ASEE's 123rd Annual • Conference & Exposition • New Orleans, LA • June 26-29, 2016

New Orleans JAZZED about Engineering Education

Paper ID #15381

A Standards-based Assessment Strategy for Written Exams

Dr. J. Blake Hylton, Ohio Northern University

Dr. Hylton is an Assistant Professor of Mechanical Engineering at Ohio Northern University. He previously completed his graduate studies in Mechanical Engineering at Purdue University, where he conducted research in both the School of Mechanical Engineering and the School of Engineering Education. Prior to Purdue, he completed his undergraduate work at the University of Tulsa, also in Mechanical Engineering. He currently teaches first-year engineering courses as well as various courses in Mechanical Engineering, primarily in the mechanics area. His pedagogical research areas include standards-based assessment and curriculum design, the later currently focused on incorporating entrepreneurial thinking into the engineering curriculum.

Prof. Heidi A. Diefes-Dux, Purdue University, West Lafayette

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. She is currently a member of the educational team for the Network for Computational Nanotechnology (NCN).

A Standards-Based Assessment Strategy for Written Exams

Introduction

Grading and assessment in higher education has been an on-going point of professional and scholarly discussion. In the traditional model of assessment, a summative approach is followed. Under such a system, a series of assignments are each assigned a score and then summed in accordance with some system of weighting before arriving at a final letter grade. Ultimately, however, the question must be asked: What is the conceptual value of a point under such a system? On what basis has a student who has collected a certain percentage of available points demonstrated his or her mastery of a topic? In fact, the value of a point is a largely arbitrary construction which fluctuates wildly from institution to institution, course to course, or even semester to semester under the same instructor. It is not difficult to imagine a scenario in which a single assignment might be graded by several instructors with different interpretations of how to assign scores, meaning that the assessed learning tied to a given point may even fluctuate.

As the latest pedagogical trends have shifted in the direction of a more holistic, experiential approach to education through methods such as project-based and active learning, the education community has sought alternative ways to assess student learning in these systems. The challenges faced by such a reform are formidable, not least of which being a pervasive mindset that the primary function of grading is differentiating between students, rather than assessing a particular student's achievement or competency.[4] However, there is a building momentum for change, as researchers and practitioners have begun to question the ability of a traditional, summative grading scheme to adequately assess student understanding and communicate student learning to students and instructors. Towards this goal, there has been an increased focus on the viability and efficacy of standards-based assessment, with a number of different interpretations of that term appearing in the literature.[11]

Standards-based assessment is an alternative to the traditional score-based grading approach, by which student assessment is conducted directly on identified course learning objectives. Students are assessed repeatedly on their achievement on these objectives while also being provided with clear, meaningful feedback on their progress.[3,12] The terminology and design-thinking behind standards- or criteria-based grading has grown tremendously in the past decade, and Sadler, Ziegenfuss, and Muñoz provide worthwhile reviews for the interested reader.[7,11,13] Ultimately, however, the overriding conclusion is that a standards-based approach is the clear next step in assessment methodologies.

Additionally, standards-based assessment has been noted as offering a number of clear advantages over the traditional approach, including personalized and meaningful feedback, clear connections between assessment and stated course objectives, and transparency in the grading process.[2,11] The benefits of such a system are only just being studied in detail, but positive impacts have been observed in both affective and cognitive behaviors, including an increase in self-efficacy and a sophistication of epistemological beliefs.[3] Researchers have observed these benefits at both large public institutions and small private colleges and the observed improvements appear to be independent of overall student performance, meaning that the observed affect for high performing students was comparable to that observed for low performing students.[1,5]

Variations of a standards-based approach have been previously applied to project-based courses and even to more traditional homework type assessments.[8,12] With Post being a notable example, most of the published case studies have as a central component of the grading infrastructure a system of well-developed rubrics.[10] Jonsson presents an argument for rubrics as a more reliable and targeted assessment tool with great potential to promote learning and reflection, making them a natural pair for standards-based grading.[6] While the applications and structure of rubrics can vary greatly across the literature, a rubric in this context includes criteria for rating student performance as well as standards for attainment of those criteria. Rubrics of this variety may be *holistic*, meaning that they include a single rating scale for the entirety of the work, or *analytical*, meaning that several scales are used to assess different dimensions of the work. Perlman offers a valuable discussion of the thought-process that goes into developing a successful rubric, as well as the different varieties which may be applied.[9]

In this work, a system of analytical rubrics were applied to traditional written exams, conducted in the context of a course assessed almost entirely through standards-based grading. We look at the nature of the exams and standards-based rubrics and how they were implemented, communicate the lessons learned, and demonstrate how the other standards-based graded elements of the course interact with the exam to provide a more complete picture of student achievement.

Purpose, Motivations, and Setting

The Purdue University First-Year Engineering (FYE) program is a foundational course series undertaken by all beginning students seeking entry into an engineering program. The FYE program, which sees an annual throughput in excess of 1500 students, is continuously seeking ways to improve the efficiency and efficacy of student assessment. Beginning in Spring 2013, all homework assessment activities were migrated to a rubric based evaluation approach, grounded in the course learning objectives. The move was motivated by a number of factors, including paperless submission and grading of assignments which necessitates clear communication of performance in the absence of writing on students' papers, transparency and perceived fairness by the student population, as well as a desire to leverage the ability to better connect assessment activities with the course and program outcomes. It was observed that student regrade requests, inquiries about minor point deductions, and other such concerns were greatly reduced upon introduction of the standards-based system.

The focus of this paper is on the second course in engineering course sequence, ENGR 132: Ideas to Innovations II, in Spring 2015. This is a required second semester, 2-credit hour course for all FYE students. In this course, students learn how to use computer tools (i.e., Excel® and MATLAB®) to solve fundamental engineering problems, learn how to make evidence-based engineering decisions, develop problem-solving, modeling, and design skills, and develop teaming and communication skills.

During the 2015 spring semester, an effort was undertaken to also transition the traditional written exams to a standards-based assessment approach and to connect them directly with the non-exam assessments. The motivation for this change was to continue the transformation of the course to an entirely standards-based approach and to better connect the exams to course outcomes. The intention was to design the exams such that they retained a similar structure and content distribution, but to wholly convert the evaluation component to a rubric-based system. It should be noted that all sections of ENGR 132 take three common one-hour exams consisting of short answer, multiple choice, and code-generation or tracking questions. Due to the size of the course, each section is graded separately by the assigned team of undergraduate and graduate teaching assistants as well as the instructor of record for the section. Typically, in an effort to make grading as consistent as possible, any given grader will be assigned one or more questions, which they will then grade for all students in their section. This approach did lead to some differences of interpretation between sections when a traditional point system was used to grade an exam, lending further motivation for the move to a rubric-based approach. While the basic system of section-by-section grading was retained, the use of a common and explicit rubric sought to normalize the grading across the course.

Process and Outcomes

The exam writing team consisted of four lead instructors, two of which were consistent across all exams and two of which were drawn from the instructor pool. Prior to writing any exam material, the instructors compiled a list of all learning objectives which were covered in the preceding weeks of the course and were appropriate for assessment on a written exam. The list of learning objectives was used as a guide for writing exam questions. Selected examples of learning objectives used in this study are listed in

Table *1*. In this table, the left hand column includes the larger course objectives while the right hand column includes the specific objectives used to guide the focus of the exam questions.

Once a list of target learning objectives was compiled, the instructors divided the topics and began to develop questions to assess one or more of the specific objectives. Certain objectives, such as “Manage text output” appeared across multiple questions while others, such as “Create an x-y plot suitable for technical presentation” appeared only once. Each question was developed with three components – question, solution, and rubric. Rubric items enabled assessment of learning objectives on between 2 and 4 achievement levels (i.e., no evidence, underachieved, partially achieved, fully achieved), depending on the complexity of the objective being assessed. Because exams were to be graded by a wide range of individuals, including undergraduate teaching assistants, the rubrics were written to be explicit in terms of what constituted evidence of each achievement level.

Table 1 - Selected Learning Objectives for MATLAB

Larger Learning Objective	**This means you can: (Detailed Learning Objective)**
Perform and evaluate algebraic and trigonometric operations	Perform algebraic computations with scalars
	Employ order of operations to perform calculations
	Use built-in functions to perform algebraic and trigonometric calculations
	Perform element-by-element operations with vectors and scalars
	Perform element-by-element operations with vectors and vectors
	Perform element-by-element operations with matrices
Import data from electronic files	Import numeric data stored in .csv, .dat, and .txt files
Create and evaluate a x-y plot suitable for technical presentation	Create a x-y plot from a single data set
	Create multiple plots in separate figure windows
	Create a x-y plot with multiple data sets in a single figure window
	Create multiple plots in a single figure window
	Format plots for technical presentation
	Close figure windows
Manipulate arrays	Convert a row vector to a column vector (or vice versa)
	Extract a single element from an array (vector or matrix)
	Extract an array from a matrix
	Concatenate arrays
	Replace elements of arrays

For example, a simple question targeted the objective "*Use relational and logical operators to test 'between' logic"* (in MATLAB) and resulted in the question below:

Write the correct logical expression to check if variable X is between 1 and 6. Solution: X > 1 & X < 6

This question was evaluated using the single line, two-tier achievement level rubric item as shown below. It is important to note that, unlike pedagogical systems which have converted wholly to an objective based assessment approach, this course still utilizes point values to aggregate scores and assign a letter grade. This is largely an artifact of the on-going transition to a standards-based approach, as the course is still being migrated to the new system. It should also be observed that there is no inherent flaw in using a points-based system. The challenge of such an approach is that the points themselves are often not tied to any learning outcome that indicates clearly what learning or skill development is being assessed and, as a result, may have a fluctuating or arbitrary value. In the sense that points are applied here, they are less a raw indication of student success and more a means by which the instructors can indicate and account for the relative importance of various topics or assessment activities.

Objective: Use relational and logical operators to test 'between' logic	
No Evidence Score: 0 pts	Fully Achieved Score: 3 pts
☐ Answer is missing ☐ Any answer not equivalent to the correct answer	☐ X > 1 & X < 6 ☐ (X > 1) & (X < 6)

A second example question targeted the objective "*Create and interpret repetition structures*" and was valued at 8 points. This question was evaluated by a three-tiered achievement level rubric item, allowing for partial credit for errors when there is still some demonstrated level of understanding of the topic.

A MATLAB program uses the following loop to iteratively perform a simple calculation.

```
count = 0;
x = 1;
while x<15
        count = count + 1;
        x = x*2;
end
fprintf('The value of x is %.0f after %.0f iterations.',x,count)
```

What values of x and count will be displayed by the fprintf statement?

A. x = ________________ ANS: 16
B. count = _____________ ANS: 4

Objective: Create and interpret repetition structures		
No Evidence Score: 0 pts	Partially Achieved Score: 4 pts	Fully Achieved Score: 8 pts
☐ x neither 8 nor 16 ☐ count neither 3 nor 4	☐ Answers are for previous iteration (x=8, count=3) ☐ Either x = 16 OR count = 4, but not both	☐ x = 16, count = 4

On the other end of the complexity spectrum, some questions involved a much greater degree of effort, both in terms of students completing the problem and instructors developing the rubric. The question shown below tested two objectives: "*Create and evaluate a x-y plot suitable for technical presentation*" and "*Manage text output*". It was evaluated using two rubric items. In this case, both of the rubric items were evaluated on four tiers of achievement.

The input and calculations section of a MATLAB script are shown below. In the Outputs section, write the necessary MATLAB code to complete the following actions:

A. On the same plot, plot the populations of Rabbits versus Time and Wolves versus Time. Plot the rabbit population using a solid blue line and circles as data markers. Plot the wolf population using a dashed red line and x's as data markers. Include appropriate axis labels and a title.

B. Use the fprintf command to output the maximum rabbit and maximum wolf populations. Format the output with no digits after the decimal place. Be sure to include appropriate text in the fprintf command, don't just output a number devoid of context! *(Note: These maximum values are already calculated in the CALCULATIONS section of the script)*

```
%% INPUTS
Time = [0:17];  % Months
Rabbits = [27 25 44 77 96 124 176 244 297 341 352 331 249 155 51 17 5 0];
Wolves =  [7 7 9 13 13 13 13 17 19 28 35 43 45 63 64 65 64 60];

%% CALCULATIONS
% Calculate maximum rabbit population and maximum wolf population
maxRabbits = max(Rabbits);
maxWolves = max(Wolves);

%% OUPUTS
% Plot populations on same graph
plot(Time,Rabbits,'bo-');
hold on;
plot(Time,Wolves,'rx--');
title('Wolf and Rabbit Population Statistics');
ylabel('Population');
xlabel('Time [Months]');

ALTNERNATE:

plot(Time,Rabbits,'bo-',Time,Wolves,'rx--');

% Display peak values
fprintf('Peak rabbit population was %.0f.\n',maxRabbits);
fprintf('Peak wolf population was %.0f.\n',maxWolves);

SINGLE LINE ALTERNATE:

fprintf('Peak rabbit population was %.0f. Peak wolf population was %.0f.\n',
maxRabbits,maxWolves);
```

Objective: Create and evaluate a x-y plot suitable for technical presentation

No Evidence Score: 0 pts	Underachieved Score: 5 pts	Partially Achieved Score: 10 pts	Fully Achieved Score: 15 pts
☐ Plot command missing ☐ More than one error from: ○ Hold is missing ○ Data not entered into vectors ○ Time vector missing	☐ Plot command used ☐ Only one error from: ○ Hold is missing ○ Data not entered into vectors ○ Time vector missing ☐ **All** these errors: ○ Time is dependent variable ○ Incorrect or missing formatting codes ○ Title missing / before plot command / not acceptably descriptive ○ Axis labels missing / before plot command / reversed	☐ Plot command used ☐ Hold is used ☐ Data entered into vectors ☐ Up to three errors from: ○ Time is dependent variable ○ Incorrect or missing formatting codes ○ Title missing / before plot command / not acceptably descriptive ○ Axis labels missing / before plot command / reversed	☐ Plot command used ☐ Hold is used ☐ Data entered into vectors ☐ Time is independent variable ☐ Both plots formatted correctly ☐ Appropriate title ☐ Appropriate axis labels

Objective: Manage text output

No Evidence Score: 0 pts	Underachieved Score: 2 pts	Partially Achieved Score: 4 pts	Fully Achieved Score: 5 pts
☐ Missing or syntax error in fprintf commands OR ☐ All of these errors: ○ Contextual text missing ○ Incorrect formatting code ○ Missing \n (new line)	☐ No syntax error in fprintf commands ☐ Variable is hard-coded in fprintf, rather than referencing variable ☐ No more than two of these errors: ○ Contextual text missing ○ Incorrect formatting code ○ Missing \n (new line)	☐ No syntax error in fprintf commands ☐ One of these errors: ○ Contextual text missing ○ Incorrect formatting code ○ Missing \n (new line)	☐ No syntax error in fprintf commands ☐ Appropriate contextual text is included in fprintf ☐ Correct formatting codes used ☐ \n (new line) included

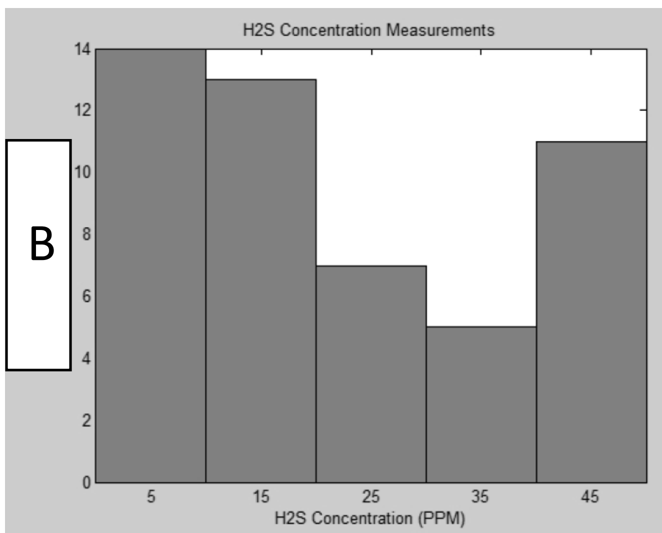
This particular example, taken from the first of the three exams, is a useful discussion point as it also exemplifies what the authors came to realize is a poor practice. Instructors observed the first rubric item to be particularly harsh in terms of penalizing students for minor mistakes. That single rubric item is worth 15 points, a drop of one tier constitutes a significant reduction in points. This can be seen in the overall performance on the question (Table 2).

Table 2 – Percentage of students assessed at each achievement level on the learning objective *"Create and evaluate a x-y plot suitable for technical presentation."*

No Evidence	Underachieved	Partially Achieved	Fully Achieved
5%	18%	51%	26%

Going forward, a different strategy for writing questions was employed so as to avoid having singular high-value rubric items and instead focusing on questions with multiple smaller components which could be assessed with greater fidelity. For example, the question shown below, taken from the third exam, tested a series of objectives, all related to the creation and interpretation of histograms. It was similarly valued at 20 points, but was evaluated using four rubric items, rather than two. As such, no single item was seen to dominate the score.

You have taken 50 random measurements of hydrogen sulfide (H_2S) levels in an industrial chemical process, measured in the range of 0-50 parts per million (PPM). Your data is in a MATLAB row vector named h2s_ppm.



A. (6 points) Write a single line of MATLAB code needed to generate the histogram shown above. NOTE: For this question you do NOT need to title or label the histogram.

hist(h2s_ppm,5)

B. (4 points) Write a single line of MATLAB code to generate an appropriate label for the histogram's y-axis (indicated as "B" in the figure below).

ylabel('Frequency')

C. (4 points) On the answer sheet, circle the bin (counting from left to right, 1 to 5) which would include a measurement of 30 PPM?

4 (In MATLAB, left edges are inclusive, except in the last bin which includes both edges)

D. (6 points) Write a single line of MATLAB code to save the bin centers and bin counts for this histogram into variables named h2s_centers and h2s_counts, respectively.

[h2s_counts, h2s_centers] = hist(h2s_ppm)

A) *Objective: Create a histogram with a specified number of bins*			
No Evidence Score: 0 pts	Underachieved Score: 2 pts	Partially Achieved Score: 4 pts	Fully Achieved Score: 6 pts
☐ None from fully achieved column	☐ Only one from fully achieved column	☐ Only two from fully achieved column	☐ Use MATLAB hist command with right syntax ☐ Specify h2s_ppm as input variable ☐ Specify 5 as second hist parameter (number of bins)

B) *Objective: Format histograms for technical presentation*			
No Evidence Score: 0 pts	Underachieved Score: 1 pts	Partially Achieved Score: 3 pts	Fully Achieved Score: 4 pts
☐ None from fully achieved column	☐ Only one from fully achieved column	☐ Only two from fully achieved column	☐ Use appropriate term for y-axis label (e.g., count, frequency, number of measurements) ☐ Use MATLAB command ylabel ☐ Use correct syntax for command ylabel

C) *Objective: Interpret histograms*		
No Evidence Score: 0 pts	Partially Achieved Score: 2 pts	Fully Achieved Score: 4 pts
☐ Answer missing OR ☐ Selected bin other than 3 or 4	☐ Selected Bin 3	☐ Selected Bin 4

D) *Objective: Compute the frequency of the data in each bin of a histogram*			
No Evidence Score: 0 pts	Underachieved Score: 2 pts	Partially Achieved Score: 4 pts	Fully Achieved Score: 6 pts
☐ None from fully achieved column	☐ Only one from fully achieved column	☐ Only two from fully achieved column	☐ Syntax to capture two outputs [] ☐ Correct variable names used ☐ Variables names in correct order [h2s_counts, h2s_centers]

It should be noted that an alternative strategy to this "step-through" approach would have been to break the concepts into smaller individual problems, unrelated to one another. This approach may even be preferable, both in simplifying the questions from a student perspective and eliminating any chance of carry-over errors and also in reducing the complexity of the associated rubrics.

For any given exam, once all of the exam questions, solutions, and rubric items were written, the point values were balanced to ensure appropriate relative weighting of the problems based on complexity and estimated time and effort on the part of the student. Any large learning objectives which were felt to be lacking in assessment resulted in additional questions in that area. Once completed, the exams were tested by undergraduate teaching assistants, to evaluate the difficulty level to get an estimate of the time to complete the exam, and to identify any questions in need of clarification. If needed, the exams were revised to clarify directions and meaning and to add or remove (as was more often the case) questions or question components.

After the exam was administered, a small subset of 10-20 randomly selected exams were graded using the rubric by the exam development team. This step was found to be critical in identifying problems with the rubrics. Issues observed included unexpected student errors not previously accounted for, areas where the rubric seemed to apply more or less harshly than intended based on the combination of errors or a lucky "shotgun" answer, or areas where the rubric was confusing or difficult to apply. Once changes were made, the rubric was distributed to the rest of the instructors.

To reduce grading complexity, the exams were constructed in two parts – a questions booklet and an answer sheet. Only the answer sheets were graded. To facilitate using the rubric during grading, the answer sheets included scoring blocks for each question. For each question, a scoring block contained the same number of rows as associated rubric items, the possible achievement levels, and the points for those levels. An example is shown in Figure 1, depicting the answer sheet section related to the histogram example problem presented above. Here there were four rubric items being assessed with three being assessed using four achievement levels and one being assessed using three achievement levels. Graders circled the correct achievement level for each rubric item on the answer sheet as shown in Figure 1. This approach enabled the graders to print only a single copy of the rubric and conduct rubric-based grading directly on the answer sheets, rather than on separate and lengthy rubric pages. Generally, the majority of exam questions were graded by undergraduate teaching assistants. The more complex and involved questions, as well as those involving a complex assessment rubric, were set aside to be graded by the instructor or graduate teaching assistant.

To communicate the scores to students, the rubrics were recreated electronically on Blackboard®. Rubric scores were then entered into this form, allowing students to access and view their achievement on each individual rubric item. This approach is central to the standards-based grading method deployed throughout the course and provides a much greater transparency of scoring as well as enabling students to more clearly identify areas of achievement and misunderstanding.

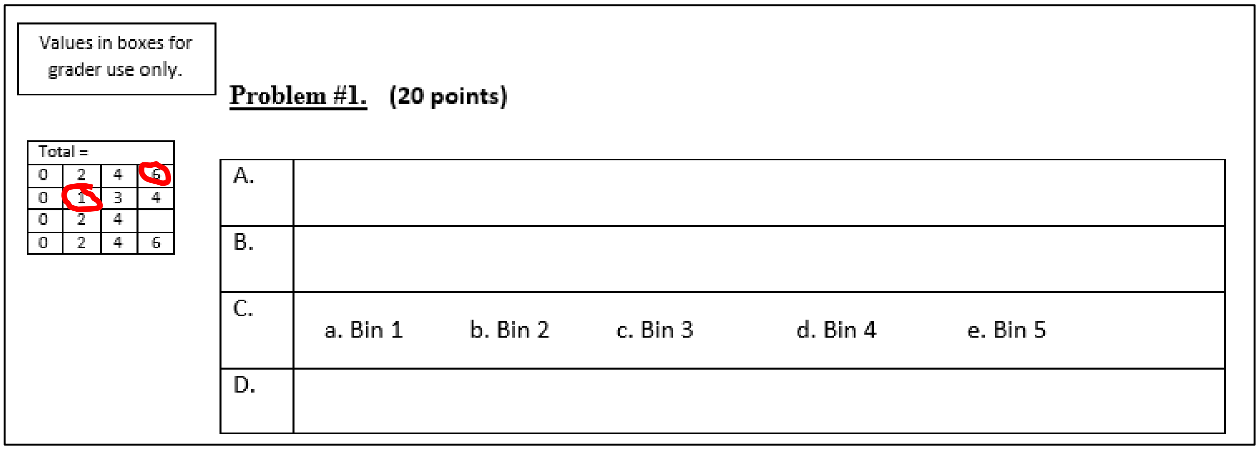


Figure 1 – Sample answer sheet section

A further benefit of this approach is that, because each section is created individually on Blackboard, instructors are able to retrieve rubric-item level reports on achievement within their section. This allows more directed follow-up instruction in particularly low performance areas as well as better instructor self-reflection and evaluation of teaching methods. Additionally, course lead instructors can, with some assistance from Blackboard support staff, extract course-wide rubric item scores. This allows a larger course-wide study of student achievement at the large objective level, either for instructional or institutional assessment purposes.

Discussion

Feedback from Students

To evaluate student perceptions of the standards-based approach to exam assessment, relevant questions were included in a larger end-of-semester survey administered via Qualtrics, a web-based survey software tool (http://www.qualtrics.com/). Survey questions were randomly assigned to students. In total, 145 students (17% of the population) received the exam related questions. In one question, students were asked to rate five items about their exam experiences using seven-point bimodal scale. The scale listed two extremes (e.g. 1 = “Exams were too easy” and 7 = “Exams were too difficult”) and students were asked to mark their opinion. Students were directed to use the middle point (four) as “just right”. The five items assessed were:

- Difficulty (The exams were too easy – The exams were too difficult)
- Length (The exams were too short – The exams were too long)
- Content (The exams had too few questions – The exams had too many questions)
- Fairness (Grading of the exams was too lenient – Grading of the exams was too harsh)
- Representativeness (My exam scores overestimated my actual knowledge – My exam scores underestimated my actual knowledge)

In addition, students were asked, in an open text box, to comment about any aspect of the exams. Overall, students tended to select the “just right” level on most of the scaled questions.

A statistical analysis was conducted on the survey results. For each item, a one sample two-tailed t-test was conducted using the null hypothesis of a mean score equal to 4 ("just right"). 95% confidence intervals were also determined. The results of this analysis are presented in Table 3. A lean towards the higher ("too harsh") end of the scale was observed across all items and found to be significant for all except the question on difficulty level. The greatest deviation from the mid-level score was observed for the fairness of the exams question, which had a median score of 4.92 out of 7 (P = 0.0001) and shows a definite right skew (Figure 2). This indicates that, while students generally were satisfied with the exams, they felt that the grading was at times too harsh. This is not an unexpected result, given the same observations made by the exam team as discussed above. One limitation of this analysis is that there is no direct comparison to traditional exams. It would not be unexpected to see similar perceptions expressed in a traditional exam environment, but a direct comparison cannot be made without that additional data set.

Table 3 – Analysis of group scores on a seven point scale
(1 being left hand side of continuum, 7 being right hand side of continuum)

Survey Item (continuum)	Group median	Standard Dev	P
Difficulty (Too easy – Too difficult)	4.15	1.19	0.0962
Length (Too short – Too long)	4.23	1.09	0.0004
Content (Too few – Too many)	4.17	1.05	0.0001
Fairness (Too lenient – Too Harsh)	4.92	1.29	0.0001
Representativeness (Overestimated – Underestimated)	4.28	1.26	0.0001

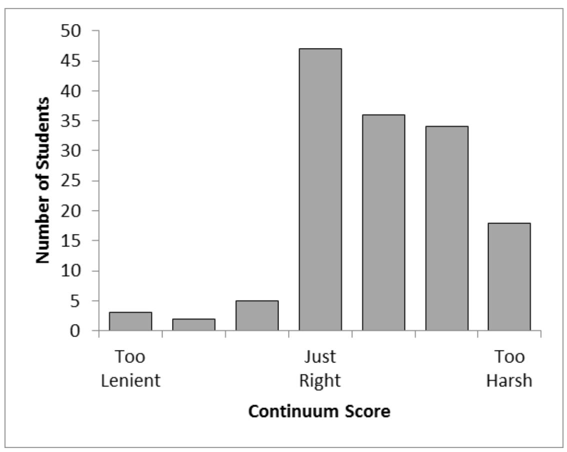


Figure 2 – Histogram of student responses on fairness of grading

To analyze the student open response comments, an emergent coding scheme was applied and responses clustered into twelve code groups. Approximately half of responding students provided only a generic positive comment (e.g. "Exams were fine") or no comment at all. Several code groups were linked to fewer than 2% of students and are not discussed further. Those codes which yielded a substantial number of responses and could provide insight into the implementation of standards-based assessment with written exams are outlined below.

Some students (9%) perceived inconsistencies in the way that rubrics were applied or that there were too many errors leading to regrades. This code group also preferred that questions be graded by the instructor rather than by a TA, in part due to a perception among the students that the instructors would either have more authority to grant partial credit, more consistency in grading, or possess the greater depth of understanding needed to identify when partial credit was warranted. There is likely some truth to these concerns, as the transition to the rubric-based approach was a learning curve for all involved. It is anticipated that more directed training of graders and experience on the part of the exam writing team in terms of rubric creation will largely mitigate these concerns in future semesters.

A group of students (5%) thought that the exams tested too many small, “nitpicky” items such as formatting or syntax. This concern is likely related to both of the above concerns. Ultimately, this is in large part a pedagogical question regarding how best to assess programming skill. It is also however a question of the rubrics themselves, albeit to a lesser extent. Further reflection is needed on the part of the course administration concerning the specific learning objectives - their specificity and whether and how they should be assessed on an exam.

Finally, 3% of students did not feel that the exams addressed the core objectives of the course. This was not unexpected, given that the written exams focused on the programming and statistics portions of the course, leaving the design and modeling objectives to be assessed in other ways. A key take-away from this objection is that perhaps the standards-based grading approach could be better utilized to help students see the larger picture of the course and to understand that exams are not the only means of assessment used to evaluate their learning.

Relation to Other Assessment Points

As this work was undertaken within the context of a course converted largely to a standards-based grading approach, it is possible to also examine how student achievement on the exams mapped to achievement on other assessment activities.

For example, assessing student work on both homework and exams using learning objectives allows us to see where learning is improving or unstable. Consider the case of the learning objective concerning understanding how to “*Use relational and logical operators to test whether x is between a and b*”. When comparing the relevant homework scores (Homework 4) to exam scores (Exam 1), we can see a clear improvement. It is relevant to note that the exam assessment occurred several weeks after the selected homework but without any direct instruction on the topic in the intervening time frame.

Table 4 – Homework vs exam comparison showing improvement on the learning objective “*Use relational and logical operators to test whether x is between a and b*”.

Assessment	No Evidence	Fully Achieved
Homework 4	21%	77%
Exam 1	10%	90%

In the case of learning objectives associated with learning how to code and track an indefinite loop, we can see that understanding is perhaps not as stable (Table 5). The surrounding context for this comparison is the same as for the previous case, albeit with a slightly shorter time delay between assessment points.

Table 5 – Homework vs exam comparison showing unstable student understanding

	Assessment	No Evidence	Underachieved	Partially Achieved	Fully Achieved
Code an indefinite looping structure	PS05	6%	3%	6%	85%
Create and interpret repetition structures [track while]	Exam 1	12%	0%	19%	68%

As there are clear limitations to this type of analysis of course results, interpretation of results needs to be carefully considered. Each learning objective assessment is attached to a single homework problem or exam problem, or potentially only a portion of an exam problem. Design flaws in the problem or the assessment and variability in graders' interpretation of the rubric item or problem solution can lead to an erroneous conclusion that the students are doing poorly with a particular learning objective. For example, a learning objective on assigning variables in MATLAB was assessed on the first exam, with the results for three sections shown in Table 6.

Table 6 – Achievement level comparison across three sections

Section	No Evidence	Underachieved	Partially Achieved	Fully Achieved
A	17%	0%	21%	63%
B	7%	0%	4%	89%
C	0%	0%	1%	99%

Why do the results for section A differ so greatly from those of the other two? For Section A, is there a grading issue or a teaching or learning issue? For Section C, do the graders not understand when evidence of achievement is present or not, and thus fail to appropriately use the rubric, or did the students in this section really understand this topic so much more clearly than those in other sections? These types of questions can only be answered after careful analysis of the surrounding context and with the support of other assessment points. Even then, the answers may not be clear. This case supports the argument that any assessment scheme must be combined in aggregate with multiple other assessment points to truly paint an accurate picture of student understanding. Standards-based grading provides the framework in which to accomplish this cross-assignment analysis at the objective level, but careful design of assessment opportunities is no less important than in any other grading scheme.

Conclusions

A method was presented by which traditional written exams may be assessed within a standards-based grading framework. The approach is predicated on the creation of detailed learning objectives able to be mapped back to larger course learning objectives. Exam questions are written to assess one or more of these objectives and then graded using detailed rubrics. Design strategies for the rubrics were also discussed. It was observed that a single high value rubric line produced poor grading resolution and was perceived poorly by the students. Instead, it was suggested that large problems be divided into smaller, multi-objective rubric lines. This approach produced higher grading resolution and greater satisfaction of both the instructors and the students. The logistics of creating, proof testing, grading, and reporting were also discussed.

In addition, it was demonstrated that student achievement on a given learning objective can be compared across multiple assessment points, including both homework and exams. This approach can reveal when student understanding is improving or when it is perhaps less stable. Limitations of this type of analysis were also discussed, as question development or grading inconsistencies can have a large effect on perceived achievement on a given problem.
Based upon the results of this study, student perceptions of fairness and difficulty of exams developed and graded from a standards-based assessment approach were largely in line with what one may expect from any well-design exam, either standards-based or traditionally assessed. Student concerns centered more on the content of the exams, being largely programming based, rather than on the grading method. Although there were certainly "growing pains" involved in the transition to a standards-based assessment strategy, it was viewed as largely positive by the instructional team. Additionally, it is believed that these perceptions could be positively impacted with more directed student instruction as to how to interpret and utilize the standards-based grading feedback. There was very little such instruction in this course, meaning that students often failed to effectively leverage the rubric feedback to guide their learning.

Acknowledgements

This work was made possible 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.

Bibliography

1. Atwood, S. A., Siniawski, M. T., & Carberry, A. R. (2014). Using standards-based grading to effectively assess project-based design courses. *Proceedings of the 121st ASEE Annual Conference & Exposition, Indianapolis, IN.*
2. Carberry, A., Siniawski, M., Atwood, S., & Diefes-Dux, H. A. (2016). Best practices for using standards-based grading in engineering courses. *Proceedings of the 123rd ASEE Annual Conference and Exposition, New Orleans, LA.*
3. Carberry, A. R., Siniawski, M. T., & Dionisio, J. D. N. (2012). Standards-based grading: Preliminary studies to quantify changes in affective and cognitive student behaviors. *Proceedings of the 42nd Annual Frontiers in Education Conference (FIE), Seattle, WA.*
4. Guskey, T. R. (2011, November). Five obstacles to grading reform. *Educational Leadership*, 17–21.
5. Heywood, J. (2014, October). The evolution of a criterion referenced system of grading for engineering science coursework. *Proceedings of the 44th Annual Frontiers in Education Conference (FIE), Madrid, Spain.*
6. Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics : Reliability, validity and educational consequences. *Educational Research Review*, *2*, 130–144.
7. Muñoz, B. M. A., & Guskey, T. R. (2015). Standards-based grading and reporting will improve education. *Kappan*, (April).
8. Parker, P. J., Bocher, B., & Polebitski, A. (2014). Assessing student writing competencies in environmental engineering courses. *Proceedings of the 121st ASEE Annual Conference & Exposition, Indianapolis, IN*.
9. Perlman, C. (2003). Performance assessment : Designing appropriate performance and scoring rubrics. In Wall, J. E., & Walz, G. R. *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators*. (pp. 497–506) ERIC Counseling and Student Services Clearinghouse, Greensboro, NC.

10. Post, S. L. (2014). Standards-Based Grading in a Fluid Mechanics Course. *Proceedings of the 121st ASEE Annual Conference & Exposition*, *Indianapolis, IN*.
11. Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, *30*(2), 175-194.
12. Siniawski, M., Carberry, A. & Dionisio, J. (2012). Standards-based grading: An alternative to score-based assessment. *Proceedings of the American Society for Engineering Education Pacific-Southwest Regional Conference*, *2012*.
13. Ziegenfuss, D.H. (2007). A phenomenographic analysis of course design in the academy. *Journal of Ethnographic & Qualitative Research, 2*, 70-79.