On the other hand, student H's first two predictions are off by a factor of 10. It is unclear how this happened or how the student did not notice the error. One could give the student the benefit of the doubt and mark this as "underachieved," but one could just as easily argue failure to recognize this error deserves "no evidence of achievement."



Student F:

y=0.2939x+0.1339

For 9 *g*:	Not in model range
For 12 *g*:	0.472g
For 16 *g*:	4.836g
For 30 *g:*	8.951g

Researcher's score

Underachieved

Grader's score

Underachieved

Student G:

y=0.2939x+0.1339

For 9 *g*:	
For 12 *g*:	3.6612
For 16 *g*:	4.8370
For 30 *g:*	

Researcher's score

Underachieved

Grader's score

Fully achieved

Student H:

y=0.294x+0.133

For 9 *g*:	0.2779
For 12 *g*:	0.3661
For 16 *g*:	4.837
For 30 *g:*	8.953

Researcher's score

Underachieved

Grader's score

Fully achieved

Figure 6. Three pieces of student work illustrating possible points of confusion for graders.

Overall reliability measures. To measure the inter-rater reliability between ourselves and the official course graders, we explored two types of reliability estimates: consensus estimates and consistency estimates[18]. While each individual problem contributes very little to a student's overall grade, systematic differences in the assignment of marks could easily contribute to issues in fairness but also affect the feedback regarding learning and progress that are sent to both the students and the instructors. Therefore, this is definitely a situation in which we would like to attain consensus agreement.

We will look at two measures of consensus agreement: percent agreement and Cohen's kappa. Across the three problems and 172 pieces of student work, we agreed with the official graders' marks on 85, corresponding to a 49.4% agreement (see Table 2 for further details)—considerably lower than the recommended 90% level of agreement for sufficient reliability[26]. Further, there was not a percent agreement with any individual section for any of the learning objective instances studied that exceeded 65%. Additionally, nearly three-quarters of all disagreements (64 out of 87) were instances in which the mark we assigned was lower than the mark the grader assigned and more than half of those instances (35 out of 64) were at least two levels of achievement apart.

Table 2. Contingency table illustrating agreement between marks assigned by the researchers and the official graders.

		Grader marks				
		Fully achieved	**Partially achieved**	**Underachieved**	**No evidence**	
Researcher marks	**Fully achieved**	19	5	3	0	27
	Partially achieved	8	10	4	3	25
	Underachieved	17	11	29	8	65
	No evidence	6	12	10	27	55
		50	38	46	38	172

To account for the possibility of chance agreement, we also calculated the overall Cohen's kappa across all items analyzed to be 0.326, which can be interpreted as "minimal" agreement. Further, we calculated Cohen's kappa for each individual section for each of the learning objective instances analyzed (Table 3). It can be seen that, of the three learning objective instances and the corresponding 10 sections investigated, only two sections produced kappa values indicated greater than "minimal" agreement.

Table 3. Frequency table summarizing Cohen's kappa values across sections. Adapted from McHugh[27].

kappa	**Frequency**	**Interpretation**
Below .20	3	None
.21–.39	5	Minimal
.40–.59	2	Weak
.60–.79	0	Moderate
.80–.90	0	Strong
Above .90	0	Almost perfect

Based on the tendency for the graders to give higher marks, a reasonable question to explore is the presence of consistency reliability. While consensus may not be being achieved, if consistency is high, the grading issue may be easier to resolve. Krippendorff[26] argues that measures such as percent agreement, Cohen's kappa, or Cronbach's alpha fail to meet all of the requirements for a good reliability indicator, opting instead for the use of Krippendorff's alpha. Therefore, we also calculated the Krippendorff's alpha for ordinal data across all 172 items analyzed to be 0.422, which, while stronger than Cohen's kappa, does not meet Krippendorff's suggested level of 0.70 necessary to claim significant reliability[26]. It is possible that, if the graders had each been treated as a unique individual rather than clumped into one symbolic "grader" in performing the calculation (a decision made based on the small number of specific items graded by any given grader in our analysis), we might have been able to identify consistency for some individual graders. As it stands, however, across all metrics, we cannot feel confident that the course graders were applying the same interpretations of the rubrics that we intended, nor that their interpretations were internally consistent for any individual graders.

Assignment of disallowed marks and inclusion of feedback. One of the three learning objective instances we used for the in-depth analysis (the one shown in detail in the previous section) did not allow for one of the achievement levels in the rubric. However, when making comparisons with the actual marks assigned to student work, we noticed eight instances (out of 54) of the disallowed "partially achieved" being assigned, as our learning management system does not allow for rubric levels to be blocked from use. Based on this, we extended this investigation to all of the problems across the entire course that had disallowed achievement levels (43 of the 126 problems) to determine how frequently graders were assigning disallowed marks. This analysis revealed that every single problem with disallowed achievement levels had at least one instance of the disallowed mark being assigned and three problems had over 100 instances.

When trying to consider individual grader reliabilities, we also noticed that there were many instances where graders were assigning marks other than "fully achieved" but were not providing any feedback to the learner. We found this to be troubling; whenever a student receives a reduced mark, it should be clear why this decision was made. Consequently, we investigated this using all graded student submissions from the three problems and corresponding sections that we had previously selected and identifying every instance in which the assigned mark was below "fully achieved" and the grader did not provide feedback. Collectively, 1103 pieces of student work were graded, of which 553 received marks below "fully achieved." Of the work warranting feedback, only 65, or approximately 12%, were missing feedback. While this number is fairly low, it would ideally be zero. Further, this does not say anything regarding the quality of the feedback that was provided in the other 488 instances. It should also be noted that the tendency to fail to give expected feedback varies by section, with some sections in failing to provide feedback far more frequently than others.

Discussion

Through our analyses of standards-based grading using rubrics in a large-scale, multi-section first-year engineering course, we encountered two major challenges to grading fairness that were present in our examples and data: fidelity issues due to the design of the assignments and rubrics; and reliability issues due to insufficient training and supervision. Based on these challenges, we recommend that assignments and rubrics need to be designed using user-centered design principles, taking into account the way the students would interpret the assignments, and the way the graders would interpret the rubrics in conjunction with the work submitted by students. Further, we recommend that proper rubric usage needs to be illustrated through comprehensive training and reinforced through adequate supervision.

Assignments. First, the assignments and solutions need to be well-designed. While most of our assignments are error free, our example illustrated an error in the solution that easily could have contributed to students receiving lower marks. Of course, we all make mistakes—an error is not the end of the world—but our study shows that it is important to have checks in place to prevent, detect, and correct errors when possible, and there needs to be a culture amongst the community of graders that encourages communication of errors if and when they are identified.

Beyond this, we saw that by considering the way students will think about the assigned questions in advance can help to remove variability in student answers that contribute to challenges for graders. For instance, for the examples shown, inclusion of lines in the assignment that specify rounding or inclusion of preset units in answer sheets could yield greater consistency in reported answers. In one of the other learning objective instances analyzed, the students were required to complete an eight cell truth table and a 16 cell truth table while showing all work. Not only did we realize that this would be an overwhelming number of steps for the grader to realistically read through (which, we suspect, likely led the graders to only look at final answers), we also recognized that it was likely unclear to the student what would constitute "all work" (e.g., does every logical operation require its own line of work?). By redesigning this item so that some of the cells are pre-completed for the students, we expect that it will not only reduce burden on the graders, but also more effectively communicate expectations to the students. Since this study, we have been going through each assignment considering how students who are inexperienced with the content might misinterpret directions, so that we can more effectively guide the students to produce higher quality work.

On a similar note, we discovered that the potential need for graders to have to juggle several documents simultaneously (the solution key, the rubric, the student's answer sheet, the student's Excel file(s), and the students MATLAB file(s)) imposes a large cognitive demand on the graders, which would understandably make them more prone to grading mistakes. The more the students' submissions can be consolidated, the easier the process will become for the graders. For instance, we learned of a MATLAB function that combines all codes, comments, and outputs into a single .pdf file. By having students put what would have previously gone into a Word document into the comments of the MATLAB code, the graders can look at a single document of student work rather than flipping between documents. These measures to improve consistency of answers and reduce sources of work will likely lessen cognitive load on graders, which should improve efficiency and minimize grading errors.

Rubrics. The extensive work that has contributed to the development of rubrics for this course prior to this analysis should not be undermined. On the surface, they generally communicate grading criteria clearly, particularly to knowledgeable and experienced graders. However, the inexperience of undergraduate graders and the unpredictable nature of freshman-level work requires the rubrics to be slightly more robust. The examples hinted at issues with rounding, for instance. While an experienced grader with certain autonomy could comfortably make decisions regarding rounded answers, an undergraduate grader might reasonably ask and not know, "how far off is too far off?" Inclusion of tolerances or ranges of acceptable answers in the rubric itself can remedy this issue for questions that involve rounding. Similarly, and in reference to the truth table question mentioned previously, when the rubric specifies that "all work must be shown," it would be helpful to the grader if the rubric clearly communicated what constitutes sufficient versus insufficient work to help distinguish between achievement levels. As such, we have added these considerations into our rubrics when appropriate.

The examples from the results section also illustrated the challenge of lumping multiple parts of an answer into one collection of evidence items to achieve the outcome. That is, when a student gets three out of four parts correct, one could argue that missing one indicates that the student has not fully demonstrated achievement of the outcome. On the other hand, one could just as reasonably argue that the three correct instances do indicate understanding and achievement, depending on whether or not the fourth part represents a unique aspect of the learning objective. In cases where we feel the need to group multiple items into single rubric items, we have modified our rubrics to specify how to handle these situations by including statements such as "At least three items are fully correct OR all four items have at most one error," for instance.

Our analysis also brought to our attention the unintended consequences of the disallowed achievement levels for various rubric items. While this may be as much an issue with grader behavior as with the rubric itself, when over 100 invalid marks are assigned on a specific learning objective instance, something is clearly being communicated by the graders. Based on our analysis and the literature we identified in the background section, we have determined that it only really makes sense to disallow achievement levels, such as removing "underachieved" and "partially achieved," if there is a truly dichotomous learning outcome. If there is any potential middle-ground between two achievement levels, the graders have demonstrated that they will want to select something in between—and, often, their judgment may very well be right. In many cases, these rubric criteria with disallowed achievement levels are being re-written to allow for intermediate levels.

Finally, we determined some rubric items may not always fully align with the questions to which they are linked. For instance, for the question in which students had to fill in the truth tables, the actual rubric learning objective was to "Construct truth tables to evaluate logical expressions," which places the emphasis on the construction of the truth table rather than the evaluation of the logical expressions. However, the question and the rubric description of the evidence necessary to show achievement were solely focused on the evaluation of the logical expressions. In these situations, we have learned that either the learning objectives need to be refined, or different learning objectives should be assessed.

Training and supervision. While there were many pieces of student work that highlighted unpredicted, but understandable, ambiguities or needs for clarification in our assignments and rubrics, there were also many cases that should not have been ambiguous for which the graders assigned clearly inappropriate marks (for instance, the examples where graders assigned "fully achieved" when the students did not acknowledge the predictions were out of range of the data). While we cannot know what influenced their decision-making skills without conducting a more targeted investigation, these cases suggested that some form of structured training could improve consensus of understanding and consistency of application.

The importance of training in achieving inter-rater reliability in tasks such as grading is emphasized throughout the literature[10,18,19,24]. Training with rubrics has been shown to help inexperienced graders to establish mental models of student work at each level and to develop an iterative approach to grading[10,24]. Further, training is one of the best means to produce less variable, more accurate assessment of student work and functions well as an intervention in response to poor reliability checks[10,19,28]. However, while training often improves inter-rater reliability, it may also prevent graders from employing a full range of scores[29].

In our context, we hope that training might help graders to develop practices to improve their own consistency, efficiency, and effectiveness. We are encouraging graders to improve both their consistency and their efficiency by keeping a log of unique student answers and the mark they assigned. This ensures that if the graders see a similar answer several papers later, they can look back and assign a similar grade again without having to think as deeply or without being affected by variability in mood over time. Further, based on the issues with some graders providing insufficient feedback, keeping a log of the feedback that was given in response to different sorts of answers can help the graders give feedback more effectively and efficiently.

The examples in the Results section also showed that students often produce unimaginably unique answers for problems, which make it challenging to fully anticipate and prepare graders for what they might see. Still, graders also need to see at least some examples of student work to calibrate their grading decisions prior to actually assigning grades. By mining through previous student work, we have begun to provide training examples to expose graders to various situations in advance and calibrate their grades and feedback.

The grading of unanticipated student answers requires the strong support from a more senior individual. In the context of our course, while an instructor might feel comfortable giving the individual graders the power to handle these situations, it is probably most fair to have the one responsible for supervising grading, the graduate teaching assistant, making such decisions. Partly, this comes back to developing a culture in which the graders feel confident identifying abnormal work and feel comfortable to present that work to the graduate teaching assistants for support. However, while there were many cases of agreement with graders, we also identified several cases in which the grader grossly misjudged the level of achievement of student work. While we could easily blame this on the supervising graduate assistants, we have to remember that it is not feasible for them to check every single assigned grade. As such, we are developing data-driven ways to visualize how particular graders' scores look for specific learning objectives in comparison to the sections and course overall to facilitate the graduate teaching assistants' identification of anomalous grading. In the case of written feedback, we are attempting to do this through length of feedback and the use of word clouds. For marking, we are developing graphs that indicate proportions of assigned marks for each grader in comparison to overall averages.

Conclusion

Grading free-response work absolutely fairly across all students can be a challenging task for one individual grader, let alone a set of 75 graders. This is understandable, as grading decisions can be affected by a plethora of factors ranging from background experiences and knowledge to unclear or ambiguous criteria to the time or day or mood of the grader. Unsurprisingly, this variability is clearly demonstrated through our data from a large-scale, multi-section engineering course.

While understandable, this variability needs to be minimized in order to produce fair, high-integrity grades. While factors like time of day and grader mood are beyond our control, grader experience and clarity of grading criteria are within our grasp. As such, we have taken several actions to attempt to remedy unwanted variability. To minimize cognitive load on graders, we have modified assignments to make instructions and expectations clearer to students to produce more consistent work while also requesting all work involving MATLAB to be submitted using a function that consolidates all codes, comments, and output into a single document. To facilitate grading decisions and improve fidelity, we have clarified rubric criteria and the evidence for achievement to more effectively communicate expectations and delineate levels. Further, to improve reliability, we have been training our graders prior to the grading of each assignment using samples of past student work to calibrate their marks and feedback provided with those of the course developers.

The data identified through this study and the follow-up actions taken lead to a number of potential future research questions. For instance, most research conducted on rubrics investigate rubrics from a quantitative perspective. One possible study involves taking a more qualitative approach to explore the way graders interact with rubrics and student work to make grading decisions. This may help us to identify what aspects of rubrics are most likely to challenge graders. Additionally, our implementation of a training system and means to visualize marks and feedback assigned is providing us with ample data to investigate the results. This will also allow us to investigate relationships between how graders perform in their training and the quality of the grades that they assign to students.

Ultimately, these actions and potential areas of research are meant to help provide our students with a higher quality learning experience that is untainted by grader malpractice. For the sake of our students, we need to be assigning fair, high-integrity grades. With large-scale classes, the only way to practically achieve this (without limiting assignments to less meaningful multiple choice questions) is through teaching assistants. It is important to remember that while they may be paid to make our lives easier as instructors, it is necessary that we make their duties as easy and pain-free as possible. The easier our teaching assistants can produce high quality grades, the better the outcomes for everyone involved.

Acknowledgements

This study was funded by a grant from the National Science Foundation (NSF DUE 1503794). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

1. Armstrong, W. B. The association among student success in courses, placement test scores, student background data, and instructor grading practices. *Community Coll. J. Res. Pract.* **24,** 681–695 (2000).

2. Betts, J. R. & Costrell, R. M. Incentives and equity under standards-based reform. *Brookings Pap. Educ. Policy* **2001,** 9–74 (2001).

3. Muñoz, M. A. & Guskey, T. R. Standards-based grading and reporting will improve education. *Phi Delta Kappan* **96,** 64–68 (2015).

4. Braun, H. I. Understanding scoring reliability: Experiments in calibrating essay readers. *J. Educ. Stat.* **13,** 10–18 (1988).

5. Jonsson, A. & Svingby, G. The use of scoring rubrics: Reliability, validity and educational consequences. *Educ. Res. Rev.* **2,** 130–144 (2007).

6. Crisp, V. Judging the grade: Exploring the judgement processes involved in examination grading decisions. *Eval. Res. Educ.* **23,** 19–35 (2010).

7. Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R. & Schmitz, J. A. P. An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teach. Psychol.* **36,** 102–107 (2009).

8. Goldberg, G. L. Revising an engineering design rubric: A case study illustrating principles and practices to ensure technical quality of rubrics. *Pract. Assessment, Res. Eval.* **19,** (2014).

9. Moskal, B. M. Recommendations for developing classroom performance assessments and scoring rubrics. *Pract. Assessment, Res. Eval.* **8,** (2003).

10. Cooksey, R. W., Freebody, P. & Wyatt-Smith, C. Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educ. Res. Eval.* **13,** 401–434 (2007).

11. Hylton, J. B. & Diefes-Dux, H. A. A standards-based assessment strategy for written exams. in *Proceedings of the 123rd ASEE Annual Conference and Exposition* (2016).

12. Marbouti, F., Diefes-Dux, H. A. & Madhavan, K. Models for early prediction of at-risk students in a course using standards-based grading. *Comput. Educ.* **103,** 1–15 (2016).

13. Marbouti, F. A standards-based grading model to predict students' success in a first-year engineering course. (Purdue University, 2017).

14. Sadler, D. R. Grade integrity and the representation of academic achievement. *Stud. High. Educ.* **34,** 807–826 (2009).

15. Ashton, S. & Davies, R. S. Using scaffolded rubrics to improve peer assessment in a MOOC writing course. *Distance Educ.* **36,** 312–334 (2015).

16. Sadler, D. R. Fidelity as a precondition for integrity in grading academic achievement. *Assess. Eval. High. Educ.* **35,** 727–743 (2010).

17. Oakleaf, M. Using rubrics to assess information literacy: An examination of methodology and interrater reliability. *J. Am. Soc. Inf. Sci. Technol.* **60,** 969–983 (2009).

18. Stemler, S. E. A comparison of consensus, consistency, and measurement approaches to estimater interrater reliability. *Pract. Assessment, Res. Eval.* **9,** 1–11 (2004).

19. Pantzare, A. L. Interrater reliability in large-scale assessments – Can teachers score national tests reliably without external controls? *Pract. Assessment, Res. Eval.* **20,** (2015).

20. Griswold, P. A. Beliefs and influences about grading elicited from student performance sketches. *Educ. Assess.* **1,** 311–328 (2010).

21. Andrade, H. G. Using rubrics to promote thinking and learning. *Educational Leadership* **57,** 13–18 (2000).

22. Sadler, D. R. Interpretations of criteria-based assessment and grading in higher education. *Assess. Eval. High. Educ.* **30,** 175–194 (2005).

23. Popham, W. J. What's wrong—and what's right—with rubrics. *Educational Leadership* **55,** 72–75 (1997).

24. Adamson, K., Gubrud-Howe, P., Sideras, S. & Lasater, K. Use of the Lasater Clinical Judgment Rubric : *J. Nurs. Educ.* **51,** 66–73 (2012).

25. Millet, I. Improving grading consistency through grade lift reporting. *Pract. Assessment, Res. Eval.* **15,** (2010).

26. Krippendorff, K. Reliability in content analysis: Some common misconceptions and recommendations. *Hum. Commun. Res.* **30,** 411–433 (2004).

27. Mchugh, M. L. Interrater reliability: The kappa statistic. *Biochem. Medica* **22,** 276–282 (2012).

28. Reddy, Y. M. & Andrade, H. A review of rubric use in higher education. *Assess. Eval. High. Educ.* **35,** 435–448 (2010).

29. Meadows, M. & Billington, L. A review of the literature on marking reliability. *Rep. Natl. Assess. Agency by AQA Cent. Educ. Res. Policy* (2005).