Learning at Scale

- Assessment at Scale

Bingyu Shen
May 28, 2019

Peer and Self Assessment in Massive Online Classes

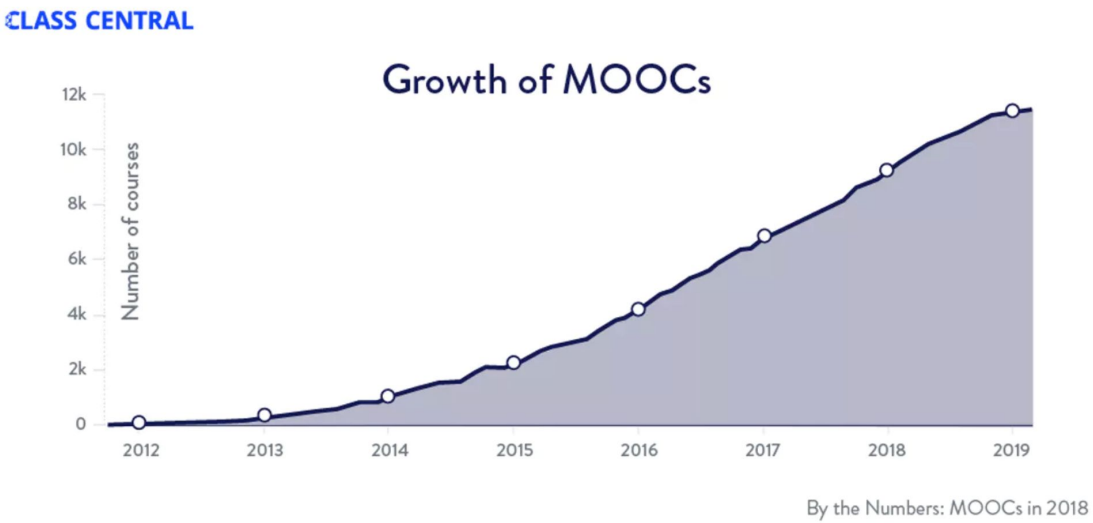
Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, Scott R. Klemmer, TCHI'13

Learning goals

- Understand pros and cons for peer assessment
- How to evaluate the accuracy of peer assessment
- Approaches to improve accuracy of peer assessment

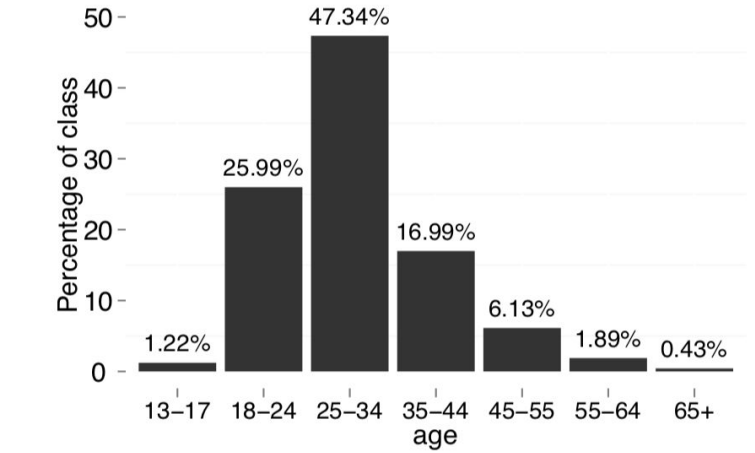
Peer Assessment in MOOCs

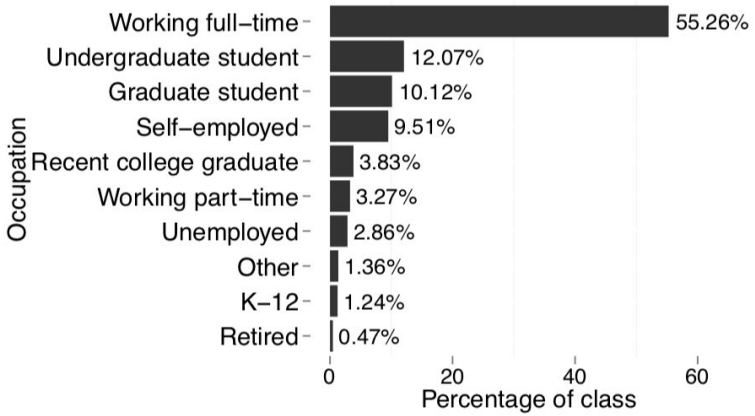
- Challenges in online MOOCs => peer assessment
- Potential issues with peer assessment?



Study setup

- Online Stanford HCI class
- 35,081 watched videos
- 2788 submissions first assignment





Problem #1:

How to establish the rubrics of grading?

Original

Table I.

Guiding questions	Bare minimum	Satisfactory effort & performance	Above & Beyond
...			
Alternate redesign—Extra credit. Have you created a fully functional alternate prototype?	0: No URL to functional prototype	3: URL present, but prototype only partially functional.	5: URL present, Alternative prototype is complete.
User testing. Photographs—extra credit. Did you submit photos from all three user testing sessions?	0: No photographs were uploaded.	3: Some photographs were uploaded (but less than 3), OR photos don't show an interesting moment in the experiment (e.g. photograph of participant signing consent form is not an interesting photo).	5: At least 3 photographs are uploaded and all photographs show interesting moments in the evaluation. Photos have meaningful captions
...			

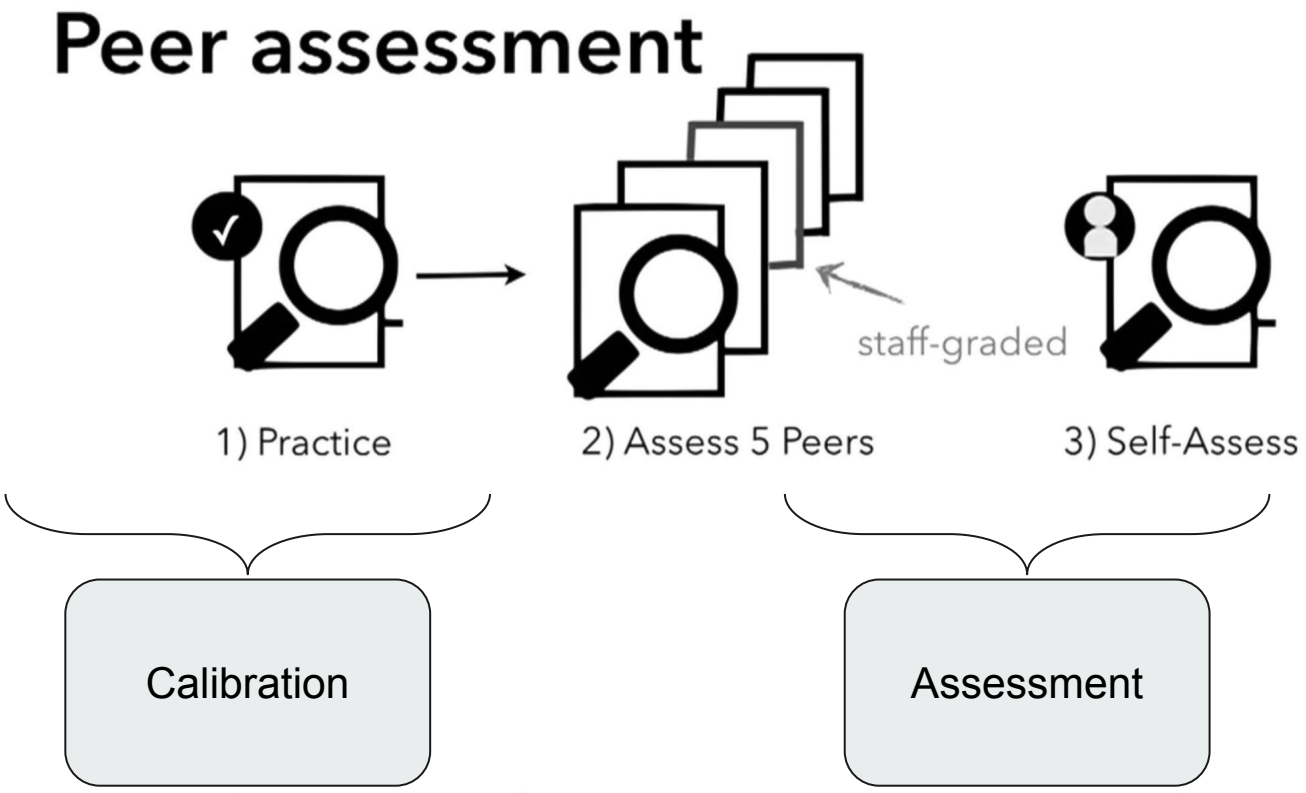
Revised

Category	Unsatisfactory	Bare minimum	Satisfactory effort & performance	Above & Beyond
...				
Extra Credit: Electronic Prototype of Redesign	0: No URL to functional prototype	1: The prototype is incomplete and barely interactive.	3: The prototype is somewhat interactive, but not ready for user testing.	5: The alternative prototype is fully interactive and ready for user testing.
Photos/Sketches	0: No photographs were submitted that showed interesting moments in the user testing process.	1: 1 photograph was submitted that showed an interesting moment in the user testing process.	3: 2 photographs were submitted that showed interesting moments in the user testing process..	5: 3 or more photographs were submitted that showed interesting moments in the user testing process.
...				

Problem #2:

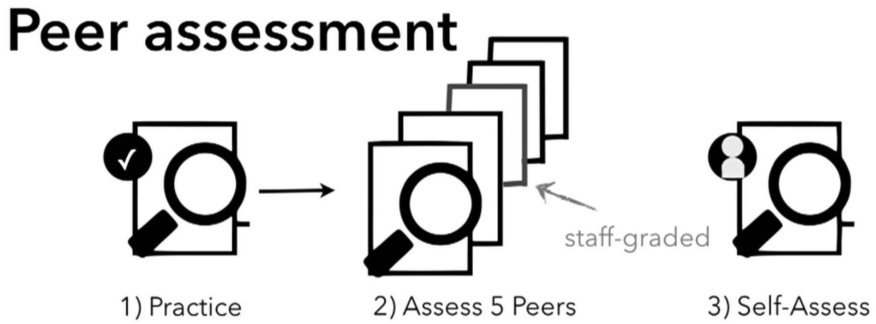
How to design the grading process?

Two phases

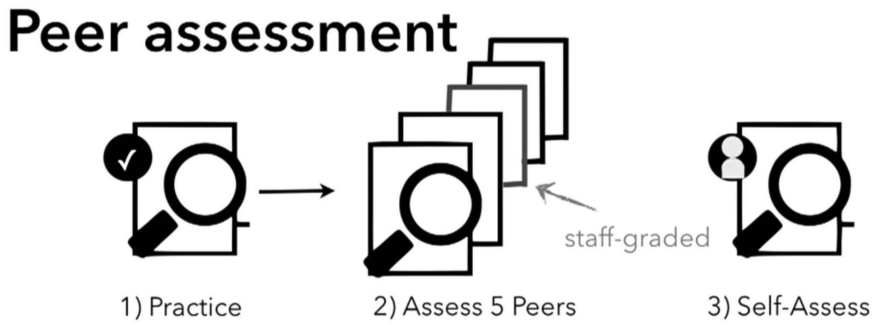

Peer assessment
staff-graded
1) Practice
2) Assess 5 Peers
3) Self-Assess
Calibration
Assessment
1. Do assignment
2. Learn to evaluate
3. Evaluate your classmates
4. Evaluate yourself
5. See results

Discussion (2 min, group of 2)

- *In what ways are peer and self assessment useful respectively?*
- *What's the point of putting self- assess after peer-assess?*



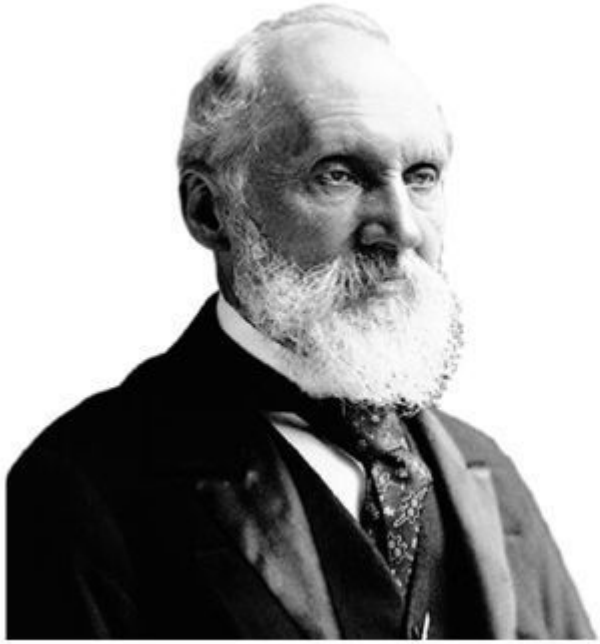
Question: How to calculate the final score?



- Median of peer assessment scores
- Self-assessment scores?

Problem #3

How to measure accuracy?



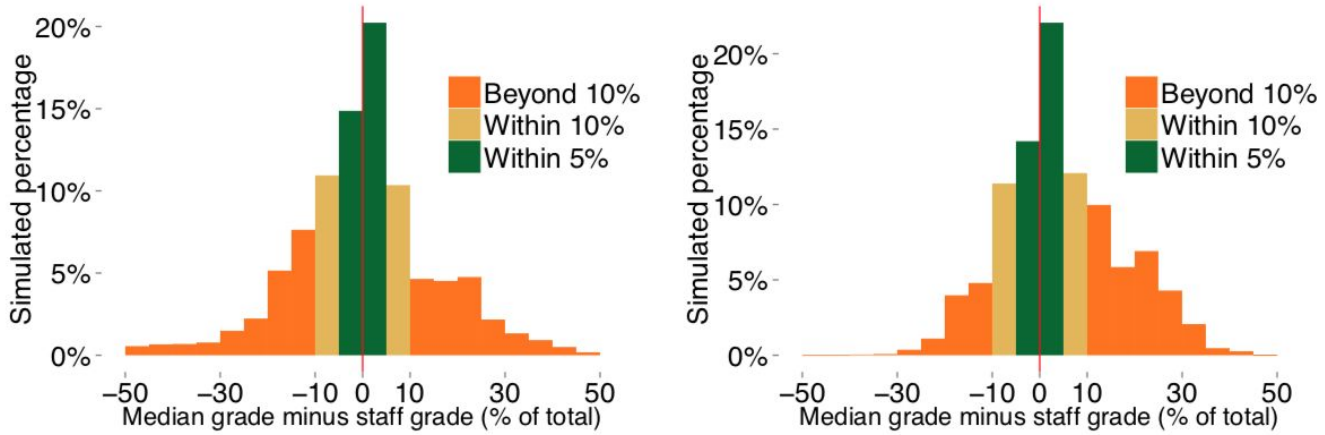
To measure
is to know.
If you can not
measure it,
you can not
improve it.
- Lord Kelvin

Methodology

- What is the ground truth?
  - Several staff-graded assignment.
  - Median grade

- Using samples with staff grade to measure accuracy
- Median score comparison with self grade

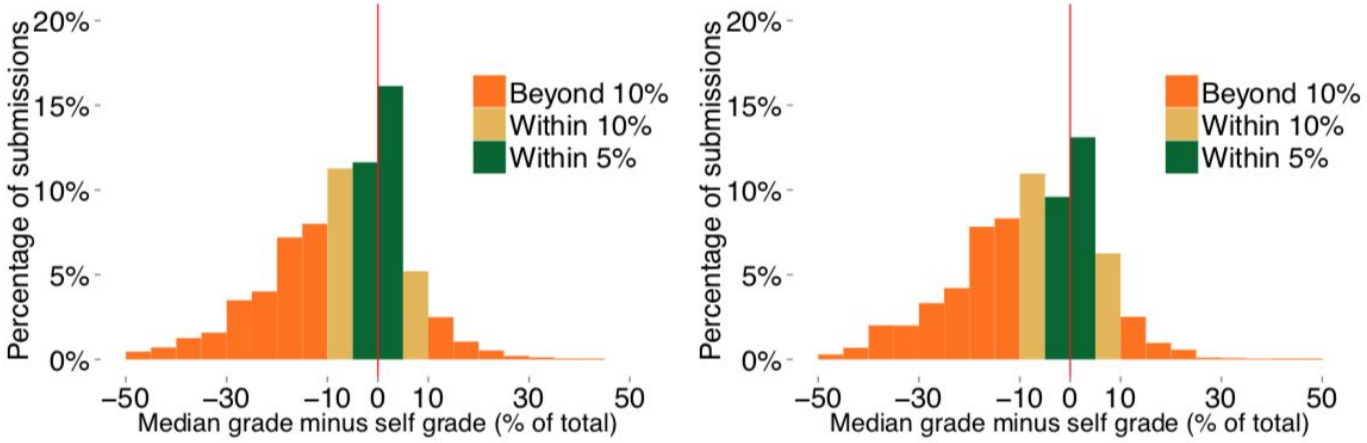
Accuracy with sampling



(a) Iteration 1: 34.0% of samples within 5% of the staff grade, and 56.9% within 10%.

(b) Iteration 2:. 42.0% of samples within 5% of the staff grade, and 65% within 10%.

Accuracy with median & self grade



Problem #4

How to improve the accuracy?

Improve Accuracy & Provide Qualitative Feedback

1. Providing Feedback (staff to grader)
2. Fortune Cookies - qualitative feedback (grader to peer)
3. Data-driven Rubric Revisions

Feedback

- About 800 participants
- Two conditions between-subject
  - No-feedback control
  - Feedback

You graded your peers' work a little low on Assignment 4. The grading rubrics are useful if you're unsure about what scores you should assign.

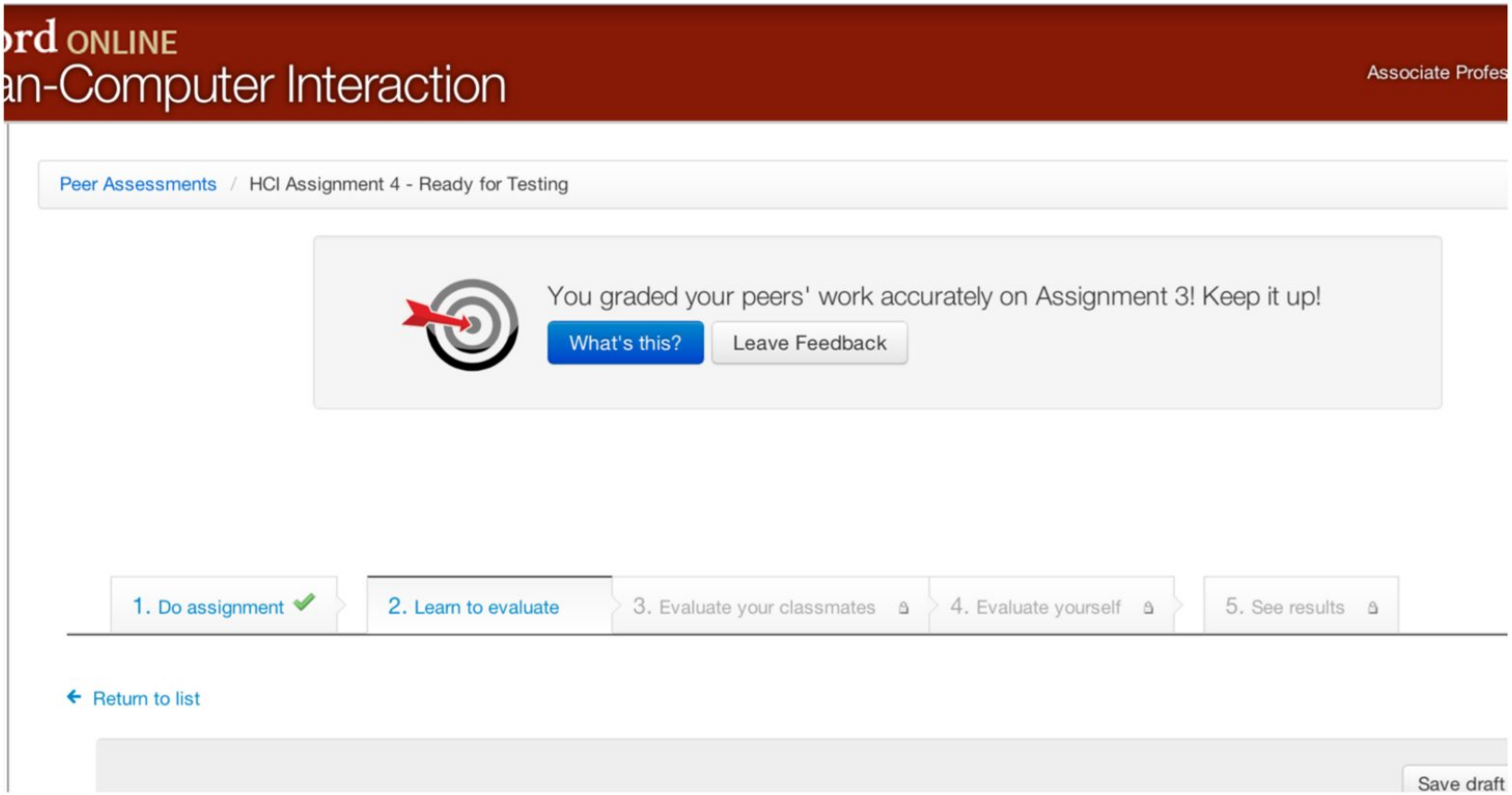
What's this? Leave Feedback

You graded your peers' work a little high on Assignment 4. The grading rubrics are useful if you're unsure about what scores you should assign.

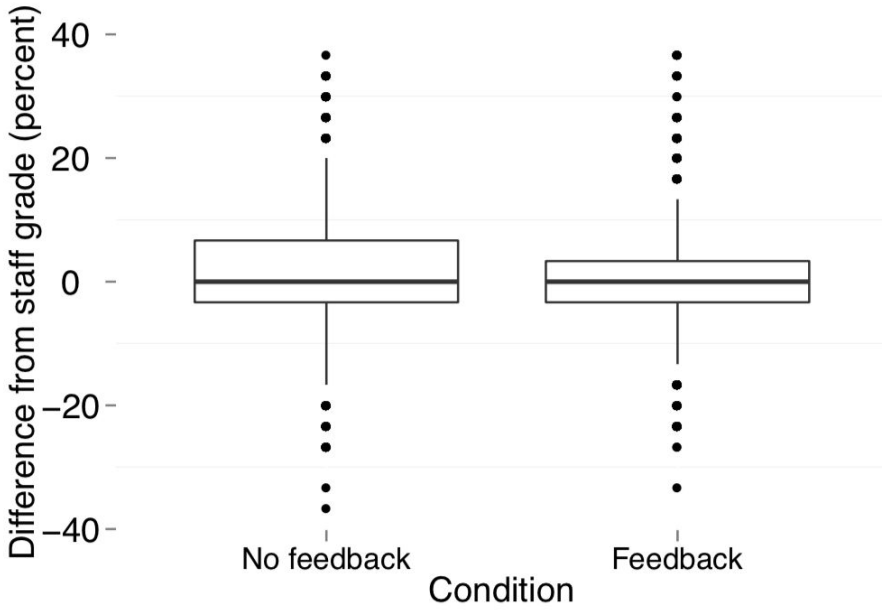
What's this? Leave Feedback

You graded your peers' work accurately on Assignment 4! Keep it up!

What's this? Leave Feedback


ord ONLINE
an-Computer Interaction
Associate Profes
Peer Assessments / HCI Assignment 4 - Ready for Testing
You graded your peers' work accurately on Assignment 3! Keep it up!
What's this?
Leave Feedback
1. Do assignment
2. Learn to evaluate
3. Evaluate your classmates
4. Evaluate yourself
5. See results
Return to list
Save draft

Feedback result



Provide qualitative feedback - Fortune cookie

- Grader to peer
  - Do not cost too much time
  - (reduce feedback cost for grader)

- Rubrics Limitations
  - Where students did poorly?
  - How to improve

Overall evaluation/feedback

Note: this section can only be filled out during the evaluation phase.

Overall feedback:

How could this student best improve his/her submission? From among the following, copy one or more pieces of advice that would help the student. Paste your advice in the feedback box below.

- Clarify the concerns, goals, and expectations of the user tests.
- Make the user tests more structured.
- Make the user tests more consistent across participants.
- Make the prototype more interactive so the user test represents a more real-life interaction.
- Determine the implications of the user succeeding (or not) on each task on the prototype.
- Make fewer assumptions about users/Reduce bias in user test.
- Other

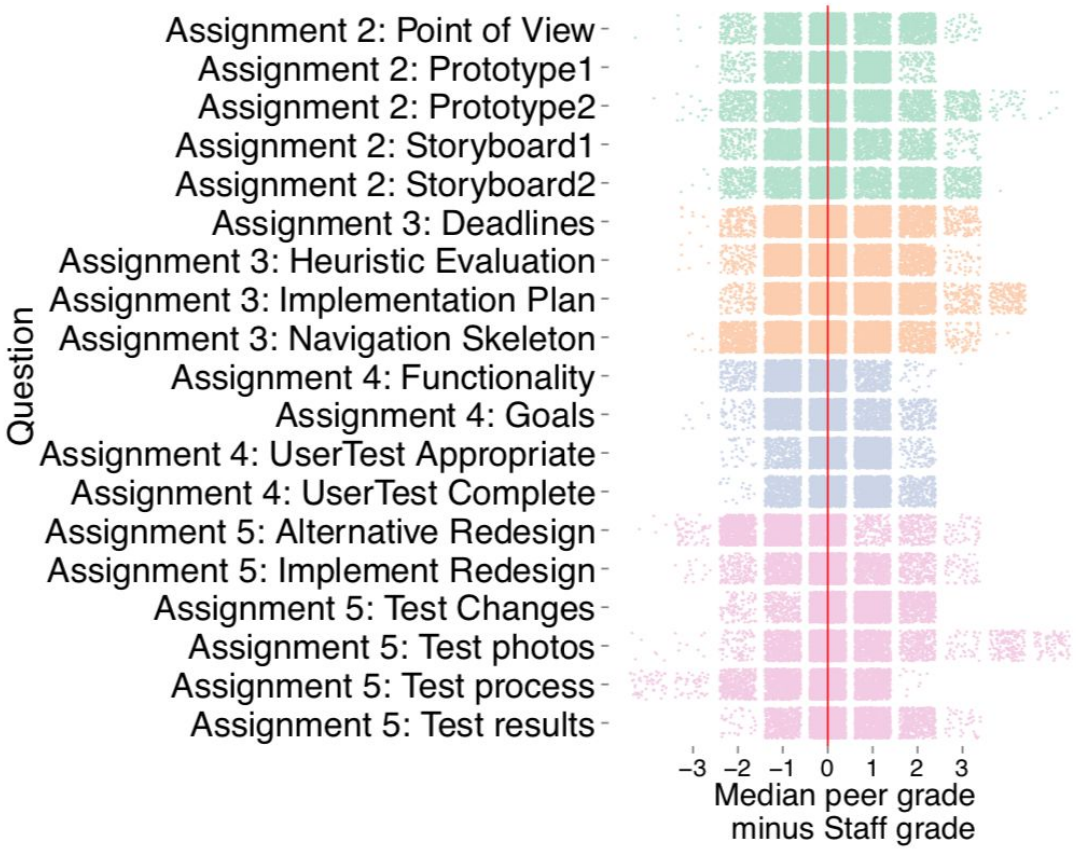
Copy, then paste

Make the prototype more interactive so the user test represents a more real-life interaction: The prototype does everything you're testing, but it couldn't hurt to make it more interactive. If the user can't possibly stray from the things you want to test, how do you know that the user can actually use the full application without making mistakes?

Discussion (2 min, group of 2)

- Could you think of the problem(s) that this fortune cookie approach may have?
- How would you improve that, and design an experiment to verify your hypothesis?

Data-driven rubrics



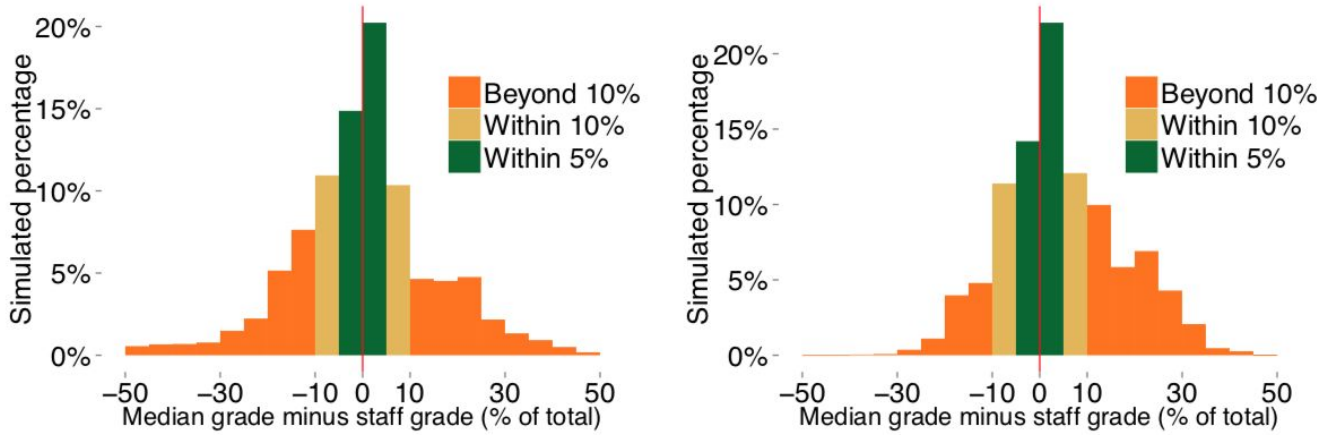
Improvements

- Parallel sentence structure
- Splitting up complex rubric items
- Using less ambiguous words

Table V. Rubric for "Ready for Testing" assignment. Students have created a paper prototype of their application in the previous assignment. Note some items have objective criteria (Did the student meet her goals?), others require subjective interpretation (Is this evaluation plan appropriate?)

Category	Unsatisfactory	Bare minimum	Satisfactory effort & performance	Above & Beyond
List of Changes	0: No changes or completely irrelevant changes.	1: The student only identified a few changes from the heuristic evaluation feedback and a large amount of feedback is ignored in the new prototype; the new prototype has some HE violations.	3: Many of the simpler suggested changes were made, but some of the more complex or difficult issues were not addressed; the new prototype does not have any obvious HE violations.	5: The user made several insightful and specific changes based on the heuristic evaluation feedback. It is hard to find any HE violations at all in the new prototype.
Interactive Prototype	0: No prototype or irrelevant prototype.	1: The prototype is not interactive, lacks many features, and has many bugs; the design does not work with the goal. OR, the student submitted a prototype URL, but the prototype wasn't viewable.	3: The prototype is mostly interactive, with only a few features missing and only one or two bugs; the design accomplishes the minimum requirements of the goal..	5: The prototype is completely interactive, reflects the feel of the final prototype, and is ready for user testing; the design accomplishes the entire goal.
User Evaluation Plan: Completeness	0: No plan or irrelevant plan.	1: User testing evaluation plan exists, but is minimal, unclear, and is not well thought out.	3: The evaluation plan is mostly complete, but does not cover all questions about testing thoroughly (what is tested, what you want to learn, when, where, participants).	5: The evaluation plan is complete, answers all questions specifically, and shows a clear process for user testing.
User Evaluation Plan: Appropriateness	0: No plan or irrelevant plan.	1: The student's evaluation plan does not choose to evaluate aspects of the design related to the design goals.	3: The evaluation plan is designed to produce some useful data, but is not justified by the student (e.g. why are you doing what you are doing?– why 6 participants? Why in a school? etc).	5: The evaluation plan is very clearly motivated or innovative in a way that will ensure rich and interesting data to address the design goals.
Development Goals	0: No goals met that were laid out on the development plan.	1: The student met a few of the goals laid out in the development plan.	2: The student met most, but not all, of the goals laid out in the development plan.	3: The student met all of the goals found in the development.

Accuracy



(a) Iteration 1: 34.0% of samples within 5% of the staff grade, and 56.9% within 10%.

(b) Iteration 2:. 42.0% of samples within 5% of the staff grade, and 65% within 10%.

Students Reaction

- Giving feedback & self assessment are valuable learning
- 20% students voluntarily did more than required assessments

Methods for Ordinal Peer Grading

Karthik Raman and Thorsten Joachims, KDD'14

Learning Goals

- Understand the distinction between ordinal and cardinal grading
- Understand the pros and cons of using ordinal feedback to scale student evaluations.

Question?

What is ordinal grading and cardinal grading?

Ordinal vs Cardinal

- Ordinal words
  - first, second, third, ...
- Cardinal words
  - one, two, three, ...

	Cardinal	Ordinal
Student A	A	1st
Student B	B+	2nd
Student C	B	3rd
Student D	C	4th

Discussion

What are some strengths and limitations of the ordinal peer grading approach?

Ordinal Peer Grading Methods

- Grade Estimation
  - Probability distribution based on rankings
- Grader Reliability Estimation

Grade Estimation Methods

- Mallows Model (MAL and MALBC)
- Score-Weighted Mallows (MALS)
- Bradley-Terry Model (BT)
- Thurstone Model (THUR)
- Plackett-Luce Model (PL)

Ordering based distributions

Pairwise preference based distributions

Experiment

- 8 Week course project
- 44 groups, 3-4 people per group
- Two assignments : Poster and Report
  - Students provided **cardinal** grades (10-point scale): 10-Perfect,8-Good,5-Borderline,3-Deficient
- Conventional grading for comparison
  - TA and instructor grading
- Percentile rank as grade (curve)

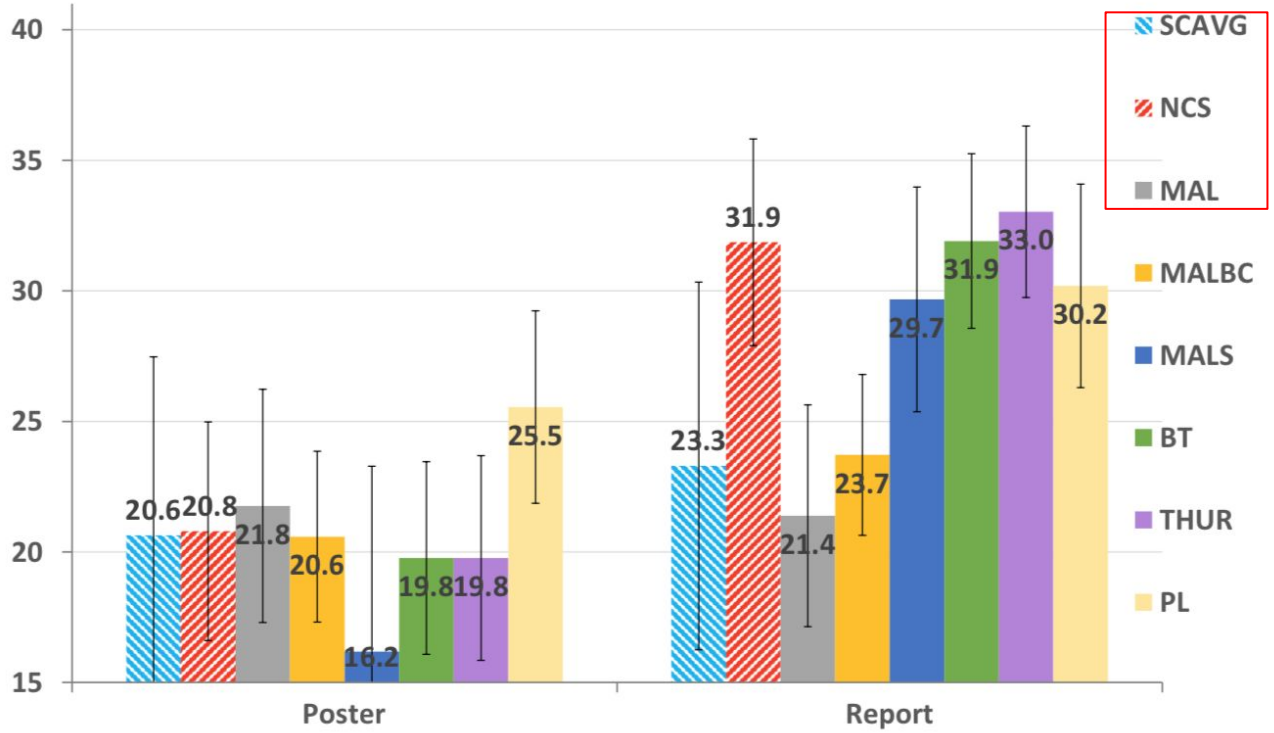
Statistics

Data Statistic	PO	FR
Number of Assignments	42	44
Number of Peer Reviewers	148	153
Total Peer Reviews	996	586
Total TA Reviews	78	88
Participating TAs	7	9
Per-Item Peer Grade Devn.	1.16	1.03

Set	Who?	Mean	Devn.
PO	Peers	8.16	1.31
	TAs	7.46	1.41
	Meta	7.55	1.53
FR	Peers	8.20	1.35
	TAs	7.59	1.30
	Instructor	7.43	1.16

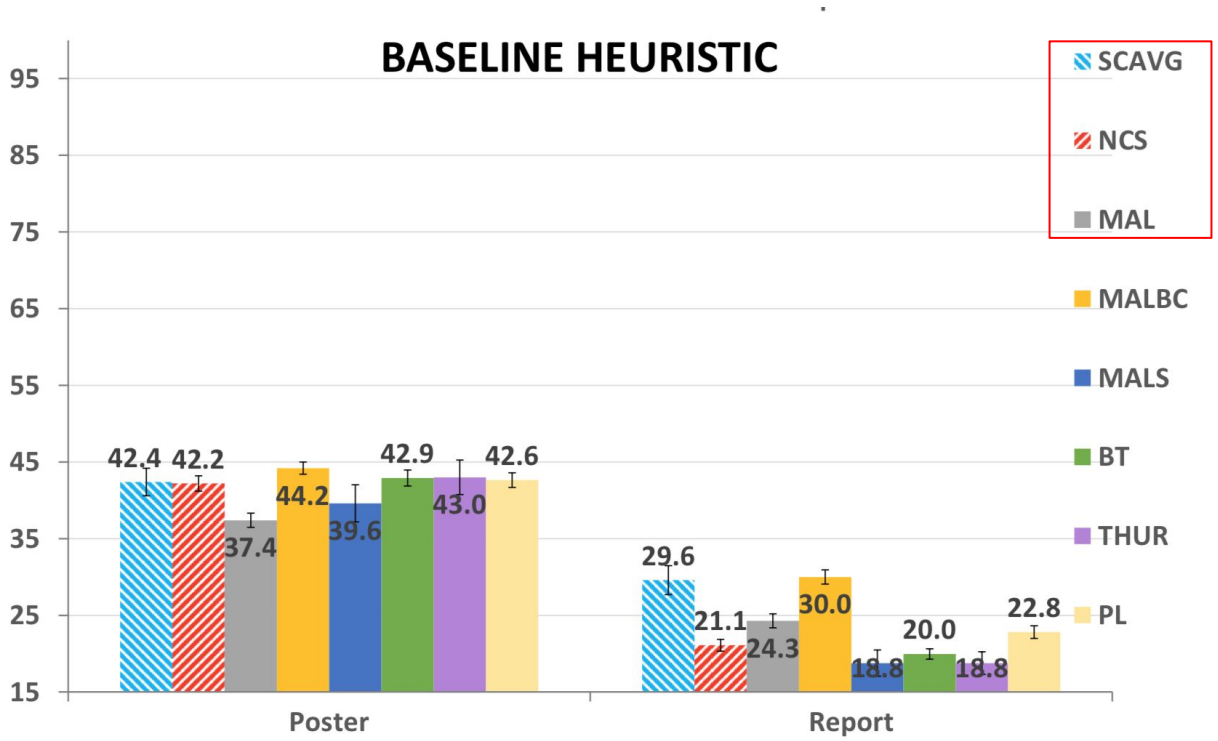
- PO = poster; FR = Report (2 hour poster session)
- Meta (TA grade based on peer grading arguments)

Peer grading vs Instructor grades



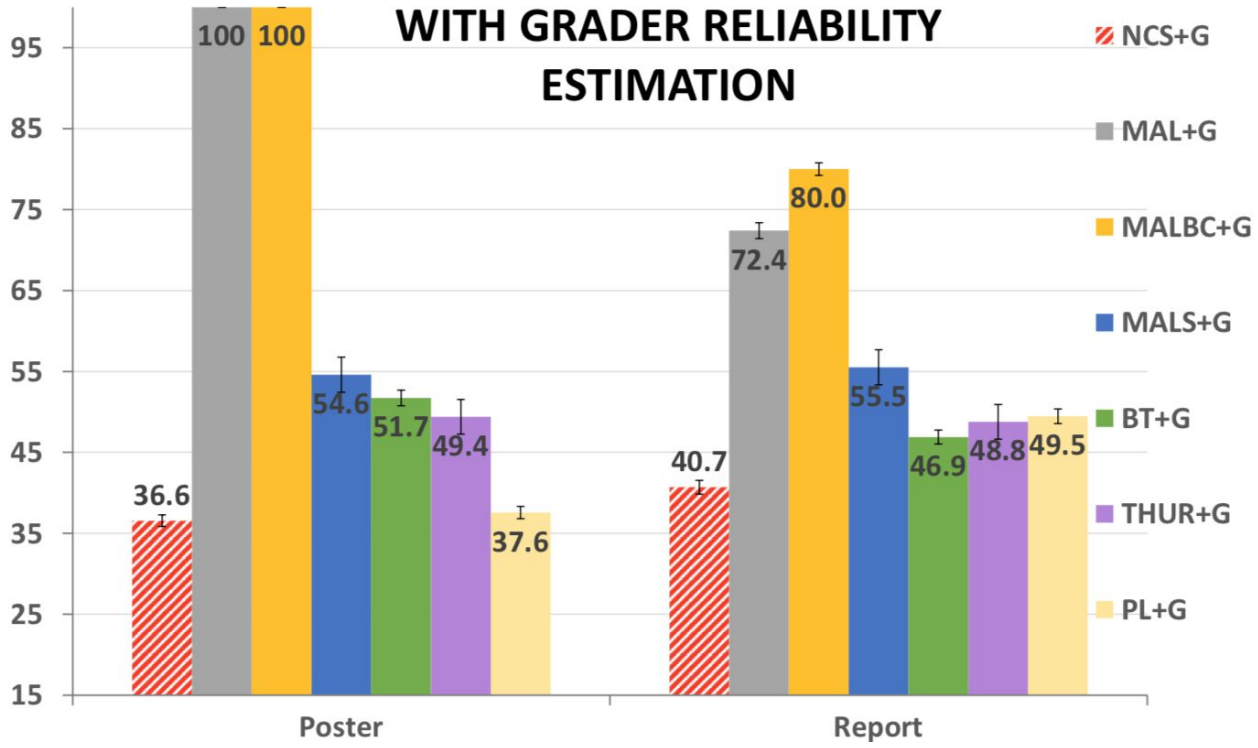
- Kendall-tau error, (lower is better)
- As good as cardinal methods (despite using less information).
- TAs had error of 22.0 ± 16.0 (Posters) and 22.2 ± 6.8 (Report).

Grader Reliability



- Percentage of times a grader who randomly scores and orders assignments is among the 20 least reliable graders (i.e.,bottom 12.5%)

Grader Reliability



- Does significantly better than cardinal methods and simple heuristics.
- Better for posters due to more data.

Question?

In the experiment, the ordinal scoring used cardinal scores to calculate ranking. Why might ranking(ordinal) be better than scoring(cardinal)?

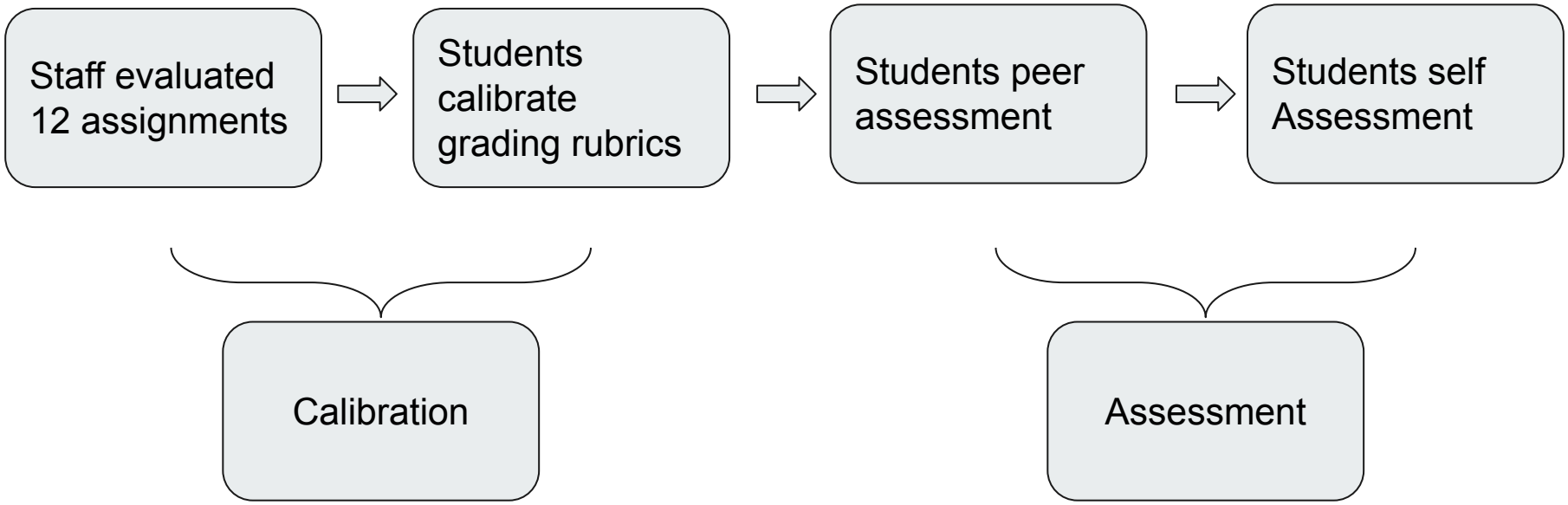
Thanks!

Personalized feedback

Personalized and actionable feedback! but

- Do not cost too much time -> reduce feedback cost for peer- grader

Grading Process



Paper reading questions

Last year: Propose an improvement to the rubric in Table 5 for subsequent iterations of the course and justify why.

My ideas:

1. How to make an effective attack on the peer and self assessment in massive online classes?

2. If you have a choice for grading your homework between peer assessment and staff assessment, which one do you prefer? Why?

3. Do you think the order between peer assessment and self assessment matters to the experiment results? Why?