Teaching a Class to Grade Itself using Game Theory

1 Introduction

Over the past few years, there has been a tremendous increase in the popularity of MOOCs (Massive Open Online Courses) and their importance to education as a whole. Popular MOOC systems such as Coursera or EdX are well funded, which explains their rapid growth: 60 million dollars were invested in EdX when it started in May of 2012 [8]. However, the main importance of MOOCs come from their scalability. MOOCs are able to educate massive numbers of students from anywhere in the world [3]: by the end of 2012, 1.7 million students had attended a course through Coursera [6]. The sheer number of students leads to high student-professor ratios that can reach 150,000:1 in some courses.

High student-professor ratios lead to problems for professors. Professors are simply unable to grade hundreds of thousands of submissions. Currently, two types of solutions are used to remedy the problem: automated grading and peer grading [7]. Automated grading relies on machines to grade assignments. However, machines can only check certain types of answers (i.e. multiple choice), severely limiting the depth of the questions asked [1]. Even though automated grading for written essays is an active area of research with much recent progress, the quality and accuracy of such systems is under heavy debate [4]. On the other hand, peer grading can grade any type of question, a system utilizing it could easily be "hacked" by the students [3]. These limitations render the students unable to effectively evaluate the mastery of the course material.

This work proposes several peer grading systems that provide students with an incentive not to hack the system using tools from game theory and mechanism design [5]. Our work introduces what previous attempts lack: a rigorous mathematical analysis of the system. It sets a concrete set of constraints that a grading system must satisfy to mitigate students' incentive to cheat, as well as a definitive benchmark to analyze the efficiency of various systems. Together, these create a measure for evaluating peer grading systems in terms of efficiency (how much time the professor and students spend grading) and fairness for the students (how accurate their final scores are). Having this measurement model is a large contribution in itself, because it allows comparison between different systems. Although a theoretical model cannot predict exactly what will happen in practice, game theory and mechanism design have a history of generally determining mechanisms that work in practice from ones that do not [5]. Mechanisms that do not follow game theoretic constraints may work in the short-term, but they will be exploited if possible in the long term [2].

Game theory allows for the creation of a unified system of assumptions that simulate student behavior in a class. The grading system can be viewed as a "game" where the set of "rules" are known to everyone and the "players" — the students — "play" the game in order to get the best possible result. Using assumptions and rules, game theory can predict how students will behave when they "play". This way, mechanism efficiency and effectiveness can be proved, allowing for comparison between different sets of rules.

The difficulty of the problem lies in finding an incentive to encourage students to spend time grading the assignments of others. Incentivizing students seems simple at first: let the professor give bonus points to students who grade, for example. However, this choice may not return a useful result because a clever student will just take the points and assign a generic grade. The professor cannot tell the difference between an accurately assigned 100 and randomly awarded 100. The next step would be to implement simple checking. For example, have each paper be graded by two students who both receive bonus points if their grades match. This may work at first, but clever students will learn over time to coordinate, assign everyone the same grade without doing any work, and pick up the bonus points (again, the professor cannot tell the difference between a randomly or sincerely assigned grade). These examples show that properly incentivizing graders can be a difficult task. The main purpose of this work is to devise mechanisms which avoid these problems, allowing for work to be distributed efficiently.

As demonstrated by the above examples, the difficulty of the problem lies in the complex behavior of selfish human beings. Because human nature dictates that people will do whatever is in their best interest and may try to outsmart the system, coming up with a strict set of rules to encourage the desired behavior can be quite challenging. A mechanism should be made as simple as possible, but the above examples show that a small amount of simplicity must be sacrificed to obtain a system that cannot be manipulated.

This research topic is also made interesting by the complexity of human nature, along with its huge practical value. This paper will cover the thought process of mechanisms leading up to the final mechanism and its implications. All mechanisms and related calculations will be compared in terms of scalability and effectiveness.

We start with a simple model and set of assumptions and gradually build upon this set to achieve an increasingly-realistic model. After every change in assumptions, we create a new mechanism that works effectively and efficiently. By the end of our research, we have contributed a highly realistic and detailed model that other game theorists can build off of to form their own solutions, as well as a set of realistic mechanisms that could be implemented in practice.

2   Materials and Methods

2.1   Scenario

In order to find a solution to the professor's problem, we consider the case of a single assignment. To apply our solution to an entire class, simply apply it to each assignment separately. In a "round" of grading, there are n students, who each bring one ungraded paper to the professor. These ungraded papers are the input for a function. Each of these papers, i has an objective score o_i where $0 \leq o_i \leq 100$. This is the score the professor would give out. Our mechanism then has the students grade the papers in a clever way, with the grades given out being the functions output. We identify the students by an index i from 1 to n and the professor as index 0.

2.2   Assumptions

To create the models that we use to predict the behavior of the students, we use a set process every time. To begin with, we create a set of assumptions. These assumptions are used to explain how students will act in a given situation; for example, one assumption could be that students want good grades. This is a logical assumption that is likely true, as all the assumptions are, and can be used to classify the behavior of the students in a model. The same set of assumptions can be used for multiple mechanisms, which will allow us to evaluate different mechanisms side by side. To begin with, we create this set of assumptions:

1. Students want good grades
2. Doing work makes students unhappy
3. Students want to be happy
4. Students only care about their final happiness
5. The happiness of a student is only affected by their grade and the amount of work they do

To re-express the above in mathematical terms:

1. Students all share a common happiness function, $H(g)$, where g represents the grade the students received, and the output is their happiness. w.l.o.g. $H(0) = 0$
2. Students can grade others' work accurately (i.e., they can find o_i) but doing so costs one unit of happiness. Therefore after grading W papers, a student's happiness can be expressed as $H(g) - W$
3. Students want to maximize their final happiness, expressed as $H(g) - W$
4. Students care only about the expectation of their final happiness, which, put in game theoretical terms, makes them risk neutral (i.e., if $H(x) = 10$ and $H(y) = 5$, then getting x and y each with 50 percent chance, gives a happiness of 7.5).
5. The final happiness depends only on the grade assigned to the student and the amount of effort exercised. It does not depend on the score of others

Two notes about the importance of a student's happiness and the irrelevance of the grades of a given student's friends. Happiness is important because it provides a measure of the combination of the work performed and grade received. The reasoning for not caring about friends is due to the relatively low impact they have on someone's grades and happiness. If a student has 100 friends in a class of 150,000, the chance that he will ever have the opportunity to affect the grades of his friends is tiny. For simplicity, we ignore this tiny probability and assume it is 0.

Do these mathematical assumptions perfectly capture human behavior? Of course not. But these assumptions are certainly a meaningful simplification of student behavior, and one should expect mechanisms that are provably good under these assumptions to be pretty good in practice as well.

2.3 Objective

The goal of the system is to make sure that every student gets the most happiness when they correctly grade every assignment they are given to grade. Without this property, a system becomes unpredictable. In other words, no student can deviate from the desired behavior to improve his or her happiness. In math terms, if by acting honestly student i grades w_i papers and gets a grade of g_i there should not be any way for the student to spend w_i' units of effort and get a score g_i' where $H(g_i') - w_i' > H(g_i) - w_i$.

2.4 Benchmark

A benchmark that will allow different mechanisms to be compared is vital to the iterative process of mechanism development. Mechanisms should be compared in terms of the amount of work done by a person as well as the accuracy of the grades produced. Therefore, the benchmark becomes the maximum work assigned to any student plus the maximum deviation of student assigned grades to the correct grades, so that a lower scoring mechanism produces a more desirable outcome. In mathematical terms, we want to minimize the function $\max_{i \geq 1}\{|H(g_i) - H(o_i)|\} + \max_{i \geq 0}\{w_i\}$ in the worst possible case. We call our benchmarking function the *Objective function*, which we use in our research.

When creating this benchmark, there are two metrics that must be considered: accuracy and workload. Accuracy is important because if we do not care, then we could give everyone a 100, which, as explained earlier, is a terrible mechanism. Secondly, ignoring the amount of work done means the professor grade can everything. This is bad because it makes the professor extremely unhappy.

In our case, we use the *maximum* deviation and the *maximum* work done. The reason why we do not use the *sum* of the deviations or the *sum* of work done is as follows. Consider the following two cases for the maximum deviation of the grade assigned by the mechanism to the hypothetical professor-given grade. Note that the grade the student receives at the end is not necessarily that given by their peers. Case 1 is as follows:

A Each of 1000 students should receive a 90, but the mechanism assigned a 91 to each of them. Maximum deviation is 1, sum is 1000.

B Each of 1000 students should receive a 90. One person is assigned a 0, while everyone else is assigned a 90. Maximum deviation is 90, sum is 90.

Clearly, outcome A is better than outcome B. The smaller maximum deviation for outcome A reflects this. For the maximum work done, consider the the following case:

A The professor grades 1000 papers. Maximum work is 1000, sum is 1000.

B Each of 1000 students grades 2 papers. Maximum work is 2, sum is 2000.

This time, outcome B is intuitively better than outcome A. Again, the benchmark captures our intuition.

Now that we have created a simple set of assumptions and a benchmark, we can create several mechanisms, and improve upon them until we are satisfied with a final mechanism. In the improvement process, we gradually refine our assumptions to better represent reality. This means that the original set of assumptions we proposed will change throughout the course of the paper to allow mechanisms to handle more complex and more realistic situations.

3 Results

Once the rules and assumptions have been defined and the benchmark determined, different grading mechanisms can be created and tested. This is where mechanism design and game theory comes into play. Mechanism design is used to create a system that satisfies a set of constraints and achieves a certain purpose. Then, game theory is used to mathematically predict student behavior to benchmark such a system. Starting from the simplest and most obvious approaches, we created, tested, and refined upon various mechanisms in an iterative process that spanned several months. The end result is fairly complete, although there are still questions to be answered in future work.

3.1 Calibration Mechanism

3.1.1 Process

First, we design a mechanism that does extremely well in a model defined by our simple assumptions, which we use as a starting point. One idea is to have the professor choose one paper that he grades. He then distributes this paper to every student, along with a second, ungraded paper. If the students fail to grade the paper the professor graded correctly, they get a 0. If they succeed, they receive their assigned grade. This is almost exactly the Calibration Mechanism, except we make a small modification. Considering the idea as is, a student whose own grade on the assignment is a 0 has no incentive to grade the paper:

Their grade will be a 0 whether or not they grade the paper, so why bother? To cope with this, we pick some minimum grade $m > 0$ and promise every student that they will receive a grade of at least m if they grade the assignment correctly. By appropriately choosing a value for m, all students should behave honestly and the mechanism should output accurate grades while maintaining a low amount of per-person work.

Definition 1. *(Calibration Mechanism) First, the professor pre-grades a paper P. Each student is asked to grade two papers: P and another paper. If the student fails to grade P, he or she gets a zero. Otherwise, the student gets the maximum of m and his or her assigned grade.*

Theorem 1. *If $m \geq H^{-1}(2)$, it will be in every student's best interest to grade both papers in the Calibration Mechanism.*

Proof. Consider the student i. If all other students behave as desired, student i will be assigned a grade $G = max(m, o_i)$ unless he or she does not grade P correctly. The happiness of student i grading one paper is $\frac{H(G)}{2} - 1$, because the student has a 50% chance of getting a 0 and a 50% chance of getting G. If student i grades both papers, his or her happiness will be $H(G) - 2$ because he or she is guaranteed to receive G by grading both correctly. If student i grades neither, his or her happiness will be 0, because he or she spends no energy and gets a 0. Therefore, if $H(G) > 2$, then student i will grade both papers. So, for the professor to guarantee that the class use effort and obtain the correct grades, $H(m) > 2$. Since we chose m to be greater than $H^{-1}(2)$, the previous always holds. $\square$

This process is summarized in Figure 2 on page 12.

3.1.2 Benchmark and Remaining Issues

We now show that the Calibration Mechanism achieves a low value in our benchmark.

Theorem 2. *The Calibration Mechanism achieves a benchmark score of 4.*

Proof. The teacher expends one unit of effort, each student spends 2, as long as the teacher set the minimum value m high enough. Everything above $H^{-1}(2)$ will be graded correctly, while everything below it will receive a score of $H^{-1}(2)$. So, maximum deviation is $H^{-1}(2)$, because this is the score that might be given to someone who deserves a 0. Expressed in happiness units, the value is 2. This means that the sum is $2 + 2 = 4$ $\square$

The benchmark results for the Calibration Mechanism can be seen in Figure 1 on page 11, together with results from other mechanisms.

This mechanism is extremely simple, and relies on the unrealistic assumption that students are unable to communicate with each other. If they were able to, students would be able to figure out which paper is calibrated by checking which one they all share.

3.2 Improved Calibration Mechanism

3.2.1 Process

In the calibration mechanism, we assumed that students could not communicate with each other, and therefore could not figure out which of the two papers they were asked to grade was calibrated. For this mechanism, we relax this assumption and instead provide a mechanism that works even if each student can communicate with all other students.

Definition 2. *(Improved Calibration Mechanism) In this mechanism, we apply the Calibration Mechanism on a larger scale. Every student is put into a cycle. Each student grades the papers of k next students. The professor also grades every kth paper in the cycle, which become the calibrated papers. The same system for assigning grades is used as in the calibration mechanism.*

Theorem 3. *In this mechanism, students will return accurate grades as long as $m \geq H^{-1}(k)$, even though they can communicate with each other.*

Proof. In this scheme, because of the overlapping papers, the students have no better way of determining which papers are calibrated than random guessing, even though they can all communicate with each other. As such, there are 3 options for the students. They can grade 0 papers, giving them a happiness of $H(0) = 0$ or grade all the papers they are assigned, giving them a happiness of $H(G) - k$ or they can grade i papers giving them a happiness of $\frac{i}{k} \cdot H(G) - i$. This means that it is always worse to grade i, because happiness from grading i is equal to i times happiness for grading 0 plus $i - k$ times happiness for grading k.

Even though the students can communicate with one another, they will have no way to determine which paper is calibrated because there are multiple calibrated papers that overlap, and are thus best off grading all the papers. As long as $m \geq H^{-1}(k)$, the papers will be accurately graded, as before.

$\square$

3.2.2 Benchmark and Remaining Issues

We now show that the Improved Calibration Mechanism receives a higher benchmark than the calibration mechanism.

Theorem 4. *The Improved Calibration Mechanism achieves a benchmark score of $2\sqrt{n}$.*

The amount of work done by the professor is $\frac{n}{k}$. For the students, this number is k. The maximum grade deviation is also k, for $m \geq H^{-1}(k)$. Therefore, the benchmark is $\max\{k, \frac{n}{k}\} + k$, a potentially higher value than before. A downside of the Improved Calibration Mechanism is the high amount of work being done, choosing the optimal value for k, which is $\sqrt{n}$. With a large number of people in a class, this number becomes massive. We can see that the benchmark becomes $2\sqrt{n}$.

7

The benchmark results for the Improved Calibration Mechanism can be seen in Figure 1 on page 11.

3.3   Deduction Mechanism

With this mechanism, we aim to fix the faults of the previous mechanism while retaining the same set of assumptions. Other than reducing the workload, we aim to accomplish the same goals as the improved calibration mechanism.

3.3.1   Process

Definition 3. *(Deduction Mechanism) In this mechanism, each paper is randomly assigned to 2 students. Students deduct as many points as they want, and write a justification for each of them. For each paper graded by students i and j, where i took of x_i points and j took off x_j, student i receives contribution score $\frac{x_i}{x_i+x_j}$, and j receives $\frac{x_j}{x_i+x_j}$. If $x_i = x_j = 0$ both contribution scores are $\frac{1}{2}$. Students may then choose to appeal their paper's grade if they are unhappy with it, at which point the professor grades it and gives the incorrect grader(s) a 0. To compute the final score for a student, the final grade on the assignment is calculated as follows: 2 points times their contribution score are added to $H(g)$. (Then the student's grade is converted from Happiness to an actual score).*

Theorem 5. *The addition of the contribution score being added encourages graders to honestly grade papers.*

Proof. The graders have two choices, they can either grade the papers, or not grade them.

If the grader does not grade the paper, what score should he or she give it?

If they take off any points that should not be taken off, the points will surely be refuted by the professor during the regrade. So if they take off any points, they will get a 0 on the grading assignment because of the professor's punishment. If the grader gives the paper 100, they might get 0.5 if the other grader also gives it 100. If he gives it anything else, the original grader will get 0. This is much better than a guaranteed zero, so they will give the paper 100.

If the grader does grade the paper, what score should he give it?

Obviously, if the student takes the effort to grade the paper, they will give the paper its true score, because doing anything else either forfeits a higher contribution score, or guarantees a 0 due to the professor's punishment. This means that they are guaranteed at least a contribution score of 0.5, so after grading both of their assigned papers, the graders will get a full contribution score of 1.

Now that we know what the students will do, lets look at what happens when we introduce the grading partner.

If the student's partner gives the paper a 100, which is the best thing they could do if they do not grade, then if the student does not grade the paper, they do not use any effort, and their contribution score is $\frac{1}{2}$. So the student's happiness will $H(g) + 2 \cdot 0.5 = H(g) + 1$.

If the student grades, he or she uses 1 unit of effort by default, but will get the full contribution score of 1. This means that their final happiness is $H(g) + 2 \cdot 1 - 1 = H(g) + 1$.

If the partner gives the paper its true grade: If the grader does not grade the assignment, they will give it a 100 and so the grader's contribution score is 0. So their happiness is $H(g) + 2 \cdot 0 = H(g)$.

If the grader does grade the assignment, they and their partner gave the same grades, so they get a contribution score of 0. This means that their happiness is $H(g) + 2 \cdot 0.5 - 1 = H(g)$.

In both of the above cases, the happiness output for honestly grading the paper is equal to the happiness output when not grading the paper. However, if we increase the 2 present in each equation by an infinitely small amount, the equalities become strict, and as our assumptions state that students will do whatever makes them happiest, so they will honestly grade the papers. For our purposes, because this new coefficient is greater than 2 by an infinitely small amount, we simply refer to it as 2.

$\square$

This process is summarized in Figure 3 on page 13.
The benchmark results for the Deduction Mechanism can be seen in Figure 1 on page 11.

3.3.2   Benchmark

We now show that the Deduction Mechanism receives the lowest benchmark out of our three mechanisms.

Theorem 6. *The Deduction Mechanism achieves a benchmark score of* 4.

In this mechanism, the most work is being done by the students, who each grade 2 papers, while the professor will not perform any work, because students will adhere to the dominant strategy behavior, honestly grading the papers. This means that no papers will ever reach the professor, so he or she will not have to grade any papers, making the most work done by any one person 2. Also, because every student will follow dominant strategy behavior, the maximum deviation from the objective score will be 2, since every student will output accurate grades, but also get a bonus 2 points thanks to the reward. This means that the final benchmark is $2 + 2 = 4$.

3.4   Final Assumptions

Now we have a model that accounts for a realistic world view. However, the Deduction Mechanism still has some unrealistic parts. To address this, we designed a final set of

assumptions that more effectively simulate a realistic world.

1. When grading, either $w_i = 0$ or $w_i = 1$
2. For all students, $g_i \neq o_i$
3. Students all share a common happiness function, $H(g)$, where g represents the grade the students received, and the output is their happiness. w.l.o.g. $H(0) = 0$
4. Students can grade others' work perfectly but doing so costs one unit of happiness. Therefore after grading W papers, a student's happiness can be expressed as $H(g) - W$
5. Final happiness is unaffected by other people, only affected by g_i and w_i
6. Students only care about the expectation of their final happiness, which, put in game theoretical terms, makes them risk neutral
7. For a given student in N, that student has contact with all other students in N

3.5 Comparison

Figure 1 on page 11 shows a comparison between our mechanisms and those currently in use. As shown, our mechanisms performs better than the others, aside from the Improved Calibration Mechanism. As can be seen in the figure, the benchmark scores of all traditional methods of grading — peer grading, hand grading by the professor, and automated grading — are higher than those our mechanisms. This means that traditional methods result in undesirable outcomes, either inaccurate or requiring a disproportionate amount of work.

This comparison clearly shows that the mechanisms we created are more efficient than any currently employed in modern classrooms.
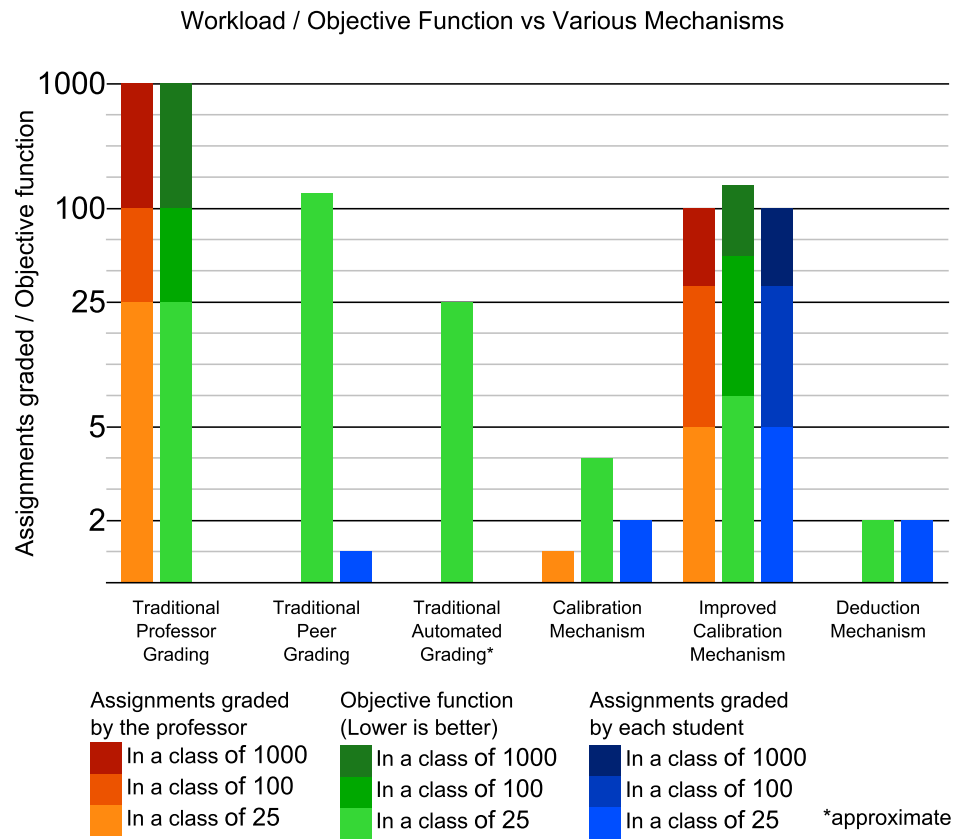
4 Illustrations

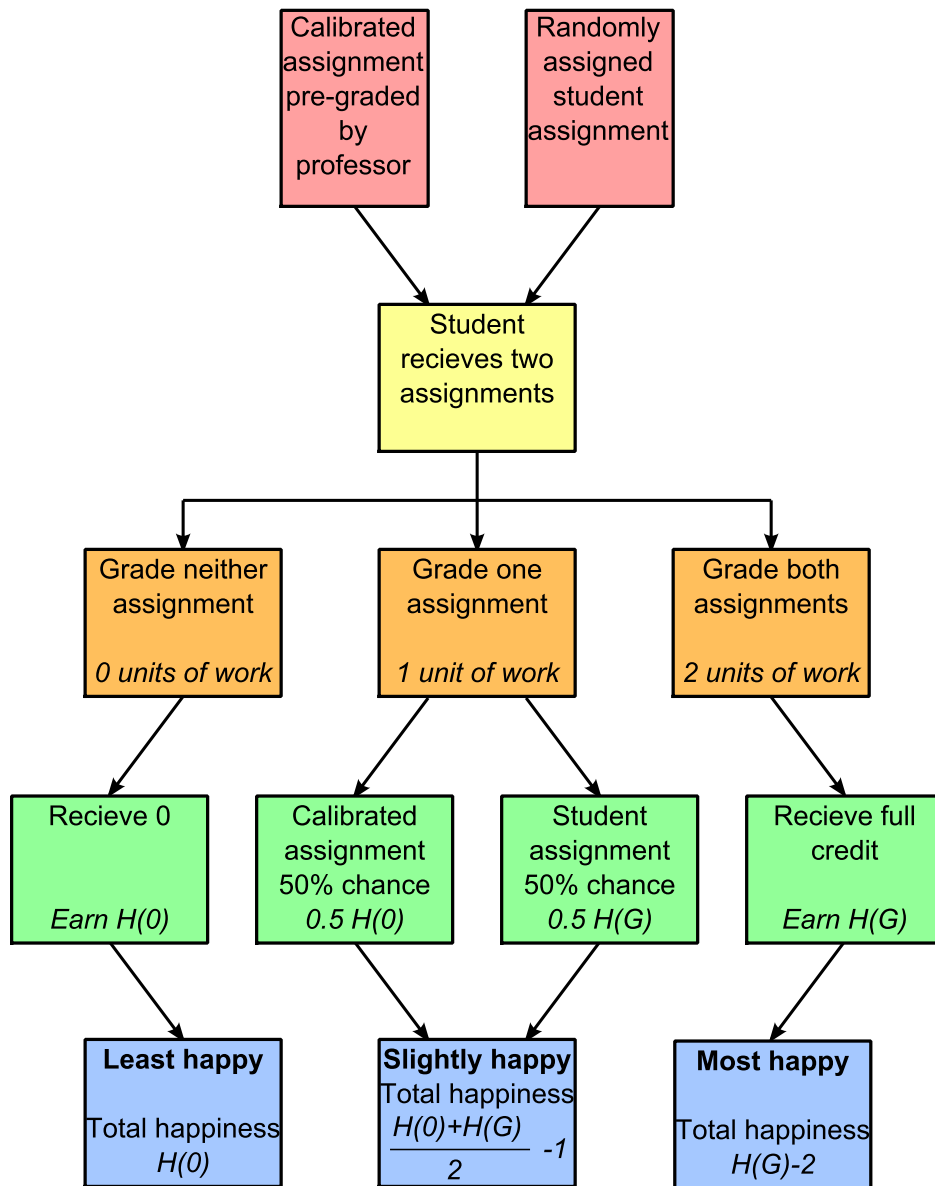Figure 1: A comparison of mechanisms in terms of effectiveness and scalability.

Figure 2: A flowchart of the Calibration Mechanism (Section 3.1, page 5) from the student perspective.
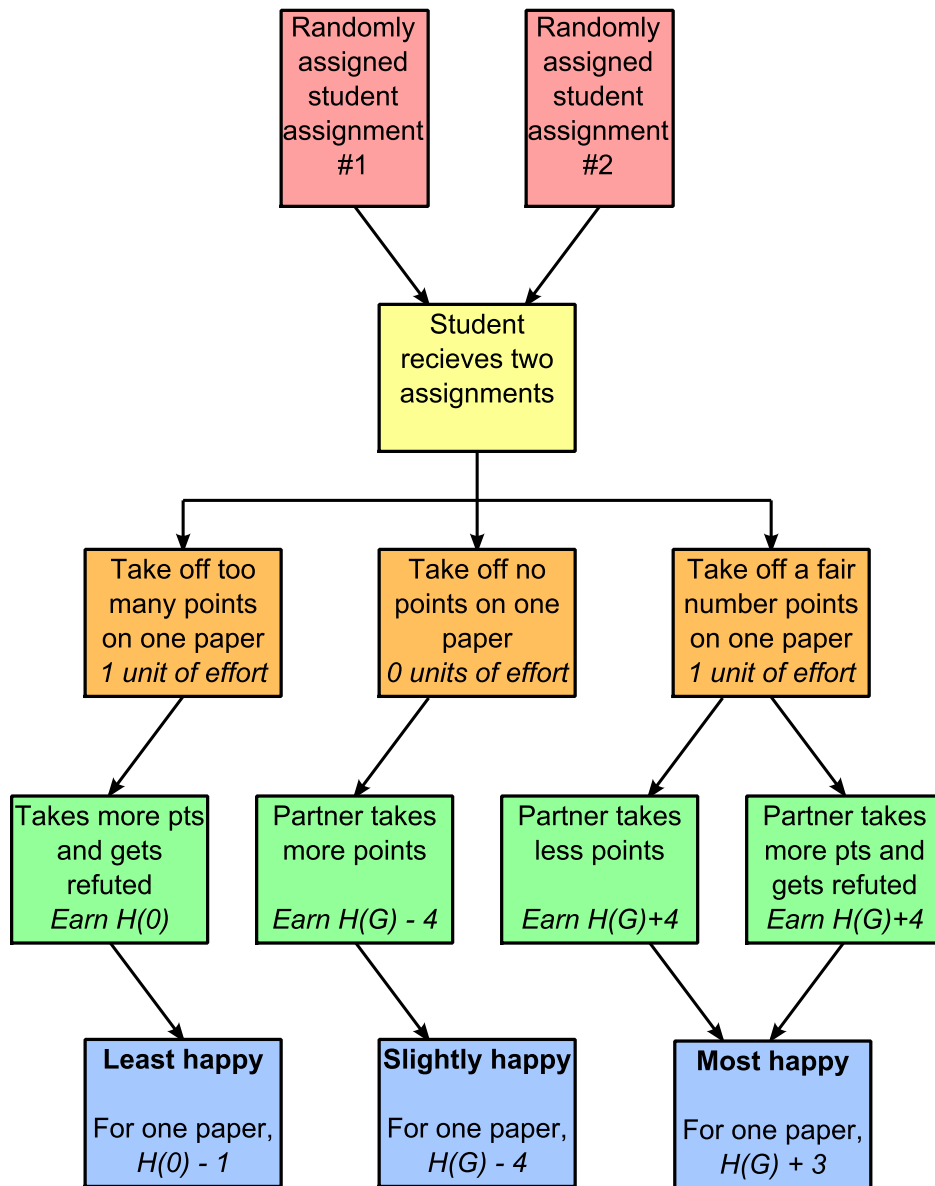
Figure 3: A flowchart of the Deduction Mechanism (Section 3.3, page 8) from the student perspective.

5 Discussion

With this research, the goal was to create a mechanism that fulfilled our goal; incentivizing students to play the "game" to correctly grade the assignments of other students. The mechanism created by the end of the research was realistic enough to be implemented in a classroom environment where all students are capable of grading others assignments efficiently.

The work done in this research furthers past solutions of grading large numbers of papers, some of which involve creating automated mechanisms that assign non-human grades. These automated solutions, as stated before, often produce unsatisfactory results. By creating a grading mechanism using game theory, we have shown that computers do not need to be relied upon to accurately and efficiently grade papers, leading the way for others to create game-theory-driven solutions to this problem.

By taking the happiness of the students into account, it can be almost guaranteed that these mechanisms are better for the students than automated grading, or any other solutions attempted before. Many student-given complaints express the lack of feedback the students can provide about their grades. This problem, found in systems such as Coursera [10], is solved by our mechanisms by allowing the students to appeal their grades and to have multiple rounds of grading. Our mechanisms address other commonly expressed problems such as inconsistency of feedback by requiring a reason for every deduction made — students must take the time to explain any deductions, or why they gave the paper a perfect score.

Traditional peer-grading schemes are prone to student manipulation: one of the issues our approach mitigates. Because such systems make grade feedback anonymous, students can easily get away with dishonestly or inadequately grading the work of others for their own benefit. Anonymity not encourage students to be honest; it gives them a reason to be *dis*honest [10]. The removal of anonymity and the application of incentives in our methods encourage positive student behavior.

Machine grading systems are known to cause unhappiness and inconsistency: one of the problems our approach addresses. Because machines always look for the same things in each paper, students who know the criteria can fool the system [9], yielding inconsistent grades. Conversely, the machine may fail to recognize alternative wordings or conceptual definitions, leading to inaccurate grades and unhappy students. The mechanisms we have created solves both these issues by using humans instead of machines.

Not only do the mechanisms developed in this research improve upon the efficiency of traditional automated and peer grading systems, but the mechanisms also fix many problems inherent in such methods. These mechanisms shed light on the alternatives to automated grading, which are more efficient and lack issues inherent in computational grading. At the moment, automated grading is not as efficient and robust as it needs to be to prevent students from creating unfair situations.

The importance of a realistic and robust model — as well as effective mechanisms — cannot be overstated. The creation of a model is essential in game theory as it provides the

outline for how the "players"of a given game will act under any set of circumstances. Many of the mechanisms mentioned would be impossible to create without a model. A well-defined model can easily be re-used and extended by others. As such, by creating a good model, this research has undergone the trouble of providing a way for anyone to predict how students will behave.

Aside from improving previous research done in this field, this research also supports the idea that many others have proposed: that peer grading is more efficient than automated grading. When EdX first introduced its automated grading system, many were skeptical of the idea, and had right to be. For this reason, systems such as Coursera have yet to adopt this, and instead default to peer grading. As many people suspected, and as this research confirms, peer grading can indeed be more efficient than automated grading when implemented correctly, simply because the artificial intelligence present in current machines cannot handle the many different ways a student can answer an open ended question.

As our mechanisms evolve, the benchmark scores yielded by each tends to decrease. However, this is not the case with the Improved Calibration Mechanism, which actually provided a higher benchmark score than the previous mechanism. This stems from the fact that the assumptions were changed, and the mechanism was not. This shows that with a change in assumptions, a new mechanism must be created to address the change, as old mechanisms will often perform badly in new models.

6 Conclusion

In this research, several novel game-theory based mechanisms are proposed to enable a class to grade itself. As there are astronomically high numbers of students in the online courses today, our solution is poised to solve the issue of grading such a large volume of assignments created in these course. Several solutions to this have been proposed, including the use of automated essay grading or multiple choice. However, the usability of automated grading is under heavy debate. Multiple choice lacks the flexibility of questions that require open-ended responses. To the best of our knowledge, our solution is the first solution of its kind, based on mechanism design and game theory.

We began our research by creating a model and a set of assumptions to mathematically express the problem, along with a benchmark to quantitatively measure the effectiveness and the scalability of various mechanisms. At first, the assumptions were relatively simple. By starting with simple mechanisms and gradually adding more realistic assumptions, we were able to design increasingly realistic mechanisms. The three we created were the Calibration, Improved Calibration, and Deduction mechanisms. Currently, the Deduction Mechanism performs relatively well in the benchmark, with reasonably realistic assumptions.

The benchmark comes in terms of a function that takes into account the happiness the students and the professor, as well as the accuracy of grades of a given mechanism. The scalability of a given mechanism can also be assessed by the objective function, as a class size grows.

The area of study is intriguing and complex, and the methods we developed are creative and innovative; to the best of our knowledge, no prior literature addresses the problem at hand in the same way that we do. As human behavior is complex, a full list of realistic assumptions is hard to create.

There still remains the question of how valid our theories are in practice. In the future, one of the experiments that would truly confirm the validity of our results would occur in the reality. Any flaws behind our mechanisms would quickly become apparent, through feedback received in classroom or online testing. After finalization, we would begin some level of real world testing. A logical next step is to extend an existing course system or classroom environment to test our ideas.

References

[1] Tony Bates. What's right and what's wrong about coursera-style moocs. *Online Learning and Distance Eductation Resources*, 2012.

[2] Zack Budryk. Dangerous curves. *Inside Higher Ed*, 2013.

[3] Sir John Daniel. Making sense of moocs: Musings in a maze of myth, paradox and possibility. *Journal of Interactive Media in Education*, 2012.

[4] Anne Herrington and Charles Moran. Writing to a machine is not writing at all. *National Writing Project*, 2012.

[5] Noam Nisan, Tim Roughgarden, E. Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[6] Laddie Odom. A swot analysis of the potential impact of moocs. *EdITLib*, 2013.

[7] Eric Randall. Edx now has software to grade your essays. *Boston Magazine*, 2013.

[8] Ruth S. Can moocs and existing e-learning paradigms help reduce college costs? *International Journal of Technology in Teaching and Learning*, pages 21–32, 2012.

[9] Theophrastus. The problems with moocs 1: Robo essay-grading. *BLT - Bible*Literature*Translation*, 2013.

[10] Audrey Watters. The problems with peer grading in coursera. *Inside Higher Ed*, 2012.