Standards-Based Grading Derived Data to Monitor Grading and Student Learning

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

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

Mr. Hossein Ebrahiminejad, Purdue University-Main Campus, West Lafayette (College of Engineering)

Hossein Ebrahiminejad is a graduate research assistant at SPHERE (Social Policy and Higher Education Research in Engineering) and a Ph.D. student in Engineering Education at Purdue University. He completed his M.S. in Biomedical Engineering at New Jersey Institute of Technology (NJIT) and his B.S. in Mechanical Engineering in Iran. His research interests include student pathways, Quantitative methods, educational policy, and relationships between education and professional practice.

Standards-Based Grading Derived Data to Monitor Grading and Student Learning

Abstract

Grading of student work is the primary practice for evaluating students' learning and performance in a course. As such, the data generated from grading can be a powerful source of evidence for course-level decision-making by stakeholders. This paper demonstrates, through a specific large engineering course example, how standards-based grading (SBG) derived data can be used to monitor student learning and grading. Three criteria for using SBG data confidently and effectively for this purpose are established. First, the grading data have to be of high quality. Second, the grading data results need to accessible through simple visual representations. Third, there needs to be a clear path forward from grading data, to interpretation, to actions.

Introduction

Grading of student work is the primary practice for evaluating students' learning and performance in a course. As such, the data generated from grading can be a powerful source of evidence for course-level decision-making by stakeholders. For grading data to be confidently and effectively used in decision-making that leads to actions, three criteria must be met. First, the grading data must be of high quality, generated through a grading system that attends to fairness, validity, fidelity, integrity, and reliability. Second, the grading data must be analyzed and presented in meaningful ways that enable easy interpretation by all stakeholders. Third, there must be a clear path for all stakeholders from data generation to interpretation of data to decision-making to actions. Here stakeholders refers, at a minimum, to students and their instructor. For large courses, stakeholders might also include the instructors for multiple sections of the course, graduate teaching assistants, undergraduate teaching assistants, and support staff.

The need for and the challenge associated with meeting each of the three criteria for confident and effect use of grading data for decision–making varies considerably from course to course. Certainly the nature of a course – its structure, content focus, teach strategies, types of assignments, etc. – influences the design of the grading system and the analysis and presentation of results to stakeholders for interpretation. But it is course size that adds incredible complexity to meeting each criteria, making workable decision-making practices for a small class impractical at scale.

To highlight the differences between small and large courses, consider two fictitious engineering instructors, Riley and Quinn. They both believe that grades should, among other things, provide an evaluation of student work that is valid, fair, and trustworthy, motivate and focus students' actions to learn, and promote data-driven student and instructor reflection [1]. For these instructors, grading is not about selecting talent - meaning the issuing of grades is primarily intended to differentiate students, but rather, grading is about developing the talent of all students through feedback. The notion that grading is feedback, and "feedback is teaching" [2] resonates with them.

Riley is the instructor of a small class and sole grader of student work. He quickly gets to know the students in his class. He is aware of their individual and collective performance through personal interactions in the classroom and office hours and the grading of their work. Irrespective of the rigor of the grading system itself and any analyzable data it might generate, he is able to, through his engagement with his students and their work, make decisions that impact what he will do in class this week to help his students' learn as well as what he will do to improve the course next time it is offered.

Quinn is an instructor in a multi-section course. She teaches one of many large sections. While active learning strategies are employed, the size of the section is such that it is difficult to know each student. The grading is performed by teaching assistants. Quinn can look at the gradebook for overall performance on an assignment and might even dig into a bit of student work, but the section size limits her ability to get a handle on the individual and collective performance at a level of detail that is actionable. She is also unable to assess whether her teaching assistants are grading reliably. Further, she has no access to the performance of students in other sections to gauge whether her section is doing on par, better, or worse than the course as a whole.

While Riley can use the intimacy of his small class to enable ad hoc decision-making processes, this is not always possible for Quinn. Quinn is considerably more removed from the details of her students' work and performance. If her course were to meet the three criteria for the confident and effective use of grading data, she and her instructional team would be better able to make evidence-based decisions. Plus, Quinn would be better equipped to enact her own teaching philosophy.

The purpose of this paper is to demonstrate how the three criteria can be met with standards-based grading derived data. The emphasis is on using grading data to monitor the quality of grading and student learning to inform change in large courses. It should be noted that the term "large course" is a bit nebulous here, as it does not mean to imply a particular number of enrolled students. Rather, it implies that implementation strategies are being employed that reduce an instructor's capacity to know, through traditional instructor course activity, individual and collective student performance details at a level that is actionable.

In this paper, the perspective of the three criteria is taken from the instructional team side of a specific course. The nature of standards-based grading data, its acquisition, and its analysis are described. The use of data visualizations to make meaning of the data to inform short and long terms actions is also described.

The perspective of the three criteria taken from the students' side is also very important but not addressed in this paper. With regards to the first criteria, researchers have looked at students' perspectives of standards-based grading (and similar) systems as compared to traditional grading systems [3] – [6]. This collection of work has essentially gotten at issues related to fairness. The second and third criteria have been explored in some initial work looking at students' access and use of standards-based grading feedback to improve their learning [6], [7].

Standards-Based Grading & Grading Quality

Standards-based grading (SBG) is a form of learning-objective based or criterion-referenced grading. SBG has been touted as being able to, among other things, provide meaning to grades, change grading practices, make grading practices transparent to stakeholders, help stakeholders understand what a quality demonstration of intended learning means, and incite instructional change and other educational reforms [8].

SBG, like all learning-objective based grading strategies, provides detailed insight into students' abilities because these strategies provide direct measures of student proficiency on well-defined course learning objectives [9], [10]. The course learning objectives explicitly state observable behaviors students should be able to demonstrate by the end of the course [11]. SBG employs rubrics, wherein the rubric items are some subset of the course learning objectives and each learning objective is assessed on a proficiency scale, for example, five levels from no evidence of proficiency to proficient. Tasks that are assessed using SBG are typically attempted only once by the students (i.e., there is no revise, resubmit, and regrade), though learning objectives may be assessed multiple times during the semester via different tasks that might increase in complexity.

SBG does not inherently ensure the fairness, validity, fidelity, integrity, and reliability of the grading system and its data. For a robust grading system, each of these aspects of a high quality grading system must be worked out and monitored over time to ensure high quality data.

Fairness is concerned with a number of issues pertaining to stakeholder perceptions, particularly students. First, grading is perceived to be fair if it is focused on an individual's work without comparison to other individuals' work (as with normative grading) [12]. This particular aspect of fairness is foundational to SBG; grades are issued based on an individual student's demonstration of their ability with each learning objective. Second, grading is perceived to be fair if students know when and how they are going to be judged [12]. This can be achieved by making a list of learning objectives accessible to students, preferably with an indication of the evidence that will be used to judge one's proficiency with each learning objective. This could mean going as far as sharing a generic (not problem-specific) rubric with students while they are working on an assignment. A third aspect of fairness is concerned with the perception of the consistency of the grading across a course and semesters [12]. Well-crafted and stable rubric items are necessary to achieve consistency.

The consistency of grading is inextricably tied to the other aspects of quality grading: validity, fidelity, integrity, and reliability. These aspects are all tied to the design of the grading system, and in the case of rubrics, the individual rubric items. Validity is about degree to which there is a match between the learning measurement and what it is supposed to measure. When an SBG rubric item is constructed, content expertise needs to be sought to both establish the demonstrable evidence of proficiency for each learning objective and design tasks that align with the learning objectives. There has to be a clear mapping between the tasks students are asked to do and the assessment of the tasks.

Fidelity, another aspect of quality grading, is about the degree to which a grade is purely representative of academic achievement versus non-achievement factors [12], [13] (e.g., effort).

As an SBG rubric item is constructed, the evidence of proficiency needs to be vetted for elements that are not directly tied to the observable demonstration of proficiency with the learning objective.

Integrity, still another aspect of quality grading, is about trustworthiness; it is concerned with the match between the proficiency level assigned to students' work and students' actual proficiency [12]. As an SBG rubric item is constructed, attention needs to be given to the extent to which some fraction or some degree of quality of the pieces of evidence for a learning objective truly indicates a particular level of proficiency.

Finally, reliability entails the degree to which multiple graders assign the same proficiency level to the same piece of student work. Reliability is a grading system implementation issue. A number of things can undermine reliability including grading fatigue, rushing to complete grading, and inattention to or misunderstanding of the rubric items. Given well designed rubrics, high reliability requires training and practice opportunities for and subsequent monitoring of those doing the grading [14].

A high level of confidence in the quality of grading data is required to use grading data in decision-making. SBG, with learning objectives as its backbone, enables the issues of fairness, validity, fidelity, integrity, and reliability to be attended in a cohesive fashion during the design of the system, including the SBG rubrics. The grading data resulting from SBG can then be analyzed and used more confidently to understand trends in student learning with respect to the learning objectives and monitor grader reliability.

An SBG Course

SBG was initiated in a large (N = 1500-1650) first-year engineering (FYE) course at a Mid-western U.S. university in Spring 2013 and has undergone revision since. The FYE course in this example was required for all engineering students. In this course, students learned how to use MATLAB to solve engineering problems as well as represent and model data. For the first two-thirds of the semester, students completed weekly problem sets. For the remaining third of the semester, the students completed weekly milestones associated with a team-based data analysis and modeling project. Students met twice a week for 110 minutes in a classroom designed for active learning.

The organizational structure of the course is shown in Figure 1. Course curators (three faculty) guided the development of the course materials and implementation strategy. The support staff handled the day-to-day operational aspects of the course (e.g., hiring and managing teaching assistants, finalizing course materials, managing the learning management system).

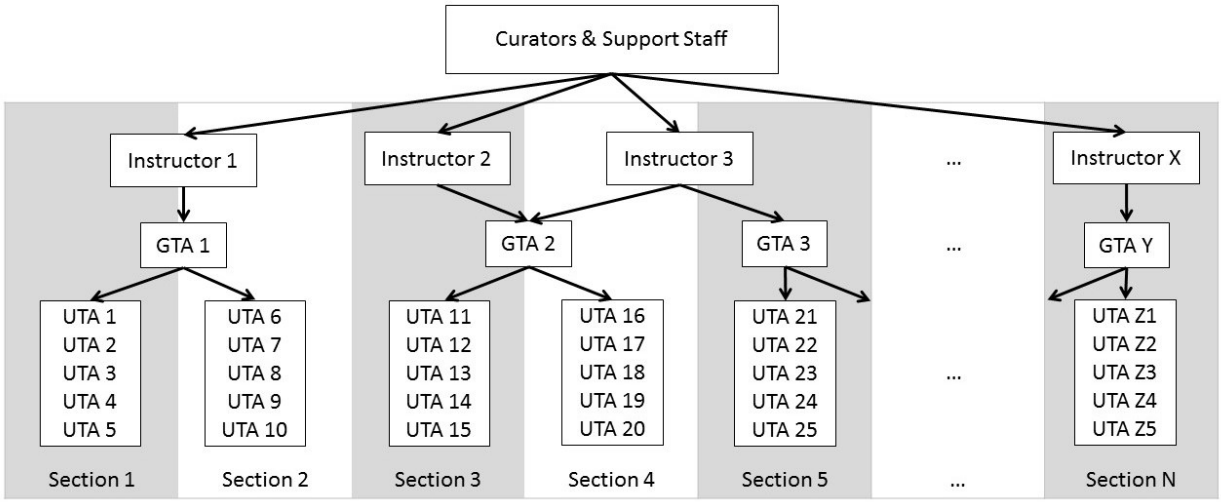

Figure 1. FYE course organizational structure.

The instructors (faculty and instructional staff) led the teaching of one to three sections of the course. A section size was 120 students. The sections used common materials (e.g. lecture slides, problems sets), had common exams, and stuck to a set schedule. The instructor of a section was supported by a graduate teaching assistant (GTA) and five undergraduate teaching assistants (UTAs). GTAs typically supported two sections, either for the same or different instructors. A GTA oversaw the UTAs in their sections; this included spot-checking the grading which was primarily completed by the UTAs. UTAs were of two types. Four were peer teachers (PTs); they attended class and supported active learning and graded student work. One was a dedicated grader; graders did not attend class but did take-on a larger share of the grading than each PT. For a given assignment, a PT likely graded 12-16 pieces of student work, while a grader graded 60 pieces. Information flow was from the curators and support staff to the instructors and GTAs, who all attended a weekly meeting. Information flow continued to the UTAs, who typically met with their instructor and/or GTA weekly. The support staff also provided the UTAs critical operational information directly.

This paper focuses on the Spring 2017 implementation of SBG to assess student work on problem sets. In Spring 2017, there were 14 sections with 10 instructors, 9 GTAs, and 69 UTAs.

Spring 2017 was the first semester in which grading data was analysed for the purpose of bringing results forward to the instructional team on a weekly basis during the semester for the purpose of immediate and long-term decision-making with regards to student learning and grader reliability. This was the first semester in which the data could be retrieved and analysed on a weekly basis without considerable cleaning and used confidently for such decision-making. Between Spring 2013 and Fall 2016, data analysis results were only presented to the instructional team after a given semester was over. In this period, much attention was given to easing the logistical issues of collecting clean data from the SGB rubrics and building confidence with the data by working towards fairness, validity, fidelity, integrity, and reliability. Confidence was built by clarifying the learning objectives and making them transparent to students, improving the alignment between problem sets and learning objectives, implementing various rubric designs for problem sets [14] and exams [15], and understanding the nature and extent of grader

reliability issues [14]. In addition, potential representations of the data were explored to prepare for bringing the data before the instructors on a more regular basis for decision-making.

Implementation of SBG

In Spring 2017, the FYE course had 19 overarching learning objectives, each with multiple sub learning objectives. Students' work on each problem set was assessed using between seven and ten sub learning objectives (heretofore just referred to as learning objectives). Detailed SBG rubrics guided the assessment of students' work with regards to each learning objective. An example of a rubric item for one learning objective, as seen by the instructional team and not the students, is provided in Table 1. Each rubric item stated the learning objective and provided a snippet of the solution that would demonstrate evidence of achievement of the learning objectives, a description of the general evidence of proficiency for the learning objective (in black text), and specific instructions for the problem being assessed (in blue text).

Table 1. Sample learning objective rubric

Item #8 Problem 3: Air Travel				
Learning Objective	07.01 Create an x-y plot from a single data set			
What to Grade:	PS02_airtravel_logins.docx > AIRFARE MODEL & DATA PLOT *Prob 3, Step 5d* Grade the plot of the data in Subplot C (airfare v fuel costs): NOTE: You need to find the location of this plot in the subplot figure – it may not be in the correct location subplot(2,2,2) % subplot 2,2,2 = subplot C, airfare v costs plot(cost,fares,'gs') grid on title('Plot C: Avg Airfare vs Fuel Price') xlabel('Fuel Cost (billion dollars)') ylabel('Airfare (dollars)')			
Proficient	**Developing**	**Emerging**	**Insufficient Evidence**	**No Attempt**
1 pt	0.8 pt	0.5 pt	0 pt	0 pt
Evidence items for proficiency: 1. Correct syntax for the plot command: plot(x, y, 'line/marker formatting') Use of the default plot format (i.e. leaving off the 'line/marker formatting', resulting in a blue line) is not correct. The formatting must be managed. 2. Correct identification of the independent (x) and dependent (y) variables X: fuel cost (variable name will differ by student; column 3 in csv data) Y: airfare (variable name will differ by student; column 7 in csv data) 3. Correct use of data markers and lines: data markers with no line (for raw data), line with no data markers (known model), data markers with overlaid line (for raw data with model) Any marker style and color are allowed as long as it's a scatter plot	1 (of 3) missing or incorrect item from the proficient list	2 (of 3) missing or incorrect item from the proficient list	3 (of 3) missing or incorrect item from the proficient list	Did not attempt the graded item

Depending on the number of pieces of proficiency demonstrated in the student work, the level of proficiency with each learning objective was indicated as Proficient, Developing, Emerging, Insufficient Evidence, or No Attempt. Written feedback was required on any learning objective that was not demonstrated at the Proficient level. The aim of written feedback was to explain the pieces of evidence that were missing or underdeveloped in the student work. This was necessary as the course learning management system did not enable checklists within a rubric.

Student work was submitted and assessed through Blackboard Learn™, the course learning management system. The Blackboard Learn rubric feature was used to communicate the assessed proficiency level and written feedback to the students. While students did not have access to the details of the rubric for any given assignment prior to submission, they did have access to a document that listed the course learning objectives and the general evidence of proficiency for each learning objective (i.e., the black text in the Proficient column of the rubric).

SBG Data & Analysis

Blackboard Learn, like all learning management systems, presents affordance and challenges when it comes to accessing, analysing, and visualizing rubric data. For the FYE course, each section had a separate Blackboard Learn site. Instructors and teaching assistants only had access to the section sites they taught. While, a summary of the results for a single problem set for a single section could be generated by the instructor with a push of a few buttons in Blackboard, there was no means by which to generate a summary report for the overall course. Further, there was no way to access through the Blackboard interface individual student data or individual grader data associated with specific rubric items (i.e., learning objectives). As a result, specific data collected by Blackboard during the process of grading via the rubrics had to be requested through the university's information technology unit.

From Spring 2013 through Fall 2016, there was some back and forth with the university's information technology unit to identify the data needed and come to an understanding of what certain available variables truly represent. These was also a considerable effort put into standardizing the FYE course rubrics entered into Blackboard from problem set to problem set and across sections so that clean data were generated (e.g., numbering of learning objectives).

Once the data to be downloaded was identified, the university's information technology unit wrote a script to automatically download all rubric data for each section once a week. In Spring 2017, this meant that 14 .csv files were generated each week, each representing all of the grading completed in a given section from the first day of class through to the download date.

The information contained in each section's data download included the proficiency level marked and written feedback provided to each student for each learning objective assessed on a given rubric. Each rubric item entry (and thus learning objective assessed) was also tagged to the person who most recently submitted the rubric (i.e., the original grader or the person who performed a regrade). In Blackboard, the entire rubric had to be resubmitted whenever one or more rubric items was updated.

Immediately prior to each weekly instructor meeting, the results of the previous week's grading was analysed. In Spring 2017, the analysis focused on answering three questions:

- What was the students' proficiency level with each learning objective at the course and section levels?
- How reliable were the UTAs at assigning the proficiency levels?
- Did the UTAs give adequate and meaningful feedback?

Python code was developed to process the .csv data files. This code generated the following results that were shared with the instructors:

- A summary of students' proficiency level with each learning objective at the course level. This was in tabular form and had to be additionally processed in Excel to create a visual (see Figure 2 for an example).
- A summary of students' proficiency level with each learning objective at the section level and by UTA. This was in both tabular and visual forms (see Figure 3 for a visual example)
- Word clouds of the written feedback for the overall assignment for the section and UTA (see Figure 4).

SBG Visualizations

Three types of visualizations were shared with the instructors and GTAs in the weekly instructor meeting. The selection of these visualizations was based on keeping attention focused on the questions needed to be answered and making the visualizations as easy to interpret as possible for answering those questions. The first visualization was the overall course performance on each of the learning objectives. A stacked bar chart was used to quickly identify the learning objectives with which students in the course appear to be having the most difficulty. An example of this is shown in Figure 2. In this example, it can be seen from the percentage of Developing, Emerging, and Insufficient Evidence marks that students had some difficulties with plotting and execution of user-defined functions. It can also be seen that as many as 8% of the students did not attempt the parts of the problem set associated with learning objectives 07.05, 11.06, and 11.08. A review of the written feedback was required to identify the exact pieces of evidence that were at the root of the difficulties with each learning objective. Again, this was necessary as Blackboard did not enable checklists within a rubric.

The second visualization was the section level view of students' proficiency with the learning objectives. An example is shown in Figure 3. In the upper left corner is the overall section performance. This can be visually compared to the course overall performance (Figure 2). While the 07.00 series of learning objectives appear similar at the course and section levels, the section appears to be doing better with the 11.00 series of learning objectives.

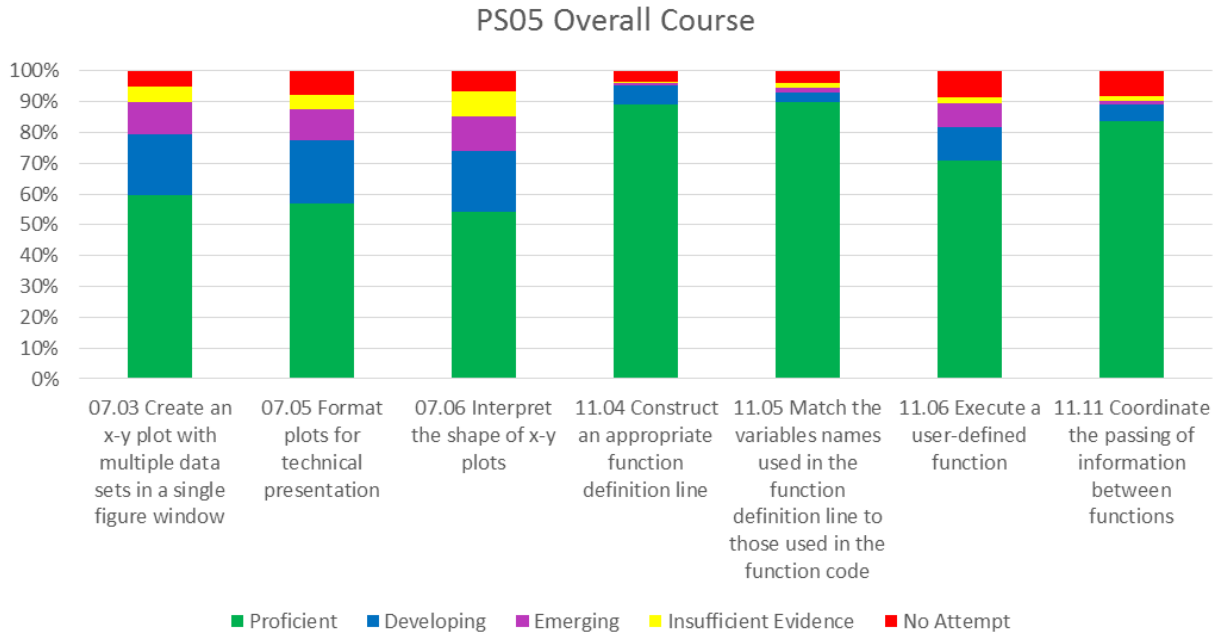

Figure 2. Sample overall course proficiency with the learning objectives on a problem set.

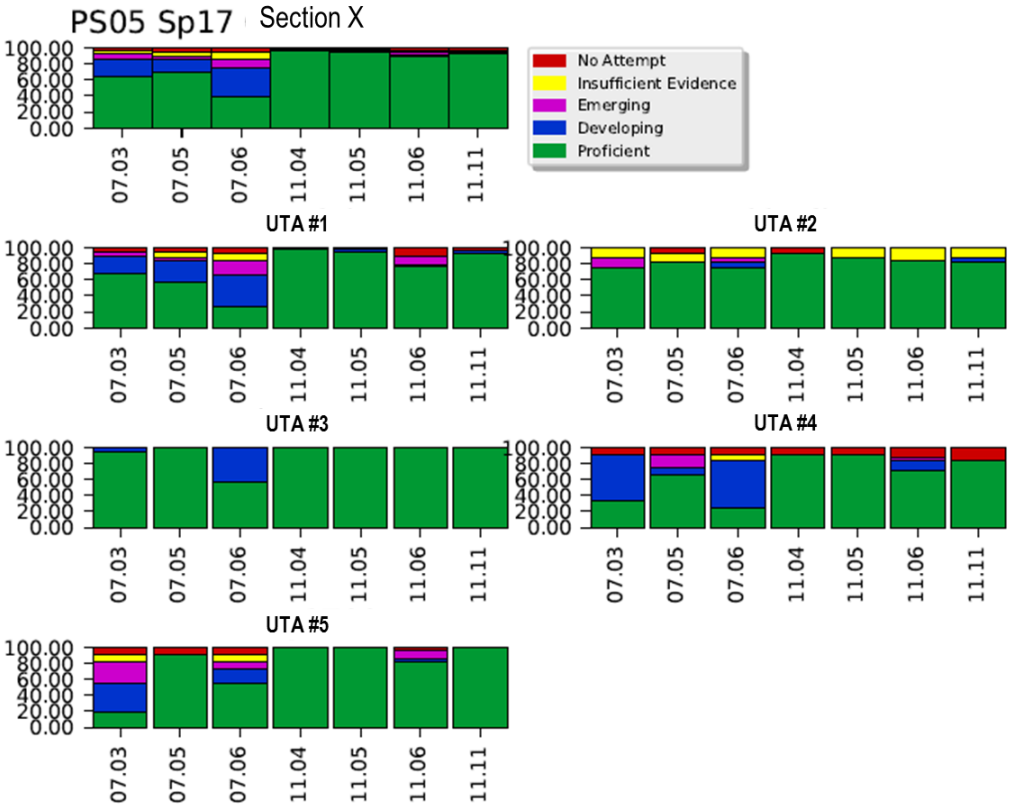

Figure 3. Sample section level course proficiency with the learning objectives on a problem set.

The remaining stacked bar charts in Figure 3 show the grading results from the individual UTAs associated with the section. These results were used to identify UTAs that may need guidance on how to apply the SBG rubrics more reliably. While variation in the results are expected from one UTA to the next due to the small number of students each UTA grades, trends can be discerned that might raise questions worth pursuing to better ensure grader reliability. For the section shown in Figure 3, there may be cause for some concerns about the grading by certain UTAs. UTA #3 only marks students' work as Proficient or Developing. Is this UTA using the rubric or rushing to complete grading? UTA #5 only marks students' work as Proficient for certain learning objectives. Is this UTA having difficult applying the rubrics for just certain learning objectives? UTA #2 uses the Insufficient Evidence mark more than the other non-Proficient levels and does so more than other UTAs. Is this UTA have trouble differentiating the levels of proficiency or is this UTA actually applying the rubric more faithfully than the other UTAs? Note that all of these questions would not likely show up in a single section, but demonstrate the types of questions that might be raised.

The third visual was a series of word clouds summarizing the written feedback students received on all of the learning objectives assessed on a given problem set. Figure 4 shows an example. The upper-most word cloud is for the section and the others are for the individual UTAs. This visualization enables a very top level evaluation of the quality and quantity of written feedback on a problem set in its entirety, rather than at the learning objective level. UTAs #2 and #3 use a wide variety of words in their written feedback and the words are related to the evidence of proficiency for the learning objectives assessed on the problem set. UTAs #1 and #5 may be overusing the words good and correct rather than giving substantive constructive feedback. UTA #4 does not provide much written feedback; this is evident from the relatively few words in the word cloud. This lack of written feedback might pair with predominantly marking the learning objectives as Proficient and No Attempt.

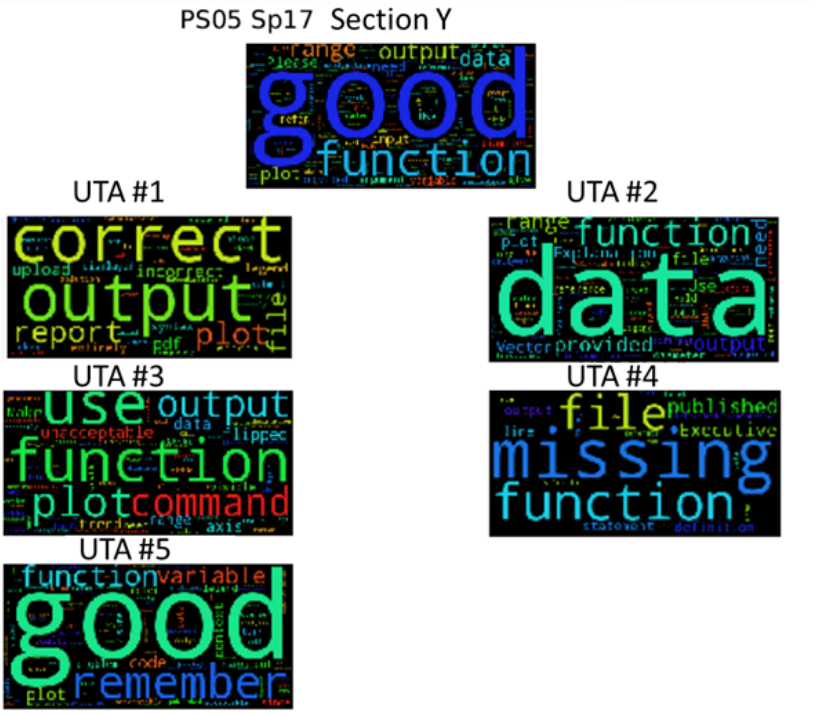

Figure 4. Sample word clouds of written feedback on the learning objectives on a problem set.

Informed Action and Change

Figure 5 represents the ultimate flow of how SBG data was collected and used for informed action and change. The course consisted of the course materials and the implementation strategy as delivered by the instructor, GTA, and UTAs. Students' work was assessed first by the UTAs and spot-checked by the GTAs. Once grading of an assignment was complete, rubric data was downloaded and analysed so that it could be interpreted collaboratively by the curators, support staff, instructors, and GTAs. The initial collaborative interpretation of data occurred in the weekly instructor meeting. Ten to 20 minutes of this hour long meeting was devoted to interpreting the overall course performance (e.g., Figure 2) - identifying learning objectives that appear difficult for students, considering UTA grading issues, and discussing strategies for going forward both in the short and long term. Instructors and GTAs were tasked with comparing the overall course performance to their section's performance (following some training on how to interpret differences).

Following the instructor meeting, instructors and GTAs took immediate action. They used the results in the following week to help their students with particular learning objectives. For instance, they might have done an additional activity around a difficult concept. Or they might have reviewed a particular part of the problem set. They also worked with their UTAs on their grading strategies and, in some cases, regraded student work associated with one or more learning objectives. Prior to SBG grading and these weekly reviews of the results, few of these immediate actions were taken on a regular basis, particularly those around improving UTA grading.

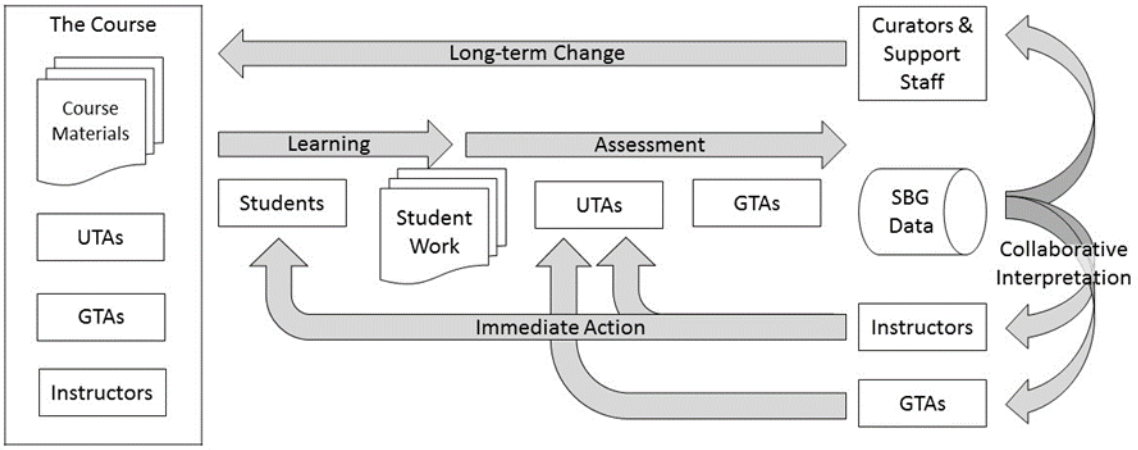

Figure 5. Flow from student work to SBG data collection to informed action and change.

The curators and support staff used the results to inform more long-term changes. For instance, revisions were made to the instructions on various problem sets to clarify what was being asked or to align the problem better with the learning objectives. Such revisions have raised the standard for the design of new problems. SBG results also prompted the development of new instructional strategies for learning objectives that were particularly difficult for students. In addition, SBG results raised awareness of a need to improve the reliability of the grading on particular learning objectives. Specifically, rubric items were clarified to make them easier for

UTAs to apply and UTA training was adjusted or created. None of these types of long-term changes were as confidently, as targeted, or as quickly pursued prior to SBG grading.

Discussion

SBG data provides a powerful means to assess student learning and monitor grading. An SBG system with strong alignment of learning objectives, tasks, and assessment yields grading data that clearly highlight problems with specific learning objectives. However, the power of SBG can only be realized through careful front-end design of the grading system to address fairness, validity, fidelity, and integrity, and reliability. Subsequent ongoing vigilance is necessary to maintain these aspects of high quality grading data because courses are not stable entities – content, assignments, assessments, graders, etc. change over time.

Vigilance is enabled through the selection of meaningful representations for grading-derived data. Simple visualizations help instructors and staff target learning objectives that need attention. While it would be optimal for the interpretation of what is happening with a particular learning objective to focus solely on issues of student learning, the reality is that the interpretation is never completely divorced from a need to consider the quality of the data. That is, whenever analysis results of SBG data is brought forward to the instructional team, one has to ask a three-part question. Is the performance result for a given learning objective a reflection of (1) student learning, (2) grader reliability, or (3) a fault in the grading system (e.g., task assigned to the students, design of the rubric item) that is undermining fairness, validity, fidelity, integrity, and reliability?

As part of a clear path forward from data representations to interpretation to actions, an open dialogue among the instructors and staff is necessary to root-out the nature of the problems with a learning objective that appears difficult. It has to be acknowledged that various course-level stakeholders have different and valuable experiences with the course content and materials, the students, and the graders. These experiences need to be brought together to meaningfully interpret the date and make decisions regarding appropriate short- and long-term actions.

The other crucial part of the clear path forward is the delineation of the roles that various course-level stakeholders have in the grading system. Ultimately, clearly articulated roles enable any action to be carried out.

Conclusion

This paper demonstrated, though a specific course example, how SBG derived data can be used to monitor student learning and grading. Three criteria for using SBG data confidently and effectively for this purpose were established. First, the grading data have to be of high quality. Second, the grading data results need to accessible through simple visual representations. Third, there needs to be a clear path forward from grading data, to interpretation, to actions.

Building an SBG system that has the power to enable confident and effective course-level change is a long-term commitment. Is it worth it? In a large, complex course with many course-level stakeholder, it is very difficult to know many details about student learning. When using a

traditional grading system, one can at best know that students performed well or poorly on a given assignment. That level of grading detail provides no information about students' abilities to demonstrate proficiency with individual course learning objectives. The lack of information leads to decision-making that can be ill-informed and unfocused, guided by intuition and hunches. What is revealed about students' learning and the course itself as SBG data quality begins to be managed and visualizations begin to be discussed on a regular basis among the course-level stakeholders is game changing. Increasingly more conversation is about particular learning objectives and related instructional strategies and grading reliability. Actions are taken on a daily basis, not just on a semester-to-semester basis. The whole course enterprise comes to value data derived from grading as a means for responsive, evidence-based decision-making to improve student learning.

Future work could proceed in a number of directions. First, the SBG data analysis methods could be refined, particularly to more rigorously identify sections and graders with reliability issues. Second, the data analysis methods could be extended to include cross assignment or complete semester analyses and report generation. These would be useful for tracking trends in student performance and accreditation purposes. Also, automated ways of analysing the written feedback, such as through natural language processing, could be explored to better identify the specific difficulties students have with a given learning objective. The more pressing need at this time is to thoroughly document the SBG data analyses methods outlined in this paper and test them with other courses so that they can be shared more broadly with others.

Acknowledgement

This work was made possible by a grant from the National Science Foundation (NSF DUE 1503794 and NSF IIS 1552288). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The authors wish to thank the whole course instructional team for coming along on this journey, particularly the support staff who have worked diligently to meet the first criteria.

References

[1] B. E. Walvoord and V. J. Anderson, *Effective Grading: A Tool for Learning and Assessment in College*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2009.

[2] Alverno College Faculty, *Feedback is Teaching*. Milwaukee, WI: Alverno College Institute, 2015.

[3] S. L. Post, "Standards-based grading in a fluid mechanics course" in *121st ASEE Annual Conference and Exposition, Indianapolis, IN, USA, June 15-18, 2014.*

[4] M. W. Roberts, A. M. Jones, and M. K. Thompson, "Work in progress: using outcomes-based assessment in an introductory structural engineering course" in *121st ASEE Annual Conference and Exposition, Indianapolis, IN, USA, June 15-18, 2014.*

[5] M. T. Siniawski, A. R. Carberry, and J. D. Dionisio, "Standards-based grading: An alternative to score-based assessment" in *2012 PSW ASEE Conference, San Luis Obispo, CA, USA, April 19-21, 2012.*

[6] H. A. Diefes-Dux, "Student reflections on standards-based graded" in *46th ASEE/IEEE Frontiers in Education Conference, Erie, PA, USA, Oct. 12-15, 2016*.
[7] H. A. Diefes-Dux, "Student self-reported use of standards-based grading feedback and resources" in *Research in Engineering Education Symposium 2017, Bogota, Colombia, July 6-8, 2017.*
[8] P. L. Scriffiny, "Seven reasons for standards-based grading," *Educational Leadership*, vol. 66, no. 2, pp. 70-74, Oct. 2008.
[9] D. R. Sadler, "Interpretations of criteria-based assessment and grading in higher education," *Assessment & Evaluation in Higher Education*, vol. 30, no. 2, pp. 175–194, April 2005.
[10] J. Heywood, "The evolution of a criterion referenced system of grading for engineering science coursework" in *44th ASEE/IEEE Frontiers in Education Conference, Madrid, Spain, Oct. 22-25, 2014*
[11] R. M. Felder and R. Brent, *Teaching and learning STEM: A practical guide*. San Francisco, CA: Jossey-Bass, 2016.
[12] D. R. Sadler, "Grade integrity and the representation of academic achievement," *Studies in Higher Education,* vol. 34, no. 7, pp. 807– 826, Nov. 2009.
[13] D. R. Sadler, "Fidelity as a precondition for integrity in grading academic achievement," *Assessment & Evaluation in Higher Education,* vol. 35, no. 6, pp. 727–743, Oct. 2010.
[14] N. M. Hicks and H. A. Diefes-Dux, "Grader consistency in using standards-based rubrics," in *124th ASEE Annual Conference and Exposition, Columbus, OH, USA, June 25-28, 2017*.
[15] J. B. Hylton and H. A. Diefes-Dux, "A standards-based assessment strategy for written exams," in *123rd ASEE Annual Conference and Exposition, New Orleans, LA, USA, June 26-29,* 2016.