Teaching a Class to Grade Itself using Game Theory

William Wu and Nicholaas Kaashoek
Matt Weinberg and Christos Tzamos

Fourth Annual MIT PRIMES Conference
May 18, 2014

Overview

Overview

- Problem
- Model
- Benchmark
- Mechanisms
  - Calibration
  - Deduction
- Experiment
- Conclusion

Problem

Problem

MOOCs - Massive Online Open Courses

Problem

MOOCs - Massive Online Open Courses

150,000:1 Student/professor ratio

Problem

MOOCs - Massive Online Open Courses

150,000:1 Student/professor ratio

Computer grading - Limited by multiple choice

Problem

MOOCs - Massive Online Open Courses

150,000:1 Student/professor ratio

Computer grading - Limited by multiple choice

Peer grading - Hackable by clever students

Model

Model

1) Let H be a function of a student's grade, returning a student's happiness, such that $H(0)=0$.
Happiness is an arbitrary numerical unit.

Model

1) Let H be a function of a student's grade, returning a student's happiness, such that $H(0)=0$.
Happiness is an arbitrary numerical unit.

2) Students want to maximize their happiness.

Model

1) Let H be a function of a student's grade, returning a student's happiness, such that $H(0)=0$.
Happiness is an arbitrary numerical unit.

2) Students want to maximize their happiness.

3) Grading an assignment costs 1 happiness.

Model

1) Let H be a function of a student's grade, returning a student's happiness, such that $H(0)=0$.
Happiness is an arbitrary numerical unit.

2) Students want to maximize their happiness.

3) Grading an assignment costs 1 happiness.

4) Happiness is not affected by external factors, such as the grades of peers.

Model

1) Let H be a function of a student's grade, returning a student's happiness, such that $H(0)=0$.
Happiness is an arbitrary numerical unit.

2) Students want to maximize their happiness.

3) Grading an assignment costs 1 happiness.

4) Happiness is not affected by external factors, such as the grades of peers.

5) Students can communicate with their peers.

Model - New Assumptions

Model - New Assumptions

6) Students are not perfect graders.

Model - New Assumptions

6) Students are not perfect graders.

7) There is no such thing as partial-grading.

Model - New Assumptions

6) Students are not perfect graders.

7) There is no such thing as partial-grading.

8) Students can report their level of uncertainty when they grade. Let this factor be equal to U.

Model - New Assumptions

6) Students are not perfect graders.

7) There is no such thing as partial-grading.

8) Students can report their level of uncertainty when they grade. Let this factor be equal to U.

9) More effort spent in grading lowers uncertainty.

Model - New Assumptions

6) Students are not perfect graders.

7) There is no such thing as partial-grading.

8) Students can report their level of uncertainty when they grade. Let this factor be equal to U.

9) More effort spent in grading lowers uncertainty.

10) When a student assigns a grade G, the chance of the grade being N off from the actual grade is proportional to U.

Benchmark

Benchmark

A *numerical score* defined by

maximum work done by any person +
highest possible error in grading.

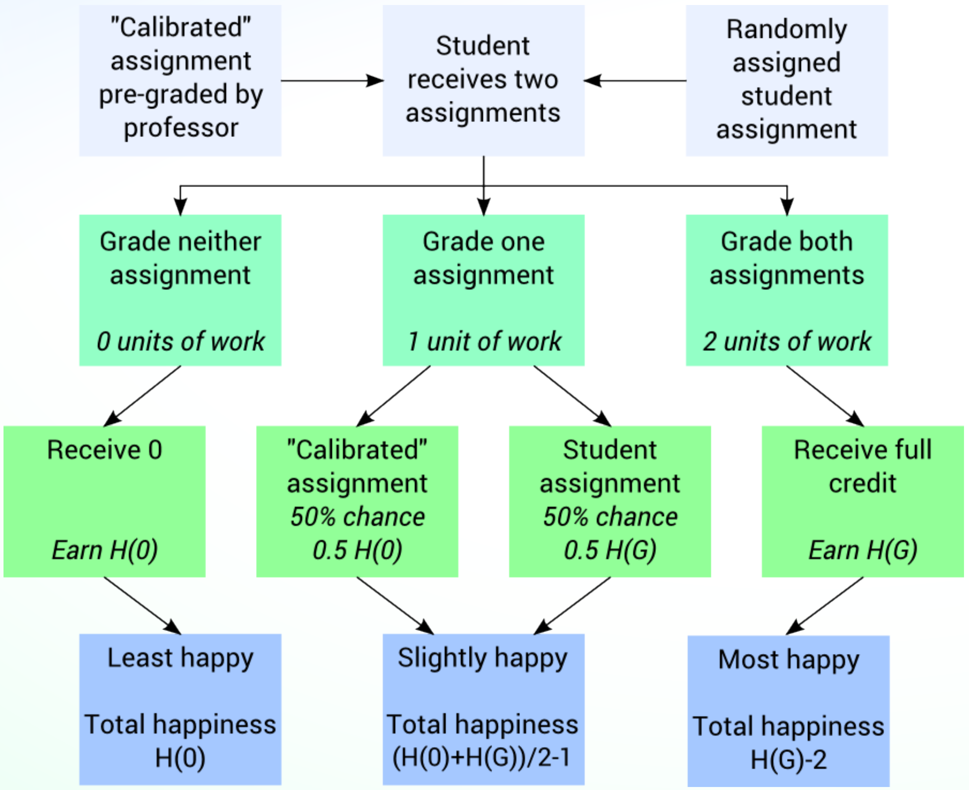
Benchmark

A *numerical score* defined by

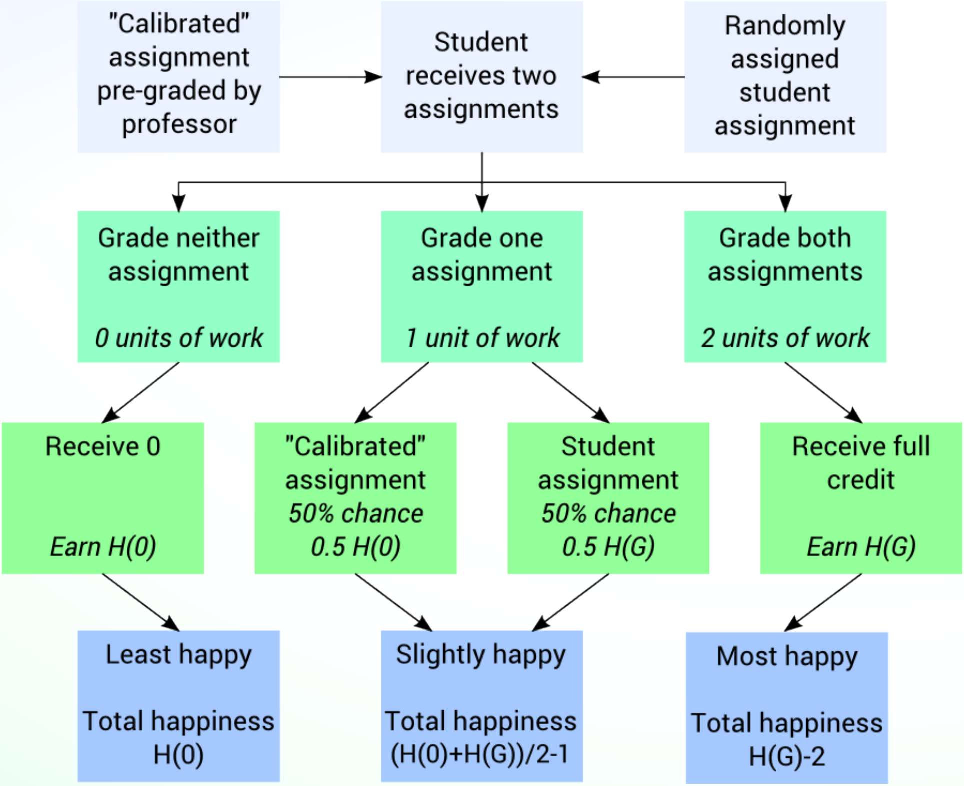
maximum work done by any person +
highest possible error in grading.

$$max_{i\geq 1}\{|H(g_i)-H(o_i)|\} + max_{i\geq 0}\{w_i\}$$

Mechanisms - Calibration



Mechanisms - Calibration

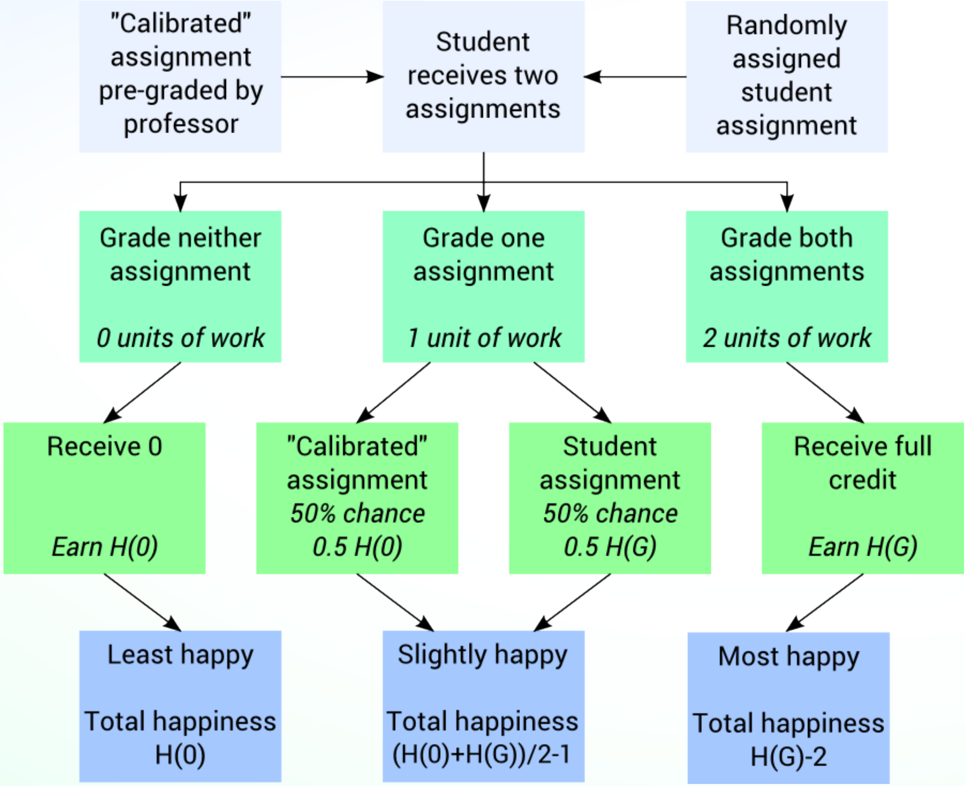


Max work: 2

Max error: 2

Benchmark Score: 4

Mechanisms - Calibration



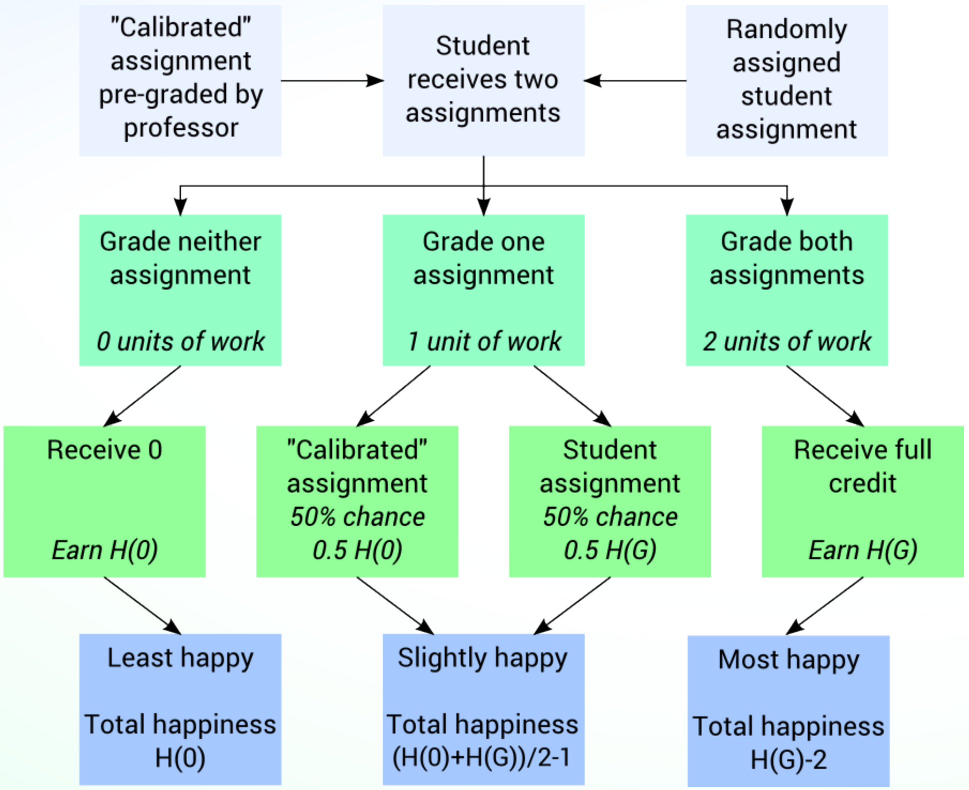
Max work: 2

Max error: 2

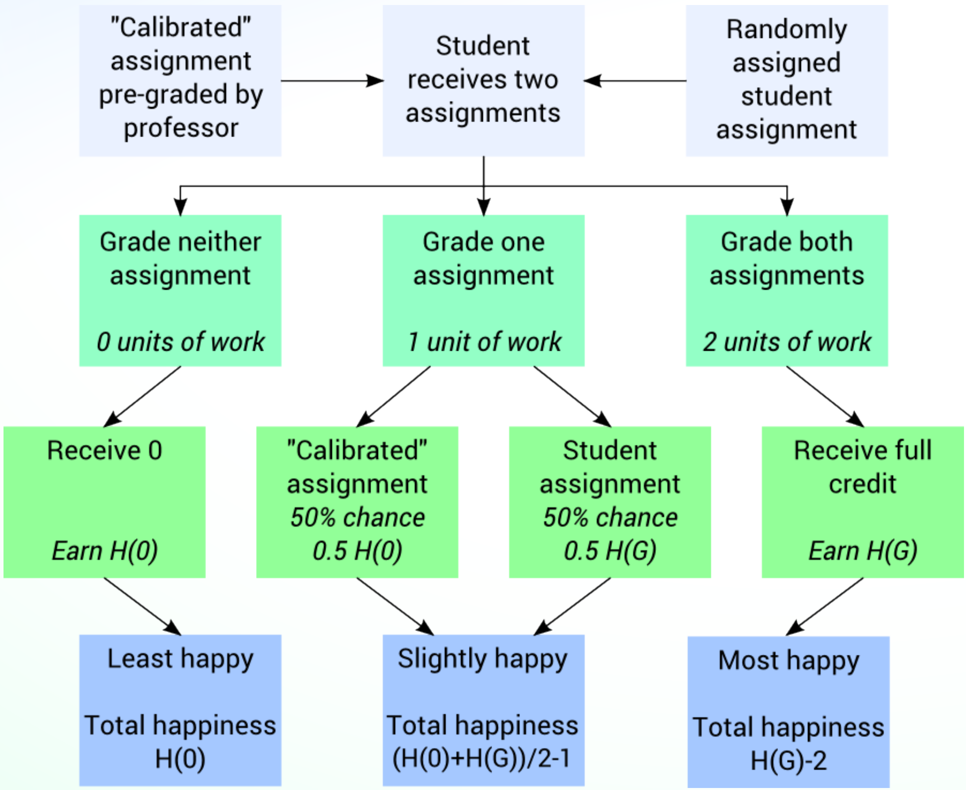
Benchmark Score: 4

What if students can communicate?

Mechanisms - Improved Calibration



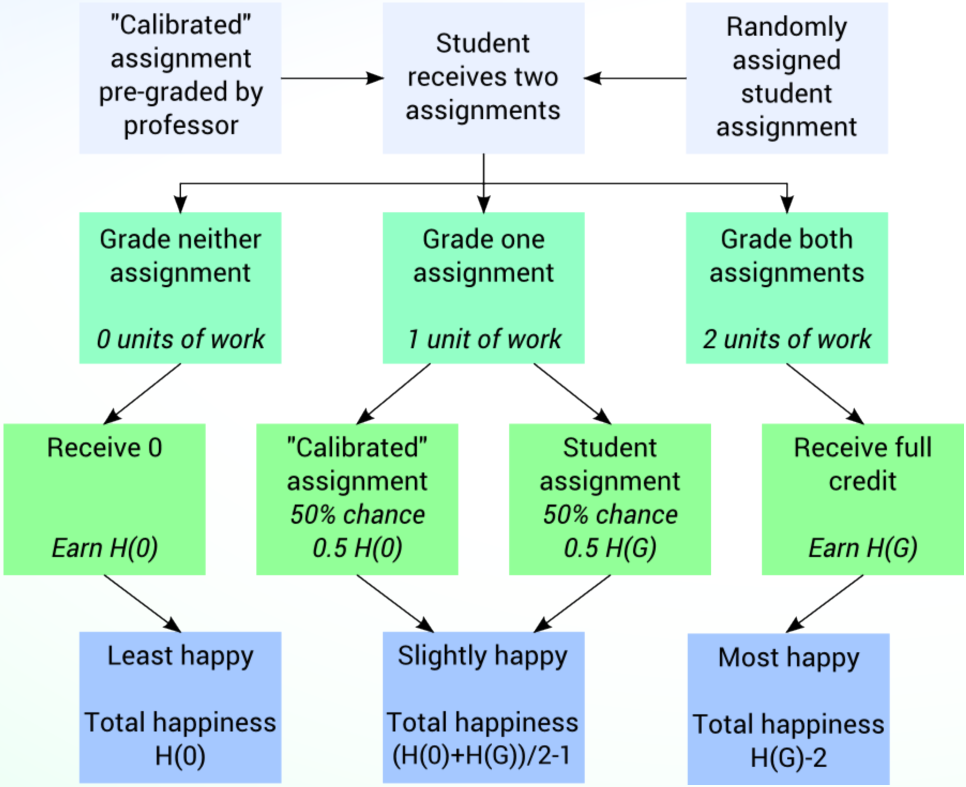
Mechanisms - Improved Calibration



Assumption added:

5) Students can communicate

Mechanisms - Improved Calibration

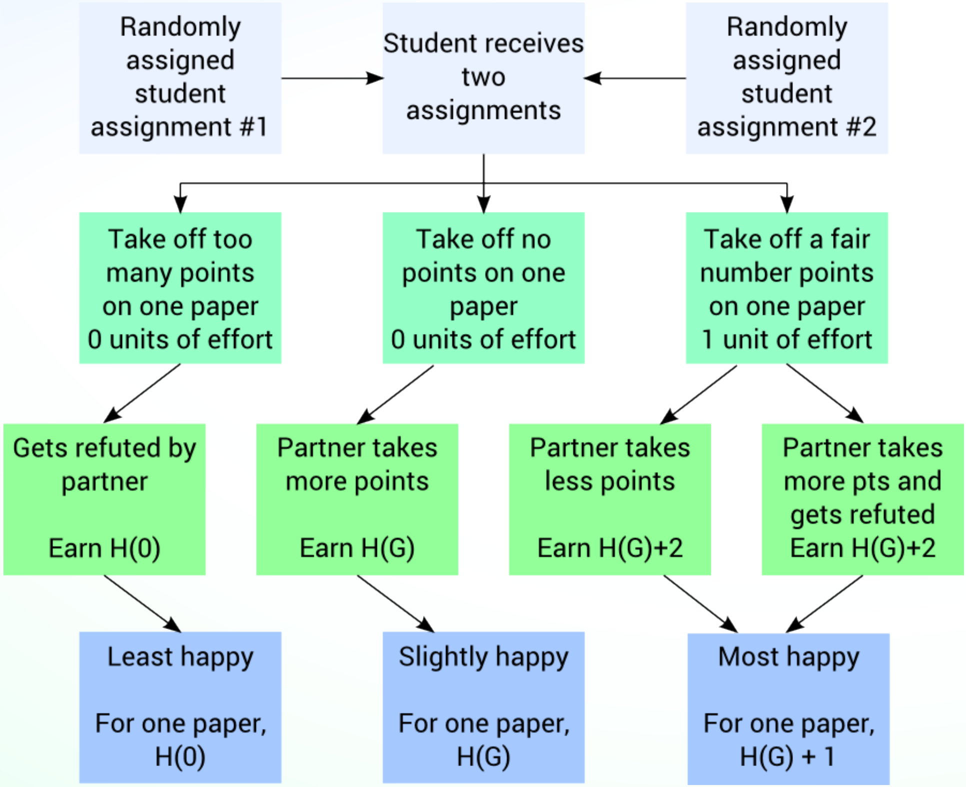


Assumption added:

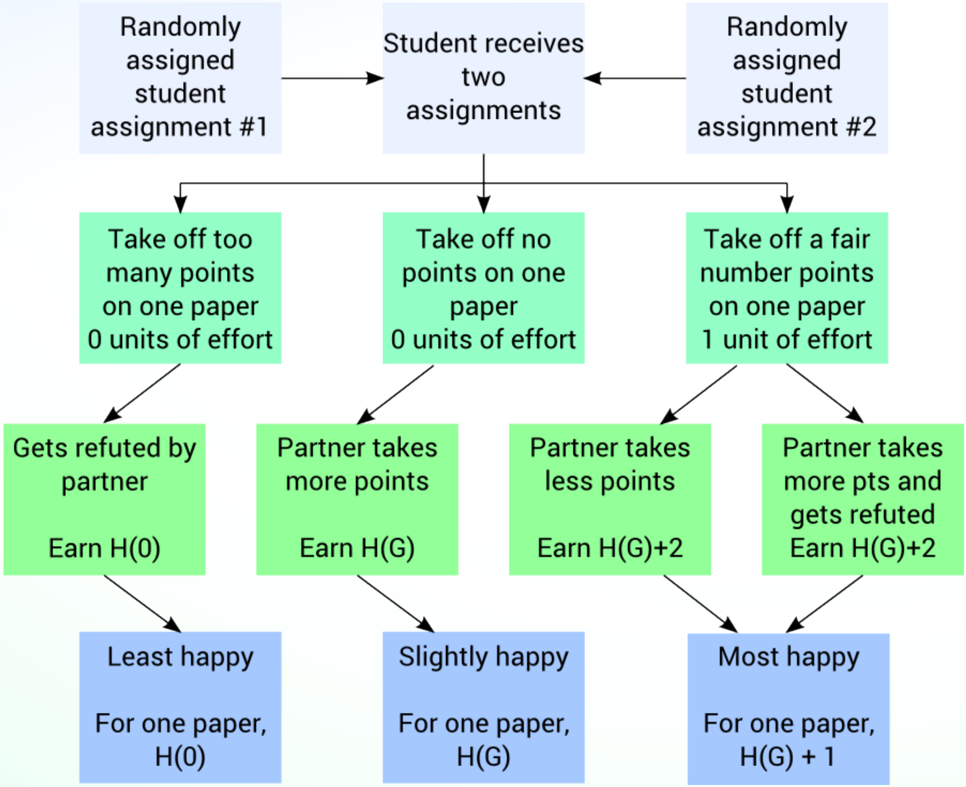
5) Students can communicate

"Improved" with multiple calibrated assignments

Mechanisms - Deduction



Mechanisms - Deduction

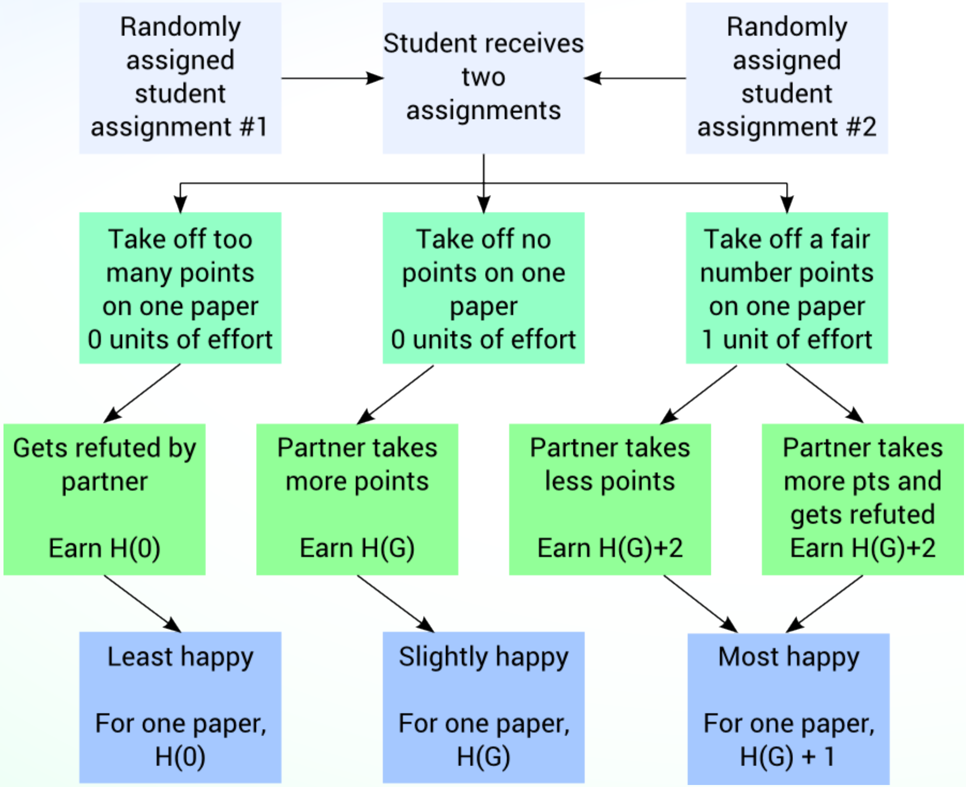


Max work: 2

Max error: 0

Benchmark Score: 2

Mechanisms - Deduction



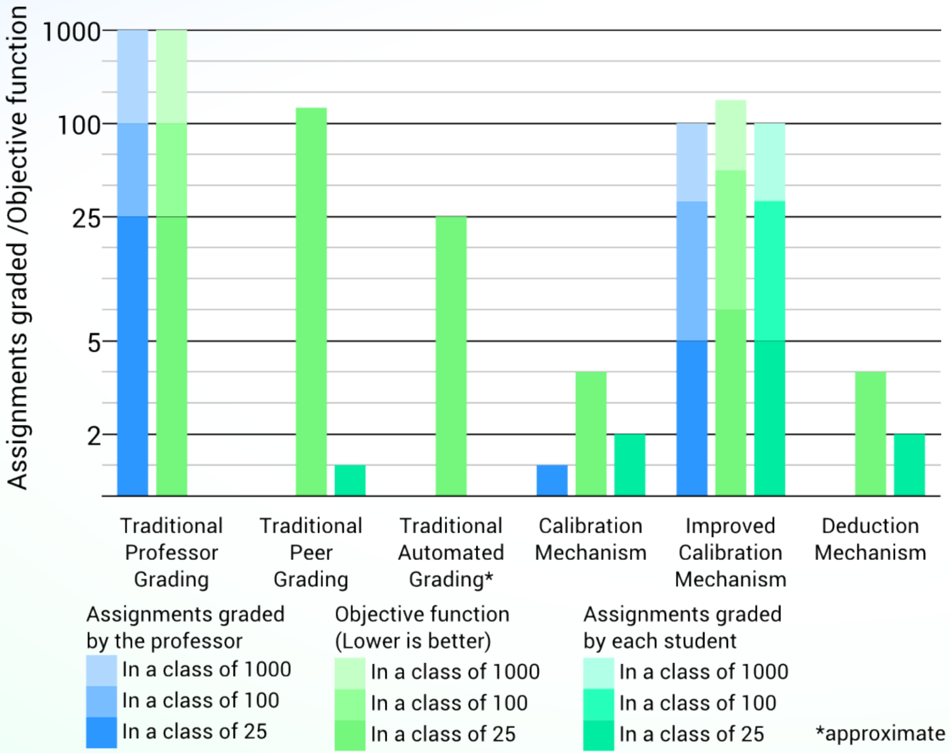
Max work: 2

Max error: 0

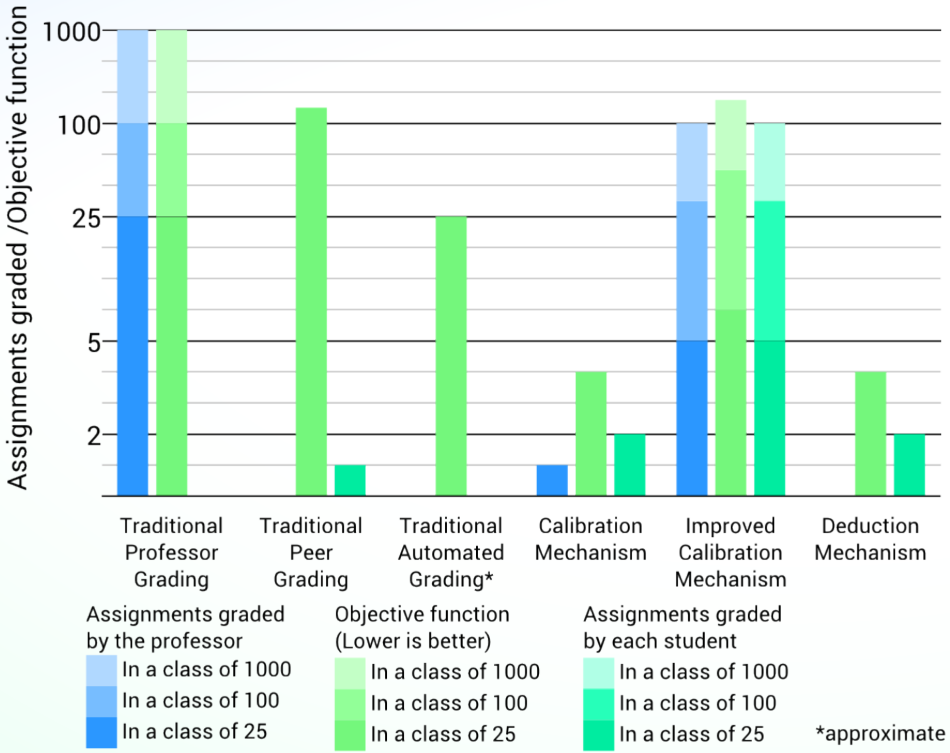
Benchmark Score: 2

Unfriendly competition

Mechanisms - Comparison



Mechanisms - Comparison



Calibration and Deduction outperform existing mechanisms

Experiment

Online, crowdsourced, and anonymous

Experiment

Online, crowdsourced, and anonymous

Designed to validate Calibration Mechanism:

Experiment

Online, crowdsourced, and anonymous

Designed to validate Calibration Mechanism:

Presented two assignments to grade,

Rewarded on one assignment

Experiment

Online, crowdsourced, and anonymous

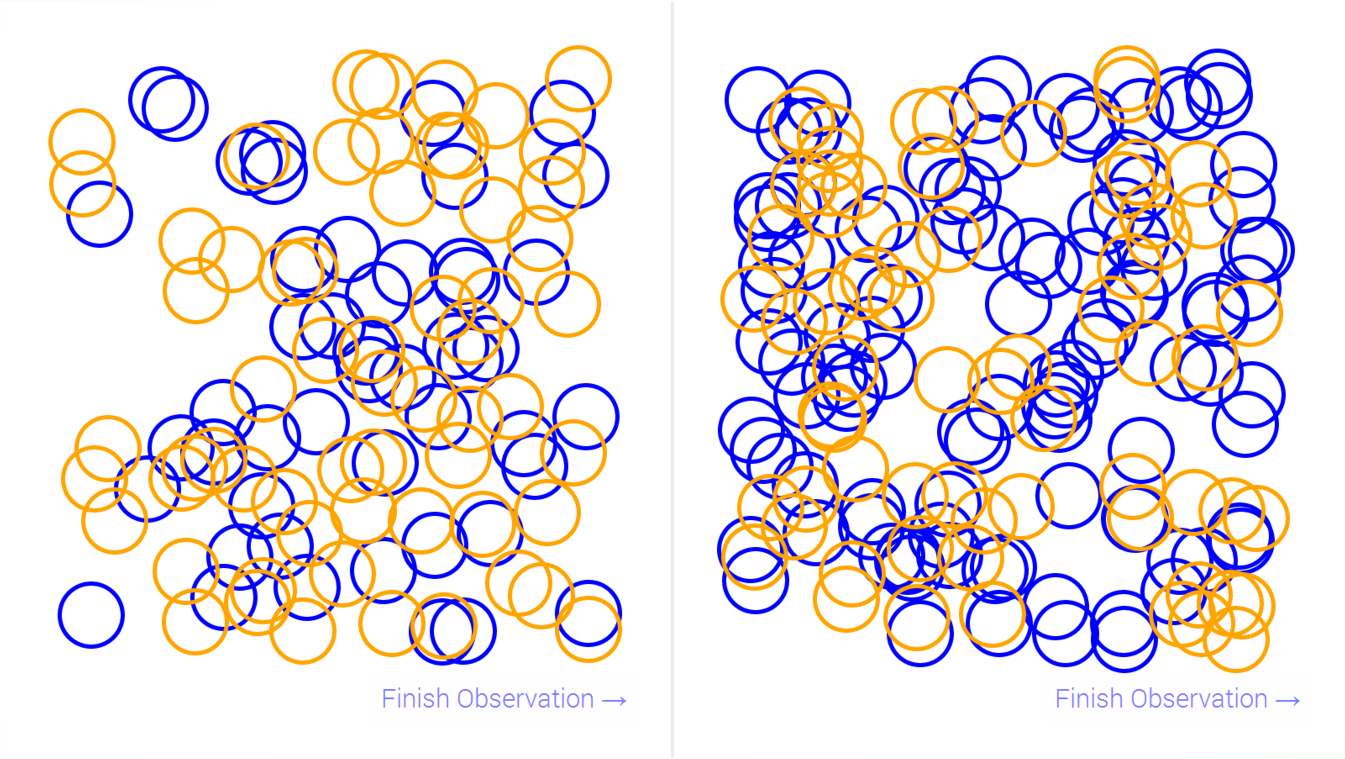
Designed to validate Calibration Mechanism:

> *Presented two assignments to grade,*
>
> *Rewarded on one assignment*

Assignment - A set of “marbles”

Grading - Counting the orange “marbles”

Experiment - Screenshot



Experiment - Reward

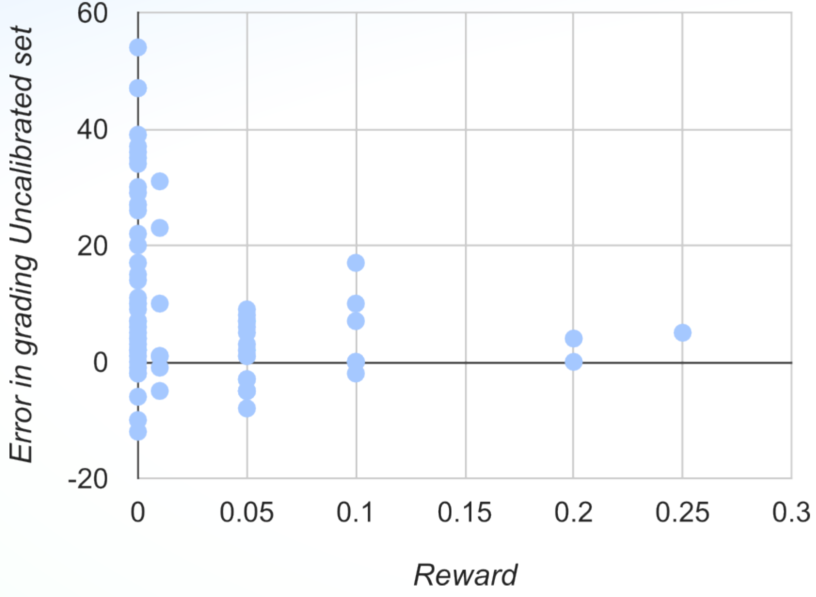
Confidence	Within	Reward
1	1 marble	$0.25
2	2 marbles	$0.20
5	5 marbles	$0.10
10	10 marbles	$0.05
20	20 marbles	$0.01

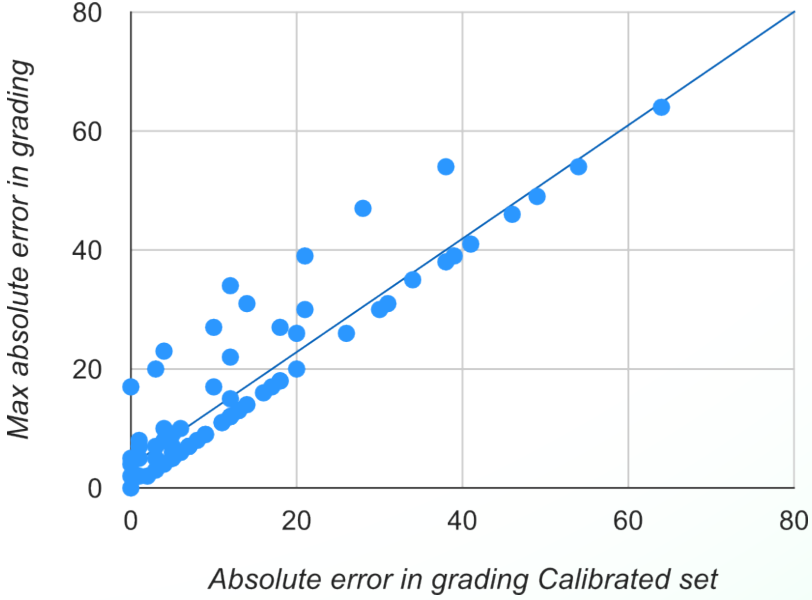
Experiment - Reward

Confidence	Within	Reward
1	1 marble	$0.25
2	2 marbles	$0.20
5	5 marbles	$0.10
10	10 marbles	$0.05
20	20 marbles	$0.01

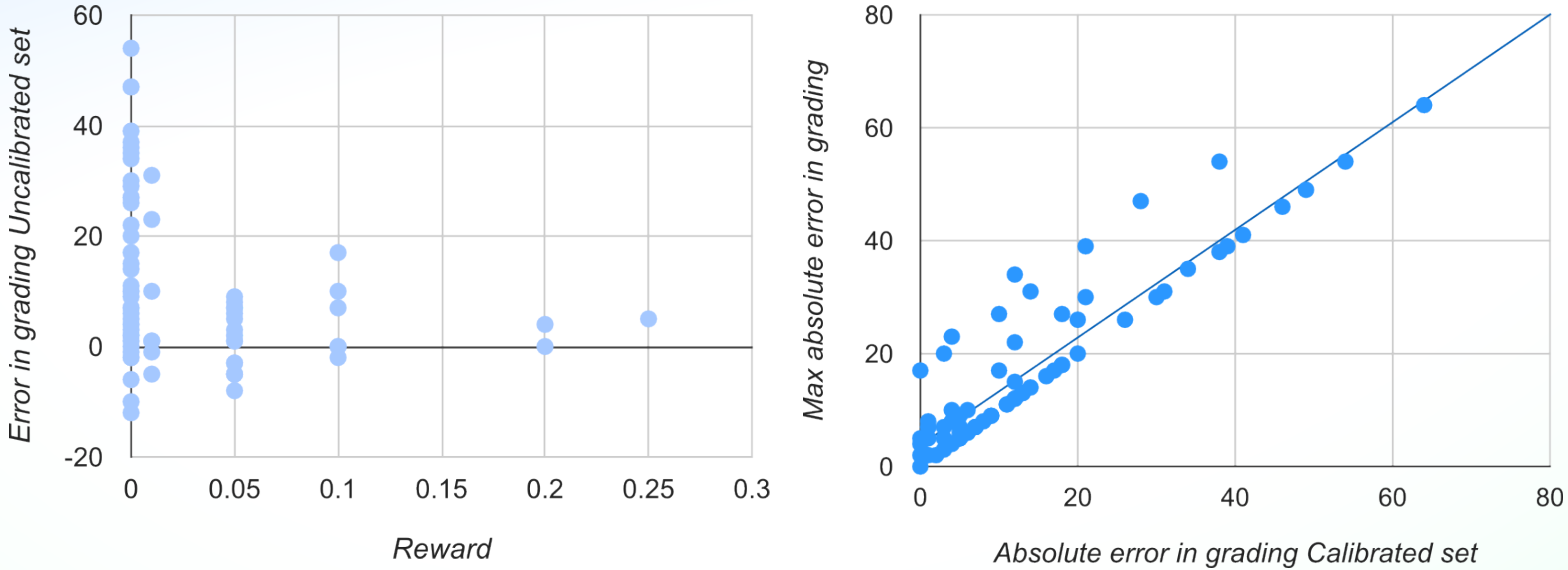
Reward is based on the reported confidence and the accuracy of the reported guess

Experiment - Data



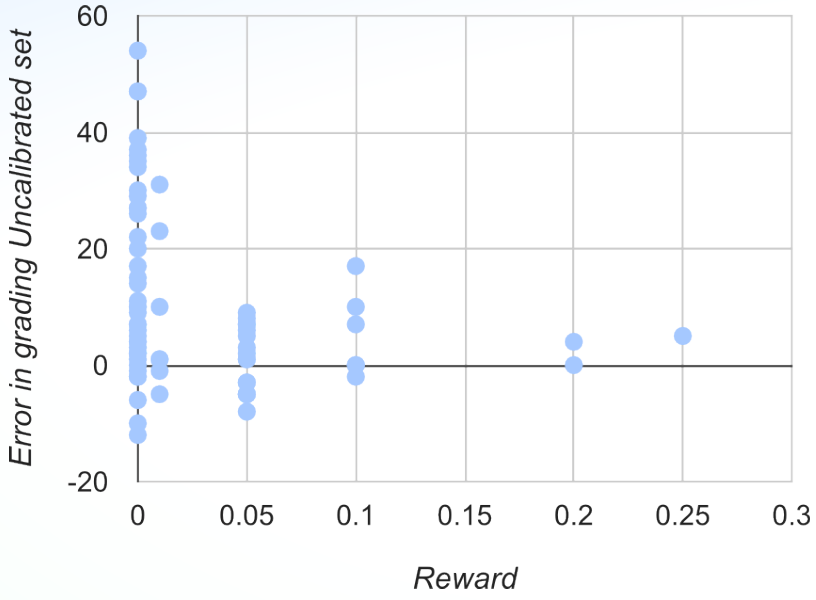


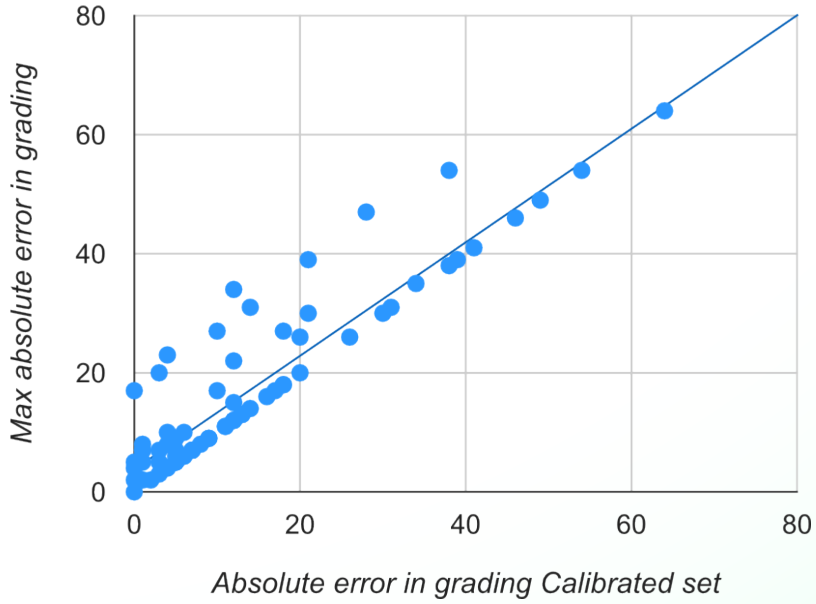
Experiment - Data



1. Greater reward → lower uncalibrated error

Experiment - Data





1. Greater reward → lower uncalibrated error
2. Calibrated set indicates grading proficiency

Conclusion

Conclusion

- Student model - approximations for student behavior

Conclusion

- Student model - approximations for student behavior
- Benchmark - score measuring efficiency and workload of various mechanisms

Conclusion

- Student model - approximations for student behavior
- Benchmark - score measuring efficiency and workload of various mechanisms
- Calibration, Improved Calibration, and Deduction mechanisms developed

Conclusion

- Student model - approximations for student behavior
- Benchmark - score measuring efficiency and workload of various mechanisms
- Calibration, Improved Calibration, and Deduction mechanisms developed
- Calibration validated by a crowdsourced experiment

Conclusion

- Student model - approximations for student behavior
- Benchmark - score measuring efficiency and workload of various mechanisms
- Calibration, Improved Calibration, and Deduction mechanisms developed
- Calibration validated by a crowdsourced experiment
- Calibration and Deduction mechanisms outperform existing grading solutions

Conclusion - Next Steps

Conclusion - Next Steps

- Improving realism - producing accurate grades from incompetent graders

Conclusion - Next Steps

- Improving realism - producing accurate grades from incompetent graders
    - Proficiency test
    - Using multiple graders to reduce error

Conclusion - Next Steps

- Improving realism - producing accurate grades from incompetent graders
    - Proficiency test
    - Using multiple graders to reduce error
- Implementation

Conclusion - Next Steps

- Improving realism - producing accurate grades from incompetent graders
  - Proficiency test
  - Using multiple graders to reduce error
- Implementation
  - User testing with Mechanical Turk
  - Eventually in Coursera / EdX

Acknowledgements

MIT

MIT PRIMES program,

Slava Gerovitch, Tanya Khovanova, Srini Devadas

Mentors, Matt Weinberg and Christos Tzamos

Professor, Costis Daskalakis

Parents