Standards Based Grading in a First Proofs Course

KATHERINE E. STANGE

Abstract. Standards-based grading, in which grading should be designed to communicate to students their current level of mastery with regards to well-articulated standards, is becoming popular at the K-12 level. As yet, the literature addressing standards-based grading at the university level is scarce. In this paper, I document my attempts to put into practice the principles of standards based grading in a lower-level undergraduate mathematics course which aims to introduce mathematical proof.

1. Introduction

Our department created a lower-level course in *Introduction to Proof* five years ago in 2008, to address a perception that students were not prepared early enough for the rigorous proof-based approach in upper-level classes such as abstract algebra and analysis. Such courses are common in undergraduate mathematics major programs around the United States, and, as did many others, our department chose as the subject matter on which to practice proof writing the general area of discrete mathematics. The course is titled *Mathematics 2001: Introduction to Discrete Mathematics*.

The course presents unusual challenges, in that it aims to teach both certain mathematical content (logic, sets, functions, and the like), and also the art of proof writing and more generally mathematical communication, which is for most students an entirely new adventure.

In Fall 2015, I taught the course for the second time. I had as one of my principal goals for this semester to incorporate *standards based grading* into the classroom. I found that in the previous semester I had failed to communicate very effectively the standards of mathematical thinking. This time, in keeping with practice on backwards design and learning goals (Wiggins and McTighe, 1998), I wanted to build my grading system around clearly communicated standards and learning goals.

Standards based grading is the implementation of one simple grading principle: *that grading should communicate to students their current mastery and progress with respect to well-defined standards*. This principle should over-ride other uses of grading, such as grading as a means to promote desired behaviour. Standards-based grading dates to 1990's K-12 educational reform (Marzano, 2010).

In practice, this means most standards based grading systems involve several practical aspects:

(1) well-articulated standards of assessment,
(2) which are graded individually, and
(3) frequent opportunities to reassess standards.

In most implementations, later assessments of a given standard overwrite previous ones, so that the current grade reflects the current level of mastery. There is a growing literature on best practices (Marzano 2010; Marzano & Heflebower, 2011; Vatterott, 2015). Standards-based grading is also closely related to *mastery learning*, in which students persevere on each topic until (ideally) mastery is reached (Block, 1971).

Standards based grading is a popular topic in K-12 education (Marzano, 2010; Reeves, 2003; Vatterott, 2015), but has yet to take root in a university setting, where the literature on

Date: August 8, 2016, Draft #1.
2010 *Mathematics Subject Classification.* Primary: Secondary:
Key words and phrases. Standards Based Grading.

the subject is so far small but hopefully growing (Beatty, 2013; Post, 2014; Duker, Gawboy, Hughes and Shaffer, 2015; Owens, 2015; Rundquist, 2011). In this article, I aim to chronicle my attempts to build Math 2001 around a standards based grading scheme. This seemingly modest goal bled into almost every area of instruction, and influenced the entire structure of the course. Discussing my successes and failures will highlight the challenges of implementing standards-based grading, particularly in a university environment.

I was inspired to write this article by a fascinating account of an attempt to use standards-based grading in an undergraduate physics course (Beatty, 2013), which I found very helpful during the lead-up to my own course.

2. Narrative: Course Design and Execution

My interpretation of the subject matter in Math 2001 included the usual core subject areas: sets, logic and boolean algebra, enumeration, relations, and functions (which I will call the *five areas*). This material, however, forms only one half of the course. The other is the art of proof writing, and more generally, mathematical communication. In order to practice this art, one is in need of a substrate, or material: the five areas, being among the most foundational topics for future classes, are an ideal choice. Other instructors sometimes include topics such as probability or graph theory. I dipped into these (particularly graph theory) to create more interesting examples as needed.

I wished to assess my students' abilities in two arenas: first, their ability to apply and reason with the basic definitions and properties covered in the five areas; and second, their ability to write logically sound and communicative proofs. I approached the design of a grading system for these two course goals separately.

Before designing the system, I made one firm decision: all assessment would be in-class assessment. My experience in past has been that homework ill reflects an individual student's understanding, resulting as it does from collaboration, seeking clues on the internet, and so forth. In the worst case, students copy work: I was mortified to discover in my previous incarnation of Math 2001 that some students had copied answers from a paywalled internet site. But even the most assiduous and well-meaning students sometimes turn in homework answers that do not fully reflect their personal understanding, for a variety of reasons, ranging from conversations with well-meaning but ineffective tutors, to a lack of study skills and self-awareness. Cognizant that my standards based grading system would probably involve a significant investment of my personal time in grading, and hoping that the assessment would strongly reflect the student's 'true' abilities, I resolved to avoid this quagmire. I hoped that by making assessment as frequent as every single class, the drawbacks of in-class assessment (namely, stress and time factors) would be sufficiently lessened. Evidence indicates that the use of very frequent assessment with feedback supports learning directly (Gibbs and Simpson, 2004) and by encouraging evenly paced studying (Kornell and Bjork, 2007), and anxiety is an important factor in learning (Boekaerts, 2010).

Finally, I should explain that the teaching style I've developed over the years is very adaptive and impromptu. As a result, I prefer not to give a very rigid system of homework, but instead to post daily updates on a course website with suggested problems or readings before the next class. Since none of this is assessed, students don't seem to mind the flowing, short-notice nature of it, and I know they adapt the load to their schedules, putting off tasks to suitable moments instead of following my schedule.

At the end of semester, the final exam in the course was given in a traditional style, and was worth 30% of the grade.

2.1. **The Five Areas.**

The five areas consist primarily of mathematical definitions and their logical consequences. The student, ideally, becomes fluent in terminology and calculation and

(1) set, element, equality of sets, empty set, cardinality of set
(2) set-builder notation
(3) ordered pair, ordered n-tuple, Cartesian products and powers
(4) subset, power set
(5) union, intersection, difference
(6) universe, complement, Venn diagram
(7) cardinality of Cartesian product or power
(8) counting subsets
(9) counting by independent choices
(10) possible overcounting
(11) Self-study sections 1.9, 1.10
(12) Evaluation of boolean expressions, truth tables
(13) Logical equivalence
(14) Converse and contrapositive
(15) Negating statements
(16) Quantifiers
(17) Logical laws
(18) Relations – definitions, ordered pairs, arrow diagrams
(19) Properties of relations
(20) Modular arithmetic
(21) Function definitions
(22) Composition of functions
(23) Inverse functions
(24) Image and preimage

FIGURE 1. List of content badges.

sufficiently familiar with the basic properties of these new characters that she can make inferences about their behaviours in new contexts. In particular, in contrast to the style in classes such as calculus, testing this material is often effectively done by 'concept-check' type problems, instead of involved calculations.

I decided on a *badges* system, the term being chosen to evoke boy scout badges earned for skills such as sewing or building a fire, which I hoped might motivate students (Boekaerts, 2010). Each student's score for one badge was either 0, 0.5 or 1, indicating little or no facility, partial mastery, and mastery, respectively. This rather coarse assessment facilitated grading, and my intention was to choose sufficiently many small topics that a finer gradation would not be needed.

In order to make room for frequent assessment, I planned to give a quiz during the last ten minutes of every 50-minute class period (3 per week). Half of these quizzes would be badges quizzes consisting of short-answer questions. Each question would test just one badge, and be labelled with that badge. Students could choose which questions to answer based on their own mastery to date (each quiz containing rather more than is reasonable for ten minutes), and their score would be updated only upward, never downward, on subsequent attempts. A 'short-answer question' typically consisted of filling out a table of properties for 3 or 4 objects, the computation of a number by use of a definition, or generating an example with required properties, etc. Some example questions are included in Appendix 9.

I began the semester with a list of badges designed for the topic of *sets*, which included 15 badges. This would seem to have projected a final total for the semester of $15 \times 5 = 75$ badges, which I never reached: I finished the semester with 24 badges, having limited *sets* to 9 badges, and choosing fewer for later topics. Figure 2.1 shows the final list of badges. I quickly discovered that, first, the badges system produced a rather heavy administrative overhead, and second, it was tedious and unhelpful to write short-answer problems for the same tiny prescribed topics. By creating larger topics, I gave myself the freedom to test some synthesis of ideas and create more novel problems day-to-day. Novel problems (not just 'recipes') seemed to me essential to making the frequent assessment an active learning tool.

Students were initially a little confused about the grading system and sought to complete every question on every quiz (which was difficult in ten minutes); I had several times to re-explain. In the beginning, I handed each student a photocopied list of their current badges

each day, but this system rapidly proved itself to be low in the value-to-time-cost ratio, and once I was sure the students understood the system, I asked them to track their own badges. About 2/3 of the way through semester, I started to use a course management system to let students have electronic access to my gradebook, where they could check which badges they had earned. Entering data into the course management system was very costly in terms of time but students found it useful that they could check their badges before class on their phones.

When I first designed the badges, I aimed, as is the suggested practice, to phrase badges as I had been taught to phrase course goals: in terms of a verb, action, ability or skill. But I gave up on this, finding that the verb (actually a collection of several verbs) was always the same. For example, a great many topics consisted, essentially, of one mathematical definition (e.g. *relation* or *power set*). My goal was for a student to be able, in the face of definition X, to be able to:

(1) state the definition
(2) give examples and non-examples
(3) compute, for a given input, the output of the definition
(4) determine if a given example has the stated property
(5) infer logically from the definition to the behaviour of examples in a given context

So I simply labelled the badge 'X', rather than writing this list over-and-over (or dividing it up into separate badges). On a given quiz, one or several of the skills above were tested, and which one varied day-to-day. This allowed me to create sufficiently varied tests of the same definition so that students were, I hoped, forced to grapple with the definition through a variety of routes.

For example, to test the student's understanding of *injective*, *surjective* and *bijective* functions, I could ask him to fill out a table of properties for some given functions, to provide an example of a function $f : \mathbb{Z} \to \mathbb{Z}$ that is injective but not surjective, to state the definition of *injective* precisely, or to adjust the codomain of a given function $f : \mathbb{R} \to \mathbb{R}$ so that it becomes surjective.

Of course for some badges which were not simply definitions, e.g. *set-builder notation* or *counting with overcounting*, the list above varied slightly. But in each case, a badge for me was really a subtopic, rather than a skill; the skills were the skills of mathematical reasoning, as applied to the topic at hand. In this way I departed in my design from established classifications of course goals (Wiggins and McTighe, 1998).

Badges were typically active for between 3 (in the case of *functions*, the last of the five areas) and 6 (in the case of *sets*, the first of the five areas) quizzes. Once students seemed to have mastered the topic as a group, I stopped writing new questions, mainly to save time. Students did significantly better on their *sets* badges than their *functions* badges: on a typical *sets* badge, all but 2 of the 26 students earned full credit (*set builder notation* proved to be a particularly challenging one, with 6 students missing full credit), while on a typical badge in *functions*, only 9 of the 26 students had received full credit when semester ended.

At the end of our 15-week semester, I had given 16 badges quizzes, somewhat less than the 22 or so which would have resulted had I used exactly half of the course meetings. This is partly due to a number of days when I declined to interrupt particularly effective in-class learning to take a quiz.

2.2. **Proof Writing.** My plan was that the other half of the quizzes would consist of proof quizzes, in which a student is asked to write one proof. These quizzes generally had two sections: *Tools* gave some relevant definitions and propositions, and *Task* stated a theorem to be proven. My grading rubric is given in Appendix 10. For each quiz, I graded *writing* out of 4 points, and *reasoning* out of 4 points. I also allowed some quizzes to have a *synthesis* grade of 2 points.

Since grading writing involves comparing what the student wrote to what they intended to communicate, I left myself the option of *ungraded* when the reasoning was sufficiently lacking,

or the writing so incomplete, as to render it impossible to assess usefully. In practice, a grade of 'ungraded' translated to a zero in the gradebook.

Somewhat surreptitiously, I also wished to test mathematical communication more generally than just proof writing. My initial grading breakdown involved reading assignments, but these were quickly and quietly dropped after the first week, when I discovered I didn't know how to effectively design or grade these. So my hope was that the ability to write clear, precise and correct proofs would entail other skills such as the ability to read a mathematical definition with precision. Simply put, I found incorporating these standards into the grading scheme individually too daunting. I aimed to do a good job on assessing the more demonstrable skill of proof writing, in the hopes that I may be able to approach this issue better after a first experience.

I had hoped to test students on approximately 15-20 proofs; I managed to assess 12 during semester (as well as one make-up quiz in which I allowed them to rewrite one of the past quizzes). I set out with the goal of averaging the best seven of their efforts, assuming this would be around half. In the end, I averaged their best 6 writing grades and their best 6 reasoning grades (separately).

Midway through semester I received feedback (through an online survey and informal classroom conversations) that students found the time pressure on proof quizzes oppressive. Specifically, they found there was not time to think through the ideas *and* write nicely. I addressed this by taking the students' own advice to give topics in advance, so that some of the thinking could happen at home. This took the form of listing possible theorems they would be asked to prove, or listing the proof type (e.g. demonstrate that a given function is surjective and/or injective). They seemed satisfied with this.

Synthesis points were aimed at assessing the student's ability to connect disparate topics in the course into a new logical argument. My motivation for including them was probably a certain guilt brought about by the ruthless way I had chopped up the course into individual standards, leaving little room for students to practice combining ideas in novel ways. At the end of semester, however, I had chosen to assess synthesis on only two of the proof quizzes. As a result, I adjusted my grading scheme so that this score declined from 5% to 2% of the final semester grade.

3. Results: Student Opinions

At the end of semester, I gave a feedback form with some survey questions specifically addressing standards based grading. In general, the vast majority of the students (19 of 21 respondents) preferred the standards based grading system to a typical university grading system.

They also agreed with statements that it helped them study more evenly throughout semester, develop confidence, learn the material, and get higher grades (for the same level of understanding). They agreed it made class less stressful, and that the grading standards were made clear, albeit agreement was more universal on this point for badges than for proofs. Finally, they seemed to feel the number of opportunities to reassess was appropriate, or slightly too high.

Given the opportunity to offer further feedback, students commented on the way the grading system caused learning to spread out across a longer time frame and evened courseload (4 comments), helped students zero in on what needed to be studied and spend more time there (3 comments), and reduced the stress of testing (6 comments) (for some, this was a negative, as stress is a motivator). Other comments included praise of the frequent feedback, the observation that it increased scores, and that it influenced student attendance (I observed higher-than-usual attendance, but some students said it facilitated truancy). At least one student felt that doing all assessment in-class without aids was unfair (as compared with a homework-based course).

As for implementation improvements, giving warning on proof quiz topics was a major topic of feedback, as was the timed aspect of proof quizzes more generally. Some students wanted more or fewer chances to attempt badges.

Students also commented on the effect on study habits, but here it was difficult to discern whether it was a net positive or a net negative. Some students found their test anxiety improved, and others found the motivation of test stress was removed. Most found they studied more evenly through semester, but others less evenly. One said the system could be misused by cramming.

Overall the feedback was positive, and is included in its entirety in Appendix 11. The end-of-semester Faculty Course Questionnaire (designed and administered by the university) included few written comments, and no standards-based-grading questions.

4. Results: Instructor Experience

4.1. **Immediate Feedback.** One of the most positive aspects of the experience, for me, was the constant feedback. Between every class, I graded the most recent quiz. If we were studying one of the five areas, this was a badges quiz, and it included the badges we were currently discussing (students had the opportunity to test on a topic already at the end of the class in which it was introduced). If we were working on proof-writing, the quiz topic sometimes lagged, but only by a day or two. As a result, I went to each lecture knowing exactly what the students hadn't understood from the previous lecture. I was able to address confusions while they were just budding, and I often saw immediate improvements on the very next quiz. The badges system also allowed me to see easily which topics students struggled with (for example, consistently challenging were *set-builder notation*, *counting with over-counting*, and *logical laws*). This was the single most impressive outcome of this semester's experiment, although any frequent-assessment system will have at least some of this benefit.

4.2. **Instructor load.** The grading load of my standards based grading system was not inconsiderable, but not as heavy as I had anticipated. I found that a proof quiz took about an hour to grade, sometimes less. A badges quiz took only 20 minutes, but data entry took an additional 20 minutes, at least (I tried tracking on paper records, one page per student, and later through a course management system; a standard computer spreadsheet, suitably set up, would be significantly faster).

4.3. **Class time.** My quiz system used significant class time. I think it may not be feasible for a course with a heavier load of material. I was very fortunate that *Math 2001* comes with a lot of space to work on skills (as opposed to covering content). I am also less efficient in covering material because I make use of various active learning techniques, which tend to take more classroom time. I found 40 minutes a significantly shorter class period than 50 minutes, and I noticed the time lost. However, I was very pleased with the constant feedback from grading, and I was also persuaded that the quiz time was useful learning time, so I felt it was justified. In the end, I gave 29 graded quizzes over the course of 45 classroom periods (not counting two quizzes which were not graded).

4.4. **Grade Inflation.** Initially, I was very worried that semester grades would be deceptively high. This definitely happened. The averages during semester were 81% on badges, 83% on writing proofs, and 85% on proof reasoning, while the average on the final exam was a huge drop to 58%. I had allowed the high semester grades to convince me the students were at a higher level of understanding than they were able to demonstrate on a traditional final. This tempted me into setting too hard a final exam. I felt a sense of disillusionment, and I fear the students did too. I think anyone attempting standards-based grading needs to calibrate their understanding of the grade averages according to the new system.

4.5. **Class Participation.** Overall, compared to other similar classes, I found this class to be noticably above-average with regards to in-class participation and attendance. I hope that this reflects a healthier classroom emotional state (Boekaerts, 2010).

5. Results: Instructor Experience with Badges

5.1. **Testing synthesis.** The nature of the badges system left no room for questions which tested some synthesis of their topics. This made all the topics appear disjointed and separate, and, I fear, unmotivated. More interesting problems might combine an understanding of different badges, but I was unable to test students on these during the semester. An alternative system may have been to allow some synthesis problems, which are then graded with respect to more than one standard.

5.2. **Variation in level.** I also felt hobbled by the inability to vary the difficulty level of the badges questions. I felt that if the overall difficulty level of a badge varied day-to-day, then the grade would be less meaningful, since students could just 'wait for an easy one'. However, this made it more difficult to assess a variety of different types of understanding, as classified for example, in the form of Bloom's taxonomy (Bloom et al., 1956). The closest approximation was to measure different levels of understanding with a single question: for example, one short-answer question might actually consist of a collection of four True/False questions varying in level, so that partial credit is likely but full credit is more difficult. I did this increasingly throughout semester.

5.3. **Optimal reassessment.** The number of tries for a given badge varied from 3 to 6 during the semester. The optimal number of tries, balancing student success with the effort and time of repeating, seemed to be closer to 6. Over the first few reassessments, students seemed to improve rapidly, especially if I gave over class time to taking up quiz problems as a group (which they often requested).

5.4. **Learning during assessment.** There is the danger that a student can manage one style of question but not others, and obtains the full badge on the first try, never attempting other questions on that topic. However, this danger is just as common in standard grading setups, and I did encourage students to attempt questions again just for further feedback, once they have attempted their currently incomplete badges. I found that they did indeed do this (unless I was just observing near-constant confusion about which badges they had earned), and they did frequently earn less than full credit on a badge they had completed in my gradebook. Comparing to a standard mathematics homework-midterm-final course, I believe the students had more chances for feedback on each topic. Furthermore, at the beginning of each class I handed back the previous day's quiz and the answers were discussed. So overall I was satisfied that the in-class time given over to assessment was useful even for students who had earned their badges.

5.5. **Skills standards for badges.** In writing this article I clarified the list of skills attached to a given badge (see Section 2.1). However, I never clarified this list to students, and I regret this, as communicating goals clearly is a well-supported practice (Wiggins and McTighe, 1998).

5.6. **Self-study badges.** During semester I cut out several badges from the *sets* topic as it became clear I needed to speed up the course a little. However, I left in the book's brief discussion of Russell's paradox and further topics as a badge titled *self-study sections 1.9, 1.10*. I never taught this in class but let it appear on quizzes with short questions like 'state Russell's paradox'. Students didn't seem to have a strong positive or negative reaction to this state of affairs, and I may repeat it in future as a way of giving the more motivated students something more significant to try for. This must surely be used in moderation, however: only 2 of 26 students received full credit for this badge, and 10 received partial credit.

5.7. **Updating grades.** Probably out of fear, I chose only to update grades upward, never downward. A grading system more true to the spirit of standards based grading might allow some downward adjustment. However, I didn't want to discourage students from attempting badges questions on quizzes for fear of lowering their grade. Therefore I would probably have needed to make questions mandatory, not optional, on badges quizzes. This would have resulted in a more rigid system, where students were not free to focus on their trouble areas.

5.8. **Badges sizes.** It would have been possible to choose much more comprehensive topics for the badges, instead of breaking them down into single definitions or ideas. For example, I could have had a badge for all the basic definitions of sets, encompassing about four of the badges I did use. I believe this would have necessitated rather longer and more carefully designed test questions and a rather more finely divided grade scale. But it would have allowed for more synthesis and more variety of levels within a given question, and it may have been closer to previous implementations of standards based grading in university physics (Beatty, 2013).

6. Results: Instructor Experience with Proof Quizzes

6.1. **Grading Writing, Reasoning and Synthesis.** My crude division of proof writing into just three topics turned out to be both too fine and too coarse.

The grading of writing and reasoning separately was not as difficult as I anticipated, in large part because I allowed myself the option of not grading writing when the reasoning was so lacking as to make it impossible. In practice, however, I did catch myself adjusting the grades so that the sum of the reading and writing grades matched my gut feeling about the value of the proof as a whole, which goes against the intention of standards based grading.

Synthesis points in practice applied to very few of the proof quizzes. This was primarily because the course is a first introduction to proof, so we focussed on the basics. Furthermore, I found myself inadequate, in practice, at effectively separating synthesis from reasoning when assigning grades, so I couldn't convince myself that the separation was conveying any information. In future I will drop them entirely.

However, this leaves the question of how to assess synthesis in any way in the course, or whether assessing it is a reasonable goal. It would seem incompatible with standards based grading systems. (citations?)

6.2. **Rubric.** The rubric stapled to each returned proof quiz was meant to reiterate the standards for proof grading. I found, however, that I used the sheet mainly for recording the grades. It was easier to write the student long comments on the proof itself than to seek out and circle the relevant advice on the rubric. I did the latter less and less throughout semester. Although the rubric attached to every quiz may have provided some benefit in itself, as it draws attention to the specific standards, it was not effectively integrated into the grading system.

6.3. **Advance warning on proof topics.** In the second half of semester, I tried to fulfil my students' requests to have advance warning on the quiz topics. My teaching style is a relatively spontaneous combination of lecture, groupwork, concept-check questions, games, etc. Between each lecture, grading the quizzes gave me feedback on what topics the class was struggling with, and I often decided on the day's lecture topic only hours before class. So I struggled with this. I was also concerned that if I warned them of the proof quiz topic, that they would produce a proof at home with the help of a tutor or friend, one that they didn't fully understand, then parrot it back during the quiz. In practice, I did not see evidence that this was happening.

6.4. **Averaging the best half.** Would it have been more true to the principles of standards based grading to average the most recent proof grades instead of the best ones? Each proof quiz was a novel type of proof, presenting new challenges. If students were improving over semester, it would not necessarily be visible in increasing scores: in other words, the target was moving. I was dismayed not to be able to re-test older methods of proof more frequently. But

I could not have graded any more proofs than I did during semester, and I have yet to find an undergraduate grader up to the task.

7. Experience-Based Suggestions

7.1. **Keep it simple.** In-class time and instructor grading time are both precious and limited. A standards based system must, if it is to succeed at all, be reasonable to implement. Sacrifices in the perfection of the feedback are preferable to a system that will be abandoned as overwhelming. During the semester I imagined many ways my system could be improved, but I hesitate to implement any of these in future, because the 'load' of the current system is already at capacity.

7.2. **Clarify skills when testing topics, and clarify topics when testing skills.** With regards to badges, I did not give any list of skills, such as those I articulated here in Section 2.1. I wish I had done so: providing these may have caused the students to reflect on their own study habits and incorporate some of these ideas into their methods of internalizing new mathematical ideas on their own.

On the other hand, with regards to proofs, I did not ever list the topics, meaning, the types of proof, e.g. *proof by contrapositive*, *proof by induction*. It may have been helpful to design these into the system in some way, so that students could isolate a type of proof they have not mastered. I do use some worksheets which practice setting-up proofs of a given type (e.g. state base case and inductive step without proving them), and perhaps a badges system could incorporate badges of this type, although I fear that 'correctness' for this exercise is not sufficiently well-defined.

7.3. **Design grading to reflect current mastery.** To avoid deceptively high grades, and to give a true assessment of *current* skills, a standards based grading system should allow grades to adjust downward in some fashion. For example, averaging a students' best with most recent grades. In the context of the *badges* system, it is not apparent how to do this without having certain unintended consequences, as discussed in Section 2.1. If the system will inflate grades, as mine did, students should be amply and seriously warned.

8. Discussion

In this section I wish to revisit the three fundamental tensions of standards based grading discussed in (Beatty, 2013), and add one more.

8.1. **Reassessment.** The fundamental advantage of reassessment is to improve learning by giving students more interaction with the material, and to improve teaching by giving the instructor more feedback. It is one of the most powerful aspects of a standards based grading system. However, it takes a significant investment in classroom time and preparation time to prepare and administer so many tests and track so many grades.

I believe Math 2001 was almost uniquely positioned to partially avoid some of these pitfalls because first, it simply had less factual material, and second, the system I designed did not seek to reassess every type of proof separately. If I had taken this approach to a course like linear algebra or calculus, the list of badges would have been prohibitively long.

Even so, the reassessment did incur a high time cost. In the end, I felt it was worth it, but it is possible many instructors may not. This remains one of the fundamental tensions.

8.2. **Grain Size and the Dead Frog Problem.** The second fundamental tension in (Beatty, 2013) concerns how to choose standards. Setting individual standards entails breaking up a subject, likened to a living frog, into little bits, which inevitably die when separated from the whole. On the other hand, if we keep the frog whole, how do we convey to students the particular standard they should address?

In the case of the badges, this tension is primarily one of grain size. If the badges are too small in scope, testing is repetitive and time consuming, but students know exactly where to focus for each badge. If the badges are too large in scope, students may attain credit for a badge while missing whole aspects of the badge, and may find it more difficult to study effectively. In the end, I was relatively happy with the grain sizes of the badges I chose for Math 2001.

This good fortune does not resolve the more fundamental issue, however. A great many subjects – perhaps all subjects – are sacred living frogs. Even without the onus of standards based grading, I think our courses tend, over time, to bloat, as we break up difficult wholes into chewable pieces, and then have to busy ourselves with assigning names and homework questions to each part. The typical calculus syllabus already suffers greatly from this defect, in my experience. Math 2001 is a blessed exception.

So this brings us to testing proof writing. This seemed to me, from the beginning, an inviolable frog. My attempted solution was somewhat outside the usual scope of standards based grading: it did not even attempt to grade separately the various standards of mathematical reasoning or writing listed in my grading rubric. I believe, in fact, that this would have been impossible.

Instead, I hoped that by developing a rigorous rubric and referring to it frequently, students would be able to isolate and reflect upon the issues they needed to work on. This was only partially successful. Every error is a unique gem, a chance to learn a new principle of clarity in writing or reasoning. Rarely are these mistakes repeated in the same form. Nevertheless, I hope that I preserved the fundamental goal of clear individual standards, while simultaneously emphasising the futility of separating them. On the other hand, by separating writing from reasoning, I aimed to emphasize that these two parts, at least, can sometimes be assessed separately, and *are both equally important.*

8.3. **The Attention Economy.** Standards based grading aims to use grading for purposes other than coercing behaviour. So students must realign their work habits to be successful in the abscence of the carrot and stick. Overall, this tension in particular is sensitive to the nuances of implementation.

In Math 2001, nothing done outside of class was directly graded. I asked the students midsemester about their hourly work outside of class. I had only nine respondents to my online survey – probably self-selected to be the more diligent students – and they reported about 4 or 5 hours per week on average. A few students addressed this issue in their feedback at the end of semester, but only a few. In fact, attendance in class was much better than average this semester. So, although I don't doubt that this was an issue for some students, I think the daily quiz system still worked as a carrot-and-stick (although it was not the intention).

8.4. **Skills vs. Topics.** I would like to add one further fundamental tension to those discussed in (Beatty, 2013). It is the tension between skills and topics in setting standards. It is easy to break up a course by its topics. In fact, more often than not, they are conveniently listed in the textbook's table of contents. But if we choose only topics, it may seem that all we accomplish is to replace one single course with fifteen parallel courses, each graded in the traditional way. (Perhaps this is sufficient.)

Of course, mathematical thinking is more than a list of definitions, and we aim to teach students a variety of skills that cut across the course topics. I was faced with this conundrum when deciding how to grade proof writing (where the possible topic standards are less evident and less relevant). For example, I have as goals for my students that they

(1) can digest a novel definition by producing examples, counter-examples and near-examples
(2) can recognize, in a theorem statement, which methods of proof are most likely to apply and succeed, before embarking on that proof
(3) can recognize recursive structure
(4) can identify logical holes in an argument by tracing a proof with an example in mind

(5) can bring together disparate skills to solve a novel problem

In fact, when as teachers we are discouraged, we lament our students' failure to develop these skills, not their failure to understand the notion of *reflexivity of a relation* or *the cardinality of power sets*. These are the goals we *really* care about. So shouldn't we design grading to assess them?

These are well-formed and articulated goals, but they are so numerous, and so perpendicular to course content, that I, at least, do not now how to apply standards based grading to them. If I were to list any one of these as a standard, I would find myself either unable to usefully test it as an individual standard, or, if I could, then I would not find time and space in the course to reassess it throughout. An attempt to do so would, at the very least, greatly alter the way the course is taught. For me, at least, it was too ambitious to attempt.

Perhaps one approach would be to use the usual test problems one uses in a non-standards-based course, and simply label each one with all the myriad standards it might test, grading the student's response separately with regards to each. Some of the standards above would be caught in this system, but the administrative overhead of such an undertaking would seem to make it infeasible for student and teacher, alike.

Furthermore, if we set standards like those above, how do we instruct students to work on a specific skill? Often part of the challenge itself is isolating which skills are needed for a given problem (e.g. recognizing recursive structure). There are no long lists of problems designed to increase facility with recognizing the correct choice of proof method (and one might argue there shouldn't be). One does this exactly once each time one attempts a proof. So how useful is it to point out to the student that this particular skill is what they are lacking?

And, to wax even more philosophical, maybe we shouldn't be trying to teach these skills in a direct way, anyway. One perspective is that all we can actually do to improve any of these skills is to read and write proofs, on any topic, with focus and discussion, many times. This point of view expounds that experience itself begets skills in an emergent way and spending time enumerating the skills is actually counterproductive to their development. In this way, we have reached a classic tension in pedagogy.

Acknowledgements

Foremost, I would like to thank my students for their participation in this project. I also owe thanks to a former student who suggested the notion of standards based grading in an anonymous feedback poll. Finally, I am also very much indebted to Stephanie Chasteen for her tutelage in the education literature, and her very detailed feedback on an earlier draft which greatly improved this article.

References

Block, J. H. (Ed.). (1971). *Mastery Learning: Theory and Practice*. Holt, Rinehart & Winston.

Bloom, B.S. (Ed.). Engelhart, M.D., Furst, E.J., Hill, W.H. and Krathwohl, D.R. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. New York: David McKay Co Inc.

Beatty, I. (2013). Standards-Based Grading in Introductory Physics. *Journal of the Scholarship of Teaching and Learning*, *13(2)*, 1. doi:?

Boekaerts, M. (2010). The crucial role of motivation and emotion in classroom learning, in (H. Dumont, D. Istance, and F. Benavides, Eds.) *The Nature of Learning: Using Research to*

Inspire Practice. OECD Publishing. doi:10.1787/9789264086487-6-en

Duker, P., Gawboy, A. Hughes, B., & Shaffer, K. (2015). Hacking the Music Theory Classroom: Standards-Based Grading, Just-in-Time Teaching, and the Inverted Class. *Music Theory Online*, *21(1)*. http://www.mtosmt.org/issues/mto.15.21.1/mto.15.21.1.duker_gawboy_hughes_shaffer.html

Gibbs, G. & Simpson, C. (2004). Conditions Under Which Assessment Supports Student Learning. *Learning and Teaching in Higher Education*, *1*, 3–31.

Kornell, N. & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review. 14(2)*, 219–224.

Marzano, R. J. (2010). *Formative assessment & standards-based grading.* Bloomington, IN: Solution Tree.

Marzano, R. J. & Heflebower, T. (2011). Grades that show what students know. *Educational Leadership*, *(69(3)*, 34–39.

Owens, K. (2015). A Beginner's Guide to Standards Based Grading. *AMS Blogs: On Teaching and Learning Mathematics*, http://blogs.ams.org/matheducation/2015/11/20/a-beginners-guide-to-standards-based-grading/

Post, S. L. (2014). Standards-Based Grading in a Fluid Mechanics Course. *121st ASEE Annual Conference & Exposition*, Indianapolis, IN, June 15-18, 2014.

Reeves, D.B. (2003). *Making standards work: How to implement standards-based assessments in the classroom, shool, and district.* Englewood, CO: Advanced Learning Press.

Rundquist, A. (2011). Standards-based grading with voice: Listening for students' understanding. (N. S. Rebello, P. V. Engelhardt, & C. Singh, Eds.) *AIP Conference Proceedings*, *1413*, 69–72. doi:10.1063/1.3679996

Vatterott, C. (2015). *Rethinking Grading: Meaningful Assessment for Standards-Based Learning.* Alexandria, VA: Association for Supervision & Curriculum Development.

Wiggins, G., & McTighe, J. (1998). *Understanding by design.* Alexandria, VA: ASCD.

9. Appendix: Example questions for Set-Builder Notation Badge

(1) Circle those of the sets below which contain the element 0:
   - $\mathbb{Z}$
   - $\{x \in \mathbb{Z} : x \text{ is odd}\}$
   - $\{x \in \mathbb{R} : x \text{ is the square of an integer}\}$
   - $\{x^2 + 1 : x \in \mathbb{Z}\}$
   - $\{x \in \mathbb{Z} : |x - 5| < 1\}$
   - $\{x^2 \in \mathbb{N} : x^2 < 0\}$

(2) Write out the elements of the following sets:
   (a) $\{x \in \mathbb{N} : x^2 < 5\}$
   (b) $\{x^2 : x \in \mathbb{Z}, |z| < 3\}$

(3) Give set builder notation for the set { 1,3,5,7,. . . } i.e., the set of positive odd numbers.

10. Appendix: Proof Rubric

MATH 2001 Proof Gradesheet

10.1. **Writing.** Grade: 0 1 2 3 4 ungraded

This is the art of writing mathematics **for an audience**. Areas that need improvement:

(1) Complete and simple sentences, appropriately sized.
(2) Do not include extraneous information.
(3) Keep structure and language in line with logical steps.
(4) State assumptions.
(5) Introduce variables appropriately.
(6) Guide the reader.
(7) Choose notation to maximize clarity.
(8) Identify the use of hypotheses.
(9) Keep structure organized on the page and legible.
(10) Precision over vagueness.
(11) Honesty about logical gaps or imprecision.
(12) Value simplicity.
(13) Observe the established culture/etiquette.
(14) Do multiple drafts as needed.
(15) Provide all necessary information to reader.
(16) Do not include examples.
(17) Do not re-use variables, or use excess variables.
(18) Correct language for calling on a definition (do not quote definition).
(19) Remark to reader the necessary things to check.
(20) Proper left-to-right flow of equations.

10.2. **Logical Reasoning.** Grade: 0 1 2 3 4 ungraded

This is the art of correct and logical reasoning from hypothesis to conclusion. Areas that need improvement:

(1) Avoid logical errors.
(2) Justify logical steps.
(3) Choose appropriately sized logical steps.
(4) Put logical steps in linear sequence.
(5) Identify logical holes in an/your argument precisely.
(6) Identify hidden assumptions.
(7) Choose the fastest or clearest route (avoid meandering).
(8) Do not include extraneous reasoning.
(9) Avoid arithmetic errors.
(10) Correct use of contrapositive or contradiction.
(11) Do not forget cases.
(12) Avoid vagueness.
(13) Check the necessary details.
(14) Complete the argument.
(15) Do not assume what you should prove.

10.3. **Synthesis.** Grade: 0 1 2 ungraded

This is the art of combining, extending and adapting previous experience to novel problems. For this proof, the type of synthesis needed was:

(1) Combine two methods in sequence.
(2) Work with a novel definition in terms of known definitions.
(3) Invent a new method by analogy to an old one.
(4) Adjust a method to a new context.
(5) Draw conclusions from the combination of known statements.
(6) Choose appropriate concepts for a given context.
(7) Recognize a known mathematical structure in a new context.

11. Appendix: Feedback Collected

11.1. **Numerical Responses.** Students were asked to rate their level of agreement with six statements about the use of Standards Based Grading in Fall 2015 Math 2001 (1 = strongly disagree / 5 = strongly agree):

A I prefer this system (to more traditional grading systems in university).
B The system helped me study more evenly throughout the semester.
C The system helped me develop confidence.
D The system helped me learn the material more effectively.
E The system helped me get higher grades for the same level of understanding.
F The system made class less stressful.
G The standards upon which grading was based were clear, for badges.
H The standards upon which grading was based were clear, for proofs.

There was one additional question, for which the meaning of the scale varied.

I The number of opportunities to reassess was (1 = too few / 5 = too many).

Twenty-one answers were collected on questions A through F, 24 on question G, and 23 on questions H and I (26 students enrolled). The resulting data are tabulated below:

Question	1's	2's	3's	4's	5's	mean	notes
A	0	1	1	6	13	4.48	
B	0	4	2	5	10	4	
C	0	0	5	4	11	4.21	one student indicated 2.5, and one put 'N/A' which I interpreted as a '3'
D	0	0	1	9	11	4.48	
E	0	2	2	5	12	4.29	
F	0	0	3	5	13	4.48	
G	0	0	1	9	14	4.54	
H	1	3	4	6	9	3.83	
I	1	3	12	4	3	3.22	

11.2. **Text Responses.** Students were given three prompts for written responses. All responses are transcribed below.

11.3. **Question: What effects did the system have (as compared to a regular grading system)?**

(1) I knew what to focus my attention on to be prepared for each class and it helped me know what I needed to practice.
(2) Stretching the content out over the course was good for retention. Better than the usual approach of hit and move on.
(3) When I studied and kept up, I was able to understand the concepts easily. However, this system removes the stress of testing, which for me is my motivator.
(4) This system significantly reduced the amount of stress that a regular class produces.
(5) I have rather severe test anxiety, so being given multiple attempts at something relieves a lot of pressure. I think I was more able to prove what I know than had I been given a midterm and panicked.
(6) There was a uniform distribution of "work" and studying for the course. Being able to have constructive feedback on my work helped me reinforce my problem areas and in turn better understand the material.
(7) It allowed me to worry less over each assignment, and it gave me more chances to prove what I knew.
(8) Since we have more than 1 chance on the problems for the same topic, I spent more time on the one that I got wrong. And that helps me understand the material better.

(9) The system was nice because it gave you credit for understanding hte material without having to do it over and over.
(10) It gives me more chance to improve my grades so I am not very nervious [sic] of this course.
(11) It made me study more on a daily basis because of the quizzes. The lack of help with proofs hurt my grade and I thought it was unfair.
(12) This system gave me more flexibility with my schedule, as there was less consequence for lack of preparation on a particular day.
(13) Easier to acheive a good grade, far less stressful.
(14) Comparatively higher score.
(15) Quizzes felt a lot more natural and more like a part of learning.
(16) I was actually allowed to get constant feedback on my proofs, which really helped since they are so nuanced.
(17) I was allowed to miss more classes when there aren't graded quizzes.
(18) Each badges quiz was less stressful than a traditional quiz.

11.4. **What aspects of implementation could be improved?**

(1) Warning ahead of which topics will be covered by proof quizzes.
(2) More frequent updates of where we stand on the badges on D2L would be helpful.
(3) Instead of dailly quizes [sic] / theorems, a longer weekly quiz would allow for more synthesis + time to go over the previous quizes [sic]. The dayly quizes [sic] can also have a delay to get back to us, so a less often quiz would allow me to profit from the previous quiz's critizism [sic].
(4) Since I am very good at putting things off, home work that is due at the end of the week helps me study.
(5) It was fixed halfway through the semester, but getting a heads up for possible topics on proof quizes [sic] was extremely helpful.
(6) I could see how the system could be abused (i.e. study night before quiz and then never study that material again)
(7) I like the badges quiz style, but I think it is unfair to base my grade so much on proofs when they only opportunity for me to get feedback on my proof writing is when we are tested on them.
(8) More attempts for badges.
(9) less
(10) More induction!
(11) I think less chances on some of the badges would force people to study more.
(12) More chances on synthesis points for proofs.
(13) I would have preferred more time and more attempts for proof quizzes.

11.5. **Other comments.**

(1) Overall I loved the system. All math classes should be done this way.
(2) I learned a lot, thank you.
(3) I very much enjoyed this class. It made it easy and fun to learn about abstract concepts.
(4) It was really helpful with the difficult content in this class!
(5) The system's solid, I'm just an abnormally apathetic student and it facilitated my truancy. Therefore I appreciated the system. To be perfectly honest.

Department of Mathematics, University of Colorado, Campux Box 395, Boulder, Colorado 80309-0395
E-mail address: kstange@math.colorado.edu