Efforts to Improve Undergraduate Grader Consistency: A Qualitative Analysis

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.

Dr. Kerrie A. Douglas, Purdue University, West Lafayette (College of Engineering)

Dr. Douglas is an Assistant Professor in the Purdue School of Engineering Education. Her research is focused on improving methods of assessment in large learning environments to foster high-quality learning opportunities.

Efforts to Improve Undergraduate Grader Consistency: A Qualitative Analysis

Abstract

In this research paper, we explore the difficult decisions faced by large-scale, multi-section courses in early undergraduate engineering education regarding fair and consistent assessment of student learning across sections. Our previous analysis of grading patterns of undergraduate graders in a first-year engineering course revealed that divergent decisions likely stemmed from two sources: insufficient grader training and ambiguities in rubrics and assignments. After revising rubrics and implementing grader training for a semester, we conducted think-aloud interviews with 17 undergraduate graders regarding grading, rubrics, and training. Qualitative analysis identified four technical aspects of rubrics that led to divergent grading decisions (wordiness, redundancies, unexpected solutions, and grade misfit) and five aspects that limited training effectiveness (length, misalignment, insufficient feedback, limited consequences, and philosophical misunderstanding). These findings contribute nuance to and extend upon aspects of rubric design and undergraduate grader training that have been previously identified in the literature. Recommendations related to issues identified are provided.

Introduction

Fair and effective assessment of engineering knowledge and skills in a way that can be instructionally useful is a formidable challenge. With calls for learning of deeper level engineering skills [1], [2], the use of open-ended problems for assessment has become of greater importance. While multiple-choice or fill-in-the blank type assessments allow for fast, reliable grading, both options severely limit the range of skills that can be authentically and accurately evaluated. Alternatively, open-ended problems enable students to demonstrate a wider range of skills but require significant time to grade. Hence, large scale courses that hope to assess a wide range of skills authentically rely on graders to help manage the heavy workload. Open-ended problems are challenging to grade consistently and fairly by even one grader, let alone many graders across multiple sections. To fairly and consistently grade all learners, course instructors must identify and employ mechanisms to minimize such variability of grading.

The challenges of providing fair and consistent grading across many students are common in engineering, particularly at the introductory level where students from each of the disciplines often take the same fundamental courses. Large introductory level engineering courses that span multiple large sections depend heavily on the assistance of graduate and undergraduate teaching assistants (GTAs and UTAs, respectively) to provide students with timely and individualized feedback [3], [4]. Use of common multiple choice or short-answer assessments across sections would allow for fast and consistent grading, but would greatly reduce the range of skills that could be assessed authentically [5, pp. 86–87]. On the other hand, open-ended problems, much like those frequently encountered in engineering, require well-established rubrics based on specific learning objectives (LOs) to achieve adequate consistency [6]–[8]. Further, the use of a LO-based approach helps to achieve greater consistency of general experiences across sections, in addition to grading.

While standardized course materials, from lecture slides to assignments and assessment, and LO-based rubrics can help immensely to reduce differences between sections, each member of an instructional team brings forth their own background and experiences that influence the way they teach and interpret course content. This problem can be further exacerbated when instructional teams include GTAs and UTAs, who are often less experienced in the classroom and not as well versed in the material, and therefore tend to be less consistent with grading [9]. To magnify the impact of this concern, evidence suggests that the grading practices a student experiences can be strongly associated with their future academic success [10]. In other words, if a student experiences or perceives unfairness in the evaluation of their abilities, there can be a lasting effect on their academic success. When these courses are meant to prepare students for their subsequent engineering studies, it is, therefore, of the utmost importance to employ whatever measures possible to maximize equality of experiences and fairness of grading for all students, regardless of course section.

To explore how well LO-based rubrics help with grading consistency, we previously conducted a study that compared the actual in-course grading decisions made by UTAs with the grades that would have been assigned to the same student work by expert graders [11]. Our investigation revealed frequent discrepancies between grading decisions made by the UTAs and the expert, with agreement occurring in less than half of the samples analyzed. In some cases, the rubrics and assignments, in conjunction with the student work, demonstrated considerable ambiguities that made the variable grading decisions understandable; in many other cases, however, despite the student work clearly fitting into specific levels of achievement specified by the rubrics, UTAs were still wildly inconsistent in their grade selections. Based on these analyses, we recommended that the rubrics and assignments be revised to reduce points of ambiguity and that UTAs receive training on the application of the revised rubrics.

In this paper, we present a follow-up to the previously discussed study. After revising assignment and rubrics and implementing a set of weekly online rubric training modules for the UTAs, we conducted interviews with a subset of UTAs asking for them to perform some grading activities and discuss their thoughts about grading and training. Focusing on a qualitative analysis of the UTAs responses, this paper addresses the following research questions:

1. What did UTAs perceive as troublesome while applying rubrics? And,
2. What were the UTAs perspectives of the training process?

Background

Rubrics. To an extent, the grading of open-ended tasks inherently includes a degree of subjectivity, which may be influenced by a plethora of factors ranging from the grader's knowledge, experiences, values and beliefs about grading, and perceptions of the grading practices of other graders [9], [12], [13]. Rubrics, which are two-dimensional matrices of criteria or standards versus levels of achievement, are intended to be tools to minimize the effects of grader judgment [14]. Previous scholars have noted that rubrics are more likely to produce errors when they are redundant, have limited options for partial credit, have uneven intervals between achievement levels, and exhibit inconsistencies in focus or form [15]. Rubrics must

also avoid being excessively detailed or excessively general and should be bias-free, well-aligned with performance tasks, and written at an appropriate level for their users [16], [17].

Training. While improved rubric design can reduce grader error, training may be the most important factor to strengthen the reliability of grading [18]. Inexperienced graders require guided practice to be able to consistently apply a rubric and having graders apply rubrics to samples of student work helps them to calibrate their judgments [19]. The process of training allows graders to establish mental models of work at each achievement level and has been shown to result in less variable, more accurate scoring [19], [20]. However, training that illustrates only a subset of achievement levels might prevent graders from using a full range of scores [21].

A survey of GTAs spanning multiple universities found that many feared grading, while many others feared not knowing answers and having to explain concepts—both of which are necessary for providing good feedback during the grading process [22]. It is, therefore, reasonable to assume UTAs might have similar fears. Also notable, another study of GTA's training needs identified that the majority believed fair and consistent grading to be their most important responsibility, suggesting they would be amenable to calibration training [23].

There are many different models for training teaching assistants [24]. The vast majority of training models tend to be single-day workshops [25], but others use week-long workshops [26], weekly courses [27], periodic training videos [28], or multiple multi-phased sessions[24], [29], [30]. These training sections are often interactive, including opportunities for discussions and group exercises [31]. In the multi-phased approach, the teaching assistants were first asked to complete the assignment they would be grading first, followed by in-person training reviewing and applying the rubric, and then feedback in the form of comparison of their grade selection with the expert selection for calibration purposes [24], [29], [30]. This strategy informed the training design used for this paper.

Context

This research investigated the UTAs in the second course in a two-semester, first-year engineering course sequence at a large, Midwestern university during Spring, 2017. The course consisted of 14 sections with up to 120 students per section. Except for a few exceptions, each section had one GTA who oversaw five UTAs (four who provided in-class support and assisted with grading and one whose sole responsibility was grading).

Rubric revisions. Following the previously discussed research investigating the consistency of UTA graders in this course, each assignment and rubric was reviewed and revised by a four-member team including an instructor, a graduate student, and two instructional support staff. The goal was to improve alignment between the two and to make the rubrics as clear as possible (i.e., to minimize what we could identify as possible points of confusion or ambiguity). Figures 1 and 2 illustrate the difference in the rubric for a specific LOs before and after revision, respectively. Note that the revised version indicates the specific portion of the student work to grade, provides greater detail through distinct evidence items, and provides a full range of partial credit scores.

Criteria	Ratings			
	No evidence of achievement:	Underachieved 50%	Partially achieved 80%	Fully achieved 100%
Linearize and plot data appropriately	Anything less than requirements for underachieved	NA	NA	Problem 1 ☐ 1.2b. Log values are taken of both x and y and plotted on linear scales
	0 pts ☐			1 pts ☐

Figure 1. Example rubric item prior to revision.

Learning Objective	13.07 Linearize and plot data appropriately			

What to Grade:	PS07_beach_logins.pdf > LINEARIZED DATA *Prob 2, Step 5:*
	Grade the linearized data on the linearized data plot.
	NOTE: do not grade the regression line or any formatting other than what is below.
	`% linearize the data for use in power function`
	`log_offshore = log10(offshore);  % log of offshore distance`
	`log_depth = log10(depth);  % log of water depth`
	
	`% Plot linearized data`
	`figure(2)`
	`plot(log_offshore,log_depth,'g*')`
	`xlabel('Log (Offshore Distance in Meters)')`
	`ylabel('Log (Depth in Meters)')`			

Proficient	**Developing**	**Emerging**	**Insufficient Evidence**	**No Attempt**
1 pt	0.8 pt	0.5 pt	0 pt	0 pt
Evidence items for proficiency: 1. Linearize the independent variable data correctly based on the diagnosed function type • Power: log of independent data 2. Linearize the dependent variable data correctly based on the diagnosed function type • Power: log of dependent data 3. Axes labels (description and units) are correct based on the plotted data You will need to see what they plotted to see if their units match what's in their plot command. You'll grade whether or not they plotted the correct information in the next LO	1 (of 3) missing or incorrect item from the proficient list	2 (of 3) missing or incorrect items from the proficient list	3 (of 3) missing or incorrect items from the proficient list	Did not attempt the graded item

Figure 2. Example of the same rubric item after revision.

Training design. The training for this study is based off of the multiple multi-phased training described previously [24], [29], [30]. However, as we wanted to train the UTAs for each new LO covered, it was necessary to have weekly training. Due to the highly variable schedules of the 69 UTAs for the course, the most logistically feasible approach was to create online training modules. While the UTAs were expected to complete their training prior to applying the corresponding rubrics to assign actual grades, this option would provide enough flexibility to allow the UTAs to complete the training whenever was most convenient. Still, we attempted to retain part of the multi-phased training structure by giving the UTAs the rubrics, a sample problem, and the solution to the problem (see Figures 3-5, respectively). The UTAs were then given an example of student work for which they would take a quiz asking their selection of the proficiency level, the evidence items they believed to be missing from the student work, and the feedback they would give the student (see Figures 6 and 7). The UTAs were given the proficiency selection, missing evidence items, and recommended feedback of an expert grader. The UTAs then repeated this process for a second example of student work. Thus, for each LO covered during the semester, the UTAs took at least two quizzes to calibrate their grading.

Learning Objective (LO): 13.05 Create plots with linear and/or log axis scales (Excel)

- The student solution must be evaluated for each of the **9** items of proficiency evidence.
- Note: When grading this learning objective, you might be directed to just look at one of the plots (rather than all four). For this training assume that all four plots must demonstrate each evidence item.

Proficient	Developing	Emerging	Insufficient Evidence
• Plots of data using different axis scales to show relationships useful for function discovery ○ Linear scale: linear scale on x-axis, linear scale on y-axis ○ Log-linear scale: log scale on x-axis, linear scale on y-axis ○ …	1-2 (of 9) missing or incorrect items from the proficient list	3-4 (of 9) missing or incorrect items from the proficient list	5 or more (of 9) missing or incorrect items from the proficient list

Not Assessed by this LO:
Format of the plot for technical presentation

Common Student Mistakes:
Students will label the x and y axis as log(x) or log(y) if log scaling is used, where x and y would be specific to the context of the problem. It must be just x and y as there has been no transformation of the data.

Figure 3. Grading instructions and rubric, including what is not assessed and common mistakes.

Problem
The student is asked to use function discovery and data transformation to model the relationship between earthquake intensity and magnitude. The student is provided with a dataset containing moment (in gigaNewton-meters or GN-m) and magnitude.

The student must plot the data on all combinations of linear and log scale axes.

Figure 4. Example problem for the LO being trained.

Solution

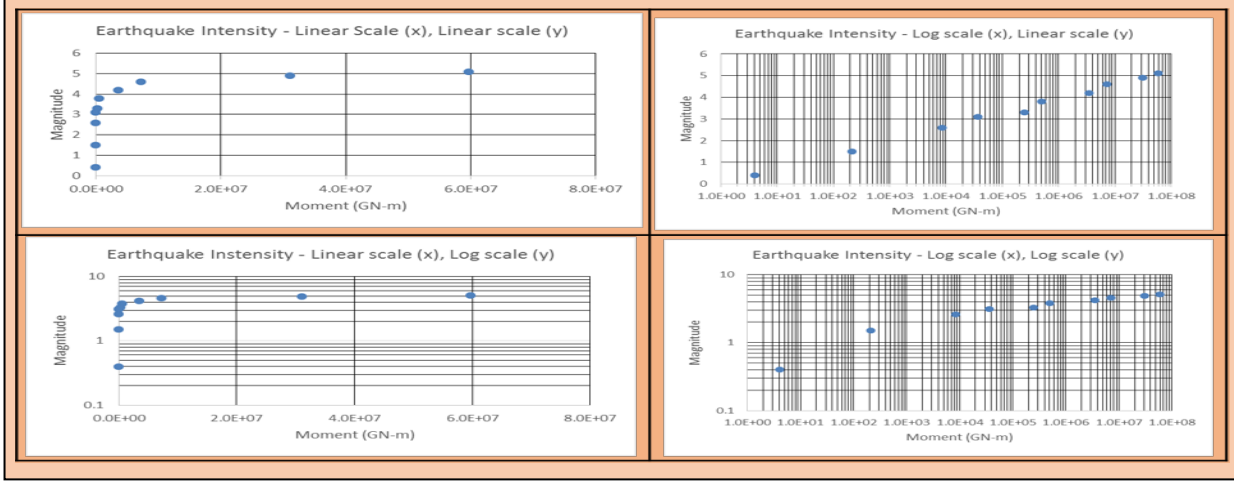

Figure 5. Solution to sample problem from Figure 4.

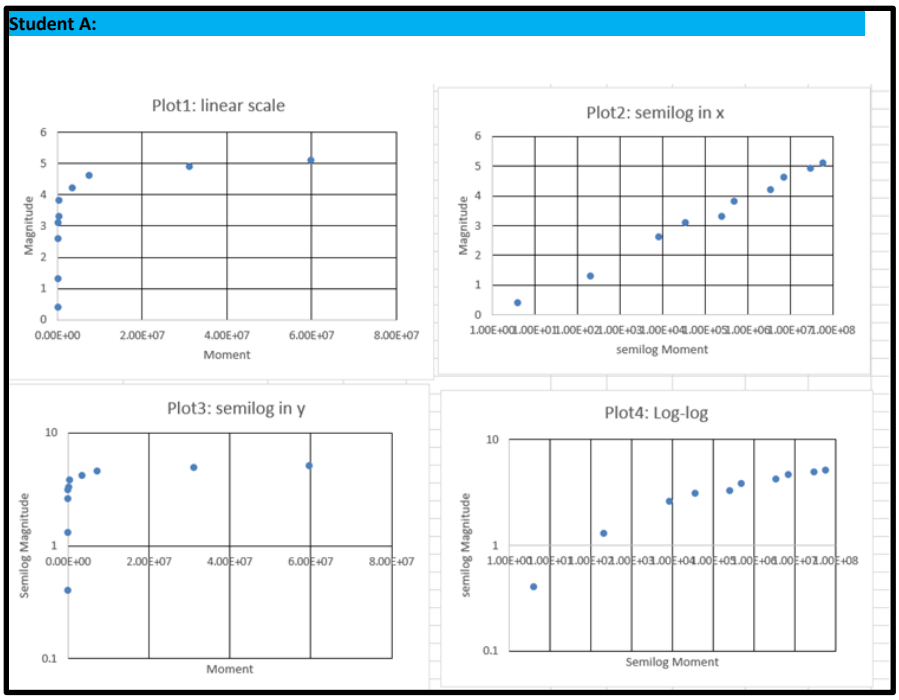

Figure 6. Sample of student work for first calibration quiz.

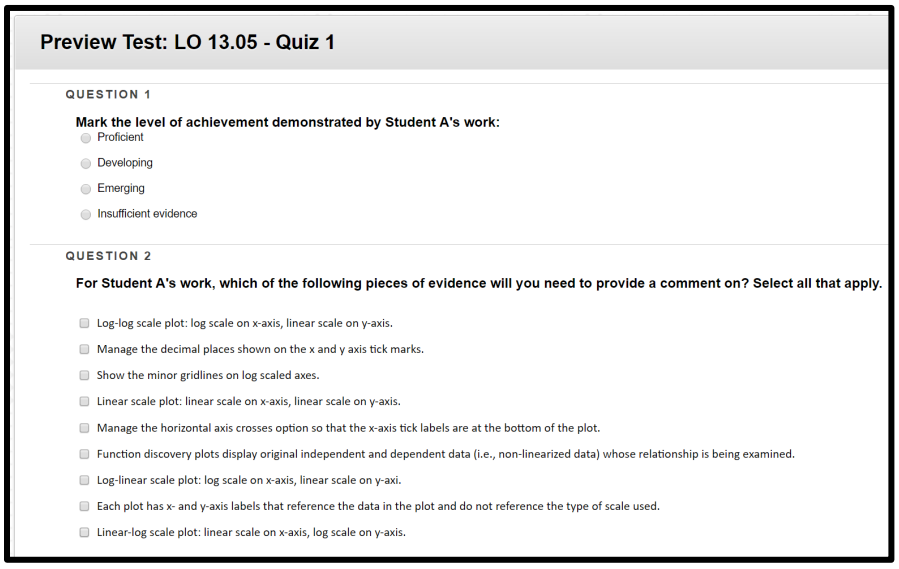

Figure 7. Training module calibration quiz.

Methodology

Paradigmatic perspective. As argued by Adamson, Gubrud, Sideras, and Lasater, "one of the greatest threats to the reliability of data produced from observation-based performance evaluation instruments is perception, or human judgment; one rater may perceive a performance differently than another rater and subsequently rate it differently" [20, p. 68]. However, even very well developed rubrics do not guarantee high reliability [12]. Thus, for this research, we adopt an interpretivist paradigm, which assumes every person constructs their own subjective reality [32], to determine differences in UTAs' understandings of the rubrics. Within the context

of the proposed research, each UTA will impose their own knowledge, experiences, and perceptions when interpreting not only the rubrics but also the work they grade. This means interpretive differences will likely occur in how the graders perceive the intentions of the students whose work they grade as well as how to apply the rubric itself. Assuming every grader intends to accurately assess student work, any instance in which two graders assign different grades to the same piece of work supports this interpretivist position. It should be noted that adoption of this perspective does not suggest that there is a single "correct" interpretation of a rubric; rather, there is an interpretation that the rubric designers had in mind. Thus, this perspective assumes that it is possible to design rubrics to minimize differences in interpretation.

Participants. At the end of the semester, we reached out to all the course UTAs to participate in one-hour interviews that consisted of a think-aloud portion and semi-structured questions about their experiences with grading and training. Of the 69 UTAs, 19 expressed interest and 17 ultimately participated. The participants were compensated $20 for their time. In addition to these students, we also interviewed three faculty instructors and one instructional support team member who contributed to the development of the assignments and rubrics. The results of their interviews are not included in the present analysis but will be used in future work.

Interviews. The interviews of this study consisted of two portions. First, the participants were given a rubric from one of the semester's assignments and three samples of real student work corresponding to each rubric item and were instructed to think aloud while making grading decisions. This think-aloud portion follows the recommendations of the Boren and Ramey, whereby participants are mostly allowed to perform tasks without interventions from the interviewer unless they have been silent for a prolonged period and need to be prompted to verbalize their thinking [33]. It is important to note, however, that the very presence of the interviewer exerts an influence on the thinking of the participant and this influence is greater if the interviewer must intervene [33]. Thus, this must frame interpretation of the results. The second portion of the interview consisted of more general questions about rubrics, grading, and the training. Writing and audio of the participants were recorded using the Notability App on an iPad.

Analysis. For the think-aloud portion of the interviews, all grading decisions were documented for each evidence item in each rubric item for each sample analyzed and comments that were made when the participants made those decisions were documented along with those decisions. Statistical analyses were conducted on the specific decisions made to identify rubric items that were applied more-or-less consistently; however, those analyses will not be presented here. The questions and answers during the second portion of the interviews were also transcribed for further analysis.

This paper focuses on the themes that could be identified purely through analysis of the comments the participants made during their think-aloud portions and their semi-structured questions about grading and training. Comments made during the think-aloud and the transcriptions of the second portion of questions were analyzed following Rubin and Rubin's approach to qualitative coding and thematic analysis [34]. The major themes identified throughout the interviews are presented in the following section, separated according to

interview portion (i.e., rubrics and training). These themes were further bolstered by themes identified within the observational memos recorded during the interviews.

Results and Discussion

Rubrics issues. Despite the revisions made to the rubrics, the think-aloud interviews demonstrated that several problems persisted with the application of rubrics. In response to the research question, "What did UTAs perceive as troublesome while applying rubrics?", we identified four major themes. It is notable that across the last three of these themes all relate to the idea of fairness, indicating how highly the UTAs who participated in our study value fairness in their job, and supporting the values identified by Cho, Sohoni, and French [23].

Length and wordiness. As Popham noted in his review of rubrics, while it is important to provide sufficient detail to fully communicate expectations, it is also important to avoid being too detailed or wordy [16]. Multiple participants noted that the rubrics made the grading process "tedious" because they take "so long" to go through. As one participant noted:

> "There are too many hoops to jump through. For each learning objective I have to find like 13 evidence items to check off, yes or no. That would take a lot of time commitment going through, one-by-one, 32 or 48, or however many students you have at that time."

While this might be perceived as a minor annoyance for the UTAs, the issue can have larger effects. For instance, observational notes taken during the interviews highlighted that for particularly long rubric items, the participants seemed more likely to skim over the text and potentially miss important information. Other participants admitted that too much text or too many LOs causes them to not read carefully:

> "It just ends up being a wall of text that I'm just going to look for the first buzzword out of the first sentence, like 'flowchart' or 'If selection structure.' Then I'll just remember that for that evidence item and the other 11 evidence items or whatever are just getting in the way."

Consistent grading requires that all graders pay attention to the same details. Rubrics that are too long increase the chances that a grader will overlook something important. As each grader is different, it is unlikely that they will all overlook the same details. Thus, the best way to ensure control over what is graded is to streamline the rubric to focus on only the most important aspects for grading and to state those aspects as succinctly and clearly as possible.

Redundancies, interdependencies, and subsets. Goldberg recommended reducing redundancies in rubrics [15]. This study, however, highlighted less obvious forms of redundancy that were not obvious when the rubrics were originally written. One example of a redundancy that causes issues with graders is when the same LO is assessed more than once on an assignment. For instance, in the problem set used for the think-aloud, the same two LOs based on the creation of a table of test cases were used for two different problems (that is, they accounted for four of the assignment's 10 LOs). One participant's observation was representative of comments made by several others: "The first test case table is similar to this test case table and I find this unnecessary." This redundancy caused concern for some participants,

as it seemed that the assignment potentially inflicted twice the penalty if a student misunderstood that one concept.

Redundancies of assessing the same LO repeatedly seemed to annoy participants but were less likely to result in variability of interpretation and grading decisions than redundancies associated with interdependent evidence items. For instance, one LO about flow charts had an evidence item that stated, "operations are connected with arrows…," another evidence item that said, "arrows must connect all flowchart elements…," and a third that said, "arrows must converge prior to stop." An exact interpretation would suggest that failure to include arrows would result in missing all three evidence items. Participants who were more confident and who displayed a greater sense of autonomy felt comfortable interpreting the latter two evidence items as being achievable without arrows, but others were either confused or conflicted about their interpretations: "They didn't do one thing. They missed one part of their structure and it just hit them. It came back to haunt them multiple times, over and over again, in the rubric. It pains me, but I've got to follow it anyway." This discrepancy supported Crisp's assertion that different graders adhere to standards with varying levels of rigidity [9].

Yet a third form of redundancy occurred as a result of an attempt, by the rubric designer, to make things easier on the graders by selecting a subset of the student's answer to be graded. For example, in this problem set, rather than asking the grader to look at and grade an entire table of test cases, two specific test cases were specified to be graded; because there were two separate LOs associated with the table, those same two test cases were graded for both LOs. While this approach reduces items for the grader to look at, the participants felt conflicted when the sample work was missing one of the specified cases. In the words of one participant:

> "I think just sometimes the way it's broken down like this where you grade for the same one each time, like they're doing something right but they just don't do *that* one right. It's unfortunate because you kind of should have to grade the whole thing. I don't know if it's just the way it's broken down in the rubric sometimes, if there's a better way to move it so we're not… like they could still get points for doing. If they only had laminar flow they could still get points for it instead of saying they didn't do invalid and the rubric asks for invalid so you're going to get it pretty much all wrong."

With this type of redundancy, we did not witness differences in interpretation, but we did see nearly ubiquitous concerns that the learner demonstrated competence with other test cases despite missing the specified case.

Unexpected or unconventional student solutions. One prominent theme in our study, which we did not identify directly in the literature, is the ambiguity caused by rubric designers unintentionally assuming students will follow a particular solution path when others are possible. In the problem set studied, one question asked the students to write code to calculate the temperature at a given height above sea level based on atmospheric layer. The rubric assumed this would be coded using an efficient if-elseif-else selection structure, but as one student's work drew to our attention, this can also be accomplished, albeit less efficiently, through a series of if-statements. However, the assignment did not clearly specify that the if-elseif-else was required.

This led to confusion and divergent decisions for our participants. As one participant stated, "So, it's like, 'Well, do I give them a zero?' Do I say, 'Well, you did this,' so there's a lot of things I feel like the LOs assume and don't tell you what to do when that assumption is missed. That's probably the biggest thing that's confusing to me." It is easy to see how different UTAs might go different directions in this situation, particularly given that the assignment did not demand a specific approach.

Many of the participants were extremely conflicted over this issue. They expressed beliefs, understandably, that in engineering we should be encouraging creativity and the ability to solve problems, regardless of the specific method used. They noted that because the students are not told specifically what LOs will be used for which problems, that it is not fair to expect or assume a particular solution path. One participant questioned, "We need some kind of consistency in there, and are we going to tell them we want them to learn and figure out and develop their own way? Or are we going to tell them, 'Hey, do it this way.' There's not consistency on either side of the board."

Misfit between grade and achievement. This theme relates to Goldberg's question of whether or not adherence to the rubric produces cognitive dissonance (i.e., does it fail the "fit" test?) [15], and possesses some overlap with the issues of redundancy and unexpected responses. Specifically, participants vocalized sizeable frustration when they felt that the grade a strict interpretation of the rubric dictated did not correspond with the grade they felt the student deserved, whether better or worse. This was generally the result of either grading only portions of a student's work, or because the participants did not feel that the LO used to assess the problem was accurately representative of the student's ability to solve the problem. As one participant stated, "There are some students who do everything right, but they don't do this one thing right, and then we grade that one thing like four times over, so they're losing all these points because they didn't do it."

Another student went on a rant, playing out a hypothetical conversation with a student who received what he considered to be a higher score than the student deserved:

> "'Well, you're going to get zero points.' 'But plot one had a perfect representation.' 'Who cares? We didn't grade plot one.' But then some students come in like, 'Why did I get points off?' 'Because you had a bad plot two.' Then they'll be like, 'But what about my plot one?' 'I didn't even look at it.' Why? Like, really, like what are we going to say? Students sometimes come in and ask, if they're preparing for an exam, they're like, 'Can I ask you a question about a homework assignment? I got full credit on this but I know I did it wrong.' Then I look at it. 'Yeah, you did this completely wrong.' 'How did I get full credit?' 'We didn't grade this problem.' Can I say that to students?"

In a similar sense, another participant felt conflicted that the structure of the evidence items was such that a student could earn full points on code that fails to run: "I feel like there are just sometimes where kids code doesn't … like, they don't run, and they did something wrong, but they still did every evidence item and they get full points for code that doesn't run. That's sometimes an issue."

Finally, a slightly different sub-theme relates to questionable item weighting. As one participant noted, the rubric often treats all evidence items equally; however, that practice does not fairly represent of the importance of each item:

> "I wonder if there's implicit … so if I'm looking at this rubric right here that has 11 items on it, implicit in this is that each of these items has equal weight in terms of quality of the solution, so that if any one or two of them, regardless of which one or two they are, that gets you developing. Any three or four gets you emerging. And, as a practical matter, that's just not true."

In other words, treating all items of evidence of achievement of an LO as equal potentially leads to misrepresentations of the extent to which someone has mastered that LO, as some aspects are, inherently, more important than others.

Training issues. The semi-structured interview questions indicated four common perspectives of our participants regarding training, but also highlighted that many of these participants suffered from a fundamental philosophical misunderstanding of the purposes of LO-based grading and training. It is important to note that while the themes identified below focus on negative perspectives and areas for improvement, many of the participants did indicate that, overall, the training helped them to provide better feedback to students and refreshed their minds of the content to better assist students in the classroom and grade more efficiently.

Length and repetitiveness. Due to limitations with learning management platforms, the structure of the training was admittedly clunky. UTAs had to download a zip file for each LO which contained separate files for the rubric, the problem and solutions, and the samples of student work. To get immediate feedback between each sample of work, they took a separate Blackboard quiz for each sample. As a result, in some weeks, the UTAs were expected to complete up to 20 quizzes. Several of the participants in our study expressed these concerns: "It was a little time-consuming for those considering there were two or three student works per item. It would probably be better just to have one incorrect one that has things wrong with it to give you a good idea of what you need to do."

Another participant admitted that the training had value, but expressed definite frustration:

> "Okay, the training is pretty simple and straightforward, and it is needed, but the thing that I didn't like is that we had to take quizzes, and it's pretty much the same thing, and we have 15 quizzes every week. I guess I don't like spending so much time of my sophomore year taking quizzes, and we are basically taking more quizzes than the students themselves, so I just don't like that part of the grading right now."

A revised training program should take measures to streamline the process so that there are fewer documents involved and the UTAs can train more efficiently.

Training and assignment misalignment. As the assignments and rubrics were all revised between the previous term and the first term with training, there were no actual samples of student work available for the training. Further, because the trainings were presented to the

UTAs before the LO topics were presented in class, the modules had to be designed even further in advance, at which point some of the assignments and rubrics had not yet been finalized. As a result, many of our participants expressed frustration that the problems used in the training were not always perfectly aligned with the assignments they would ultimately grade. One participant succinctly stated, "Sometimes the rubrics we trained for are not the rubrics we used to grade. What's the point?" Another similarly said, "How are we supposed to relate the two if it's not the actual example we're going to jump into?"

On top of some misalignment between training problems and rubrics, because no students had completed the assignments in the past, there were no authentic samples to use in the training. Instead, the samples were typically created a faculty member. Far from the mind of a student, it is understandable that one participant recognized, "Sometimes the samples we're given are just so far off that it's not what you would usually see from a student so it's kind of hard to grade."

To minimize the issue, one participant recommended that some of the UTAs be involved in helping with the rubric and training development: "First off, they need to have the problem sets done more than a day before we release them with the rubric. But I think they need to have a couple of UTAs who are willing to sit down and look at the rubric, and an example with them, be like, 'Okay, this makes sense. This doesn't make sense. Why are we making them do it this way?'"

Certainly, the perspectives of the participants lack some of the knowledge of the complexities of developing these materials for large, multi-section courses. While they likely do not appreciate the time and effort necessary to develop strong assignments, rubrics, and training materials, the graders' requests to have practice examples using the rubrics with authentic student work likely would be helpful. Further, involving some of the UTAs in the process could help to proof the materials and generate samples.

Insufficient feedback. Many of the participants of the study expressed that neither the training, nor the subsequent implementation of grading, provided them with adequate feedback to evaluate their performance. One participant noted that the feedback they received from their GTA did not correspond with how they felt they had performed:

> "I think my grading, sometimes I grade like I don't know, I might have grading inconsistencies this time. I think that way and the GTA comes to you and says you did a great job grading. Your feedback is detailed enough for students. Really? I didn't write much. Training quiz recommended feedback is like a paragraph. I write a sentence. The GTA comes in and says, 'you did great.' Well, guess what? I'm still writing sentences now."

Another participant expressed frustration with their frequently poor performance on the training quizzes and recommended a possible way to help improve the feedback they obtained:

> "I wouldn't understand why I was getting them wrong. Either meeting with your instructor or meeting as a whole group once a month and actually going through the quizzes and understanding how to grade them, as opposed to just reading the feedback

that they gave the students, I think would definitely have helped and left me with a better understanding of the problem. Rather than just, 'Oh, lost another point. Got to go onto the next quiz.'"

Perhaps supplementing the online training modules with group-based reviews of the content and expert grades would bring the training design closer to the original design upon which this model was based (e.g., [29]).

Limited consequences for improper training. The participants in the interviews also recognized one of the biggest problems that the instructional support team identified with the training program—there was little to no leverage to enforce the training on the UTAs. The instructional support team was not in a position to let go of UTAs for not training or not taking training seriously, as replacing them mid-semester would be extremely difficult and losing them would place undue burden on other members of each their instructional teams (i.e., the other UTAs and the GTA). As such, the participants noted that there were little to no consequences for not taking the training seriously:

> "The problem is, we don't care that much. Let's say we did the training and then we … So, it's one out of one point, each training quiz is one point. We got one wrong, so it's a zero out of one, right? So, what? Nothing happens. It's just training. Training is done. It's there, but it's practically useless. Returning people know how to do it, no matter what the rubric says, no matter what the training quiz says, we're probably not going to change our behaviors because our professors and GTAs have been fine with it."

Philosophical misunderstandings. Various aspects of the previously discussed themes have shown a glimpse of the fact that many of the participants we interviewed demonstrated a fundamental philosophical misunderstanding regarding our use of LO-based grading in the course and the purposes of training. The fact that many of the participants felt frustrated and wanted to give students credit for writing functional code while the rubric specified insufficient evidence of achievement shows that the UTAs do not seem to understand that LO-based grading is a means to demonstrate achievement of specific skills, knowledge, or abilities. The UTAs, like most students, have a product-oriented perspective, while the instructional designers have a process-oriented perspective.

One participant questioned, "Let's say there were four questions. We did one, three, and four. We grade one, three, and four. What about two? What if students don't do it? If students somehow … 'I don't feel like doing two this week. Oh, wait, I still got a full grade. How did that happen?'" Another participant complained, "I think sometimes the rubric is too focused on what exactly their solution is, as opposed to whether it runs correctly. Because it could still run correctly. But if you don't have exactly what the rubric calls for, then you get points marked off. I think that's something that's really caused an issue with me and the other TAs in my section." Adopting a mindset of LO-based grading, these concerns of our participants are moot points. It is notable, however, that the way our rubrics focus on specific portions of work rather than allowing general identification of achievement of the LO prevents this overall perspective from being perfectly implemented.

In terms of the training, several comments indicate that the participants view training as a means to "teach them how to grade," as if they do not understand the process, rather than a process to calibrate their interpretation of the rubrics and student work. One participant said, "This is my third semester doing this. I kind of know this stuff. I've seen this stuff. I know where kids mess up. If we already understand this stuff and how we're grading, why do we need to be taking these quizzes?" Yet, "knowing the stuff" does not guarantee consistent selection of proficiency levels, which was not obvious to the graders. Similarly, another participant's statement shows the perception that they believe it is training how to grade, "it's not like a new problem set will have a different way of grading it. It's just all the same grading."

Participants also demonstrated a lack of understanding the calibration process by explaining that, "if I get something wrong when I'm taking the quiz, I think there's something wrong with the quiz or the rubric." This participant is not reflecting on the results of the training in a way that will improve consistency. They perceive training as being quizzed to prove ability rather than the calibration process it is intended to be.

Implications and Recommendations

Each of the themes identified in the previous section indicate potential root causes for inconsistencies in grader interpretations, decisions, and behaviors. While some of these issues have been identified previously in the literature, these findings present a few new ideas and provide additional nuance to refine or extend upon old ideas. Notably, these findings highlight the importance of emotions on grader behavior. For instance, the study shows that even when the right grading decision may seem objectively unambiguous, assuming direct and literal rubric interpretation, a UTA's feelings of annoyance, frustration, confusion, or perceived unfairness of scores may result in divergent decisions. Further, this study shows that the assumption of direct and literal rubric interpretation may itself be flawed, as different each UTA's sense of autonomy affects their comfort in taking liberties with specific wording. While training can theoretically reduce these issues, this study shows that issues with training design or enforcement may reduce the intended calibration of decision making.

The findings of this paper reinforce the idea that design is iterative, which presents a challenge in educational contexts. Even using extant literature to guide rubric and training design, actual implementation uncovers previously unidentified assumptions or complications. Seeing how students interpret assignments provides both material for future training and insights for revising or clarifying assignments and rubrics. Students' unexpected or unconventional approaches to solving problems should be documented along with the corresponding LOs to improve robustness of future assignment and rubric iterations.

In addition to these implications, we can make some recommendations regarding each of the themes identified in this study. Table 1 and Table 2 present the rubric-related themes and training-related themes, respectively, with short descriptions and related recommendations.

Table 1

Rubric-related issues and corresponding recommendations

Issue	Description	Recommendation
Length and wordiness	Rubric items that have too much text or too many evidence items may reduce grader attention and focus.	Rubrics need to be succinct and clear. Eliminate any non-essential pieces of information or evidence items.
Redundancies, interdependencies, and subsets	UTAs feel conflicted by lack of fairness when evidence items, rubric items, or portions of work to grade are repeated.	Minimize the potential that a single mistake or misunderstanding by a student will be repeatedly counted against them.
Unexpected or unconventional student solutions	Unexpected student solutions may not be handled well by the rubric, causing confusion and divergent decision making.	For each LO, keep documentation of common student errors for refinement of future assignment and rubric iterations. Clearly specify in assignments when specific approaches are expected.
Misfit between grade and achievement	UTAs do not feel all rubric scores are representative of actual student performance (either too high or too low).	Clearly communicate to graders whether grading is meant to be based on achievement of LOs or functionality of solution. Also, carefully consider weight of each evidence item.

Table 2

Training-related issues and corresponding recommendations

Issue	Description	Recommendation
Length and repetitiveness	Amount of training and overwhelming number of documents makes training tedious and time-intensive.	Streamline the process as much as possible both in terms of number of files necessary and number of calibration quizzes.
Training and assignment misalignment	As training occurs in advance, training may occur before questions have been written or completed by students. Authentic examples for training do not exist.	Make sure questions are fully developed prior to training. Use the documented common mistakes for each LO or use students or TAs to help generate examples to use for training.
Insufficient feedback	UTAs do not feel that they receive sufficient feedback in order to evaluate the accuracy of their grading, indicating they do not perceive sufficient oversight.	Feedback from quizzes should clearly and explicitly explain the expected grading decision. GTAs should give UTAs individualized feedback based on grading and quiz responses.
Limited consequences for improper training	When there are little to no consequences for not training or not taking training seriously, there is little to no incentive to push UTAs to do so.	Plans and contingencies should be developed from the start of the term to reward authentic training participation and punish inauthentic or non-participation.
Philosophical misunderstandings	Many UTAs do not have a proper understanding of the intentions of LO-based grading or calibration training.	Repeatedly communicate the intentions of grading (identifying LO competence) and training (calibration, not how-to-grade).

It is important to emphasize that a culture of strong communication toward and from UTAs is crucial for improving all of the issues identified. It might be helpful to allow even a few UTAs, as the ultimate users of the rubrics, to provide feedback regarding the clarity and usability of a given rubric. Further, UTAs can contribute by generating bad solutions to identify weaknesses in the assignments or rubrics related to unrecognized assumptions of solutions or unidentified redundancies or interdependencies. Further, it is important to communicate to the UTAs why the grading system is the way it is and the purposes of training (to obtain buy-in and commitment), as well as the consequences for improper participation. Finally, UTAs want their voices to be heard and want a consistent and reliable source of support with grading. As such, it is necessary to have a clear structure and to monitor UTA training and grading so that sufficient feedback can be provided to improve performance.

Conclusion

The qualitative analysis of these interviews revealed several concepts related the design of effective rubrics and UTA training. While many of the rubric-related themes correspond to previous claims made in the literature, each theme slightly extends upon what has been previously presented.. Ultimately, all of the rubric-related themes identified highlight that good rubric design is an iterative process that requires testing rubrics with actual graders using authentic student work to identify all of a rubric's shortcomings, as the weaknesses are exposed by flawed or unexpected student solutions.

Our attempt to transition a traditional in-person training program to online modules certainly showed room for improvement. While general sentiments were that the training was beneficial, the process needs to be streamlined and better connected to the assignments and rubrics for which they are training. However, the most important themes identified through these interviews for making training more effective relate to the ways the UTAs understand and value the course's design and the purposes of training. The reasons for using LO-based grading and the idea that training is intended to calibrate grading decision needs to be strongly and effectively communicated to the UTAs for them to approach the training program with a proper mindset. Additionally, UTAs can be used as a resource to facilitate rubric and assignment development.

This research is situated well within a line of future research endeavors. First, this qualitative analysis of comments made regarding the rubrics and training barely scratch the surface of the data obtained through the think-aloud interviews. A future study will provide a more quantitative analysis of the way UTAs made grading decisions, framed by similar decisions made by multiple faculty, to identify technical aspects of rubrics that contribute to the greatest amount of variability. Further, additional data has been collected that will allow for a comparison of reliability of grading in the semester prior to the implementation of training versus that of the semesters that used training. In the future, this data will be used to further modify grading and training procedures, and data will continue to be collected and analyzed.

References

[1] ABET, "Criteria for accrediting engineering programs, 2017 - 2018," 2016.

[2] G. W. Clough, "The engineer of 2020: Visions of engineering in the new century," Washington, DC, USA, 2004.

[3] P. E. Dickson, T. Dragon, and A. Lee, "Using undergraduate teaching assistants in small classes," *Proc. 2017 ACM SIGCSE Tech. Symp. Comput. Sci. Educ.*, pp. 165–170, 2017.

[4] S. Ashton and R. S. Davies, "Using scaffolded rubrics to improve peer assessment in a MOOC writing course," *Distance Educ.*, vol. 36, no. 3, pp. 312–334, 2015.

[5] E. J. Hansen, *Idea-based learning: A course design process to promote conceptual understanding*. Sterling, VA, USA: Stylus Publishing, LLC., 2011.

[6] J. R. Betts and R. M. Costrell, "Incentives and equity under learning objective based reform," *Brookings Pap. Educ. Policy*, vol. 2001, no. 1, pp. 9–74, 2001.

[7] M. A. Muñoz and T. R. Guskey, "Learning objective grading and reporting will improve education," *Phi Delta Kappan*, vol. 96, no. 7, pp. 64–68, Apr. 2015.

[8] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," *Educ. Res. Rev.*, vol. 2, no. 2, pp. 130–144, 2007.

[9] V. Crisp, "Judging the grade: Exploring the judgement processes involved in examination grading decisions," *Eval. Res. Educ.*, vol. 23, no. 1, pp. 19–35, 2010.

[10] W. B. Armstrong, "The association among student success in courses, placement test scores, student background data, and instructor grading practices," *Community Coll. J. Res. Pract.*, vol. 24, no. 8, pp. 681–695, 2000.

[11] N. M. Hicks and H. A. Diefes-Dux, "Grader consistency using learning objective based rubrics," in *The 124th ASEE Annual Conference & Exposition*, 2017.

[12] M. A. Stellmack, Y. L. Konheim-Kalkstein, J. E. Manor, A. R. Massey, and J. A. P. Schmitz, "An assessment of reliability and validity of a rubric for grading APA-style introductions," *Teach. Psychol.*, vol. 36, no. 2, pp. 102–107, 2009.

[13] P. A. Griswold, "Beliefs and influences about grading elicited from student performance sketches," *Educ. Assess.*, vol. 1, no. 4, pp. 311–328, 2010.

[14] H. G. Andrade, "Using rubrics to promote thinking and learning," *Educational Leadership*, vol. 57, no. 5, pp. 13–18, Feb-2000.

[15] G. L. Goldberg, "Revising an engineering design rubric: A case study illustrating principles and practices to ensure technical quality of rubrics," *Pract. Assessment, Res. Eval.*, vol. 19, no. 8, 2014.

[16] W. J. Popham, "What's wrong—and what's right—with rubrics," *Educational Leadership*, vol. 55, pp. 72–75, Oct-1997.

[17] B. M. Moskal, "Recommendations for developing classroom performance assessments and scoring rubrics," *Pract. Assessment, Res. Eval.*, vol. 8, no. 14, 2003.

[18] Y. M. Reddy and H. Andrade, "A review of rubric use in higher education," *Assess. Eval. High. Educ.*, vol. 35, no. 4, pp. 435–448, 2010.

[19] R. W. Cooksey, P. Freebody, and C. Wyatt-Smith, "Assessment as judgment-in-context: Analysing how teachers evaluate students' writing," *Educ. Res. Eval.*, vol. 13, no. 5, pp. 401–434, 2007.

[20] K. Adamson, P. Gubrud, S. Sideras, and K. Lasater, "Assessing the reliability, validity, and use of the Lasater Clinical Judgment Rubric: Three approaches," *J. Nurs. Educ.*, vol. 51, no. 2, pp. 66–73, 2012.

[21] M. Meadows and L. Billington, "A review of the literature on marking reliability," *Rep. Natl. Assess. Agency by AQA Cent. Educ. Res. Policy*, 2005.

[22] A. Melvin and L. Bullard, "Tips from the trenches: Preparation and implementation of an experience-based TA training session," in *115th ASEE Annual Conference & Exposition*, 2008.

[23] Y. Cho, S. Sohoni, and D. P. French, "Need assessment for graduate teaching assistant training: Identifying important but under-prepared roles," in *2010 Midwest Section Conference of the American Society for Engineering Educaiton*, 2010.

[24] K. M. Kecskemety, A. H. Theiss, and R. L. Kajfez, "Enhancing TA grading of technical writing: A look back to better understand the future," in *122nd ASEE Annual Conference & Exposition*, 2015.

[25] J. P. Kurdziel, J. A. Turner, J. A. Luft, and G. H. Roehrig, "Graduate teaching assistants and inquiry-based instruction Implications for graduate teaching assistant training," *J. Chem. Educ.*, vol. 80, no. 10, pp. 1206–1210, 2003.

[26] M. Komarraju, "A social-cognitive approach to training teaching assistants," *Teach. Psychol.*, vol. 35, no. 4, pp. 327–334, 2008.

[27] R. Essick, M. West, M. Silva, G. L. Herman, and E. Mercier, "Scaling-up collaborative learning for large introductory courses using active learning spaces, TA training, and computerized team management," in *123rd ASEE Annual Conference & Exposition*, 2016.

[28] L. Long, A. Snyder, R. Stech, B. Jelen, C. Allison, and J. Merrill, "First-year engineering program: Student instructional leadership team - expanded and testructured," 2013.

[29] F. Marbouti, K. J. Rodgers, H. Jung, A. Moon, and H. A. Diefes-Dux, "Factors that help and hinder teaching assistants ' ability to execute their responsibilities," in *120th ASEE Annual Conference & Exposition*, 2013.

[30] M. A. Verleger and H. A. Diefes-Dux, "A teaching assistant training protocol for improving feedback on open-ended engineering problems in large classes," in *120th ASEE Annual Conference & Exposition*, 2013, p. 23.121.1-23.121.12.

[31] S. C. Roberts, K. A. Hollar, and V. M. Carlson, "Looking back: Lessons learned from ten years of training teaching assistants," in *1997 ASEE Annual Conference*, 1997.

[32] P. Lather, "Paradigm proliferation as a good thing to think with: Teaching research in education as a wild profusion," *Int. J. Qual. Stud. Educ.*, vol. 19, no. 1, pp. 35–57, 2006.

[33] M. T. Boren and J. Ramey, "Thinking aloud: Reconciling theory and practice," *IEEE Trans. Prof. Commun.*, vol. 43, no. 3, pp. 261–278, 2000.

[34] H. J. Rubin and I. S. Rubin, *Qualitative interviewing: The art of hearing data*, 3rd ed. Thousand Oaks, CA, USA: Sage Publications, 2012.

Acknowledgements

We would like to thank Nikhil Dingra for assisting with qualitative analysis and the many undergraduate teaching assistants who participated in the study. We would also like to thank the first-year engineering program for their support and allowing us to study their materials and procedures.