



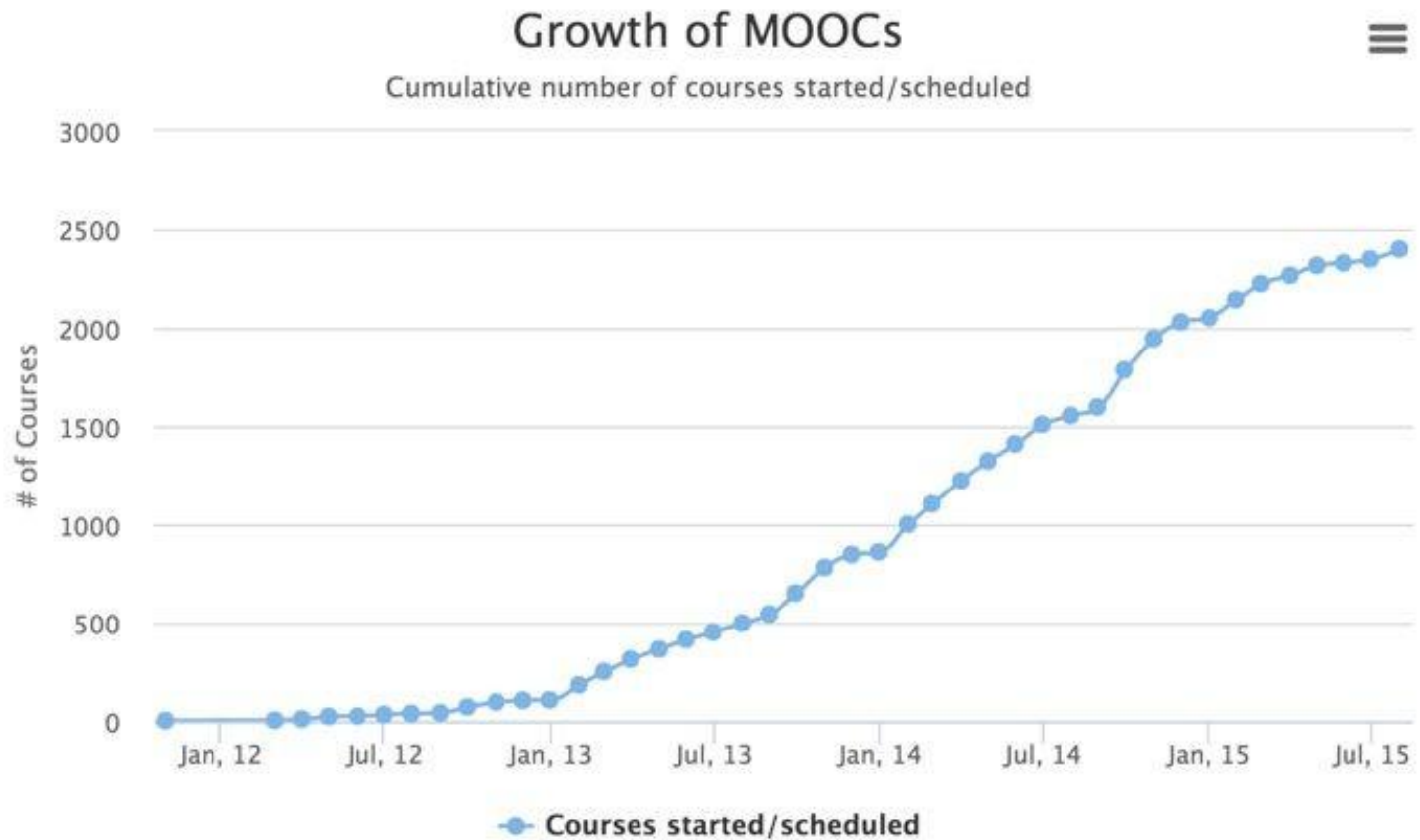
Learning at Scale

Daniel Pan

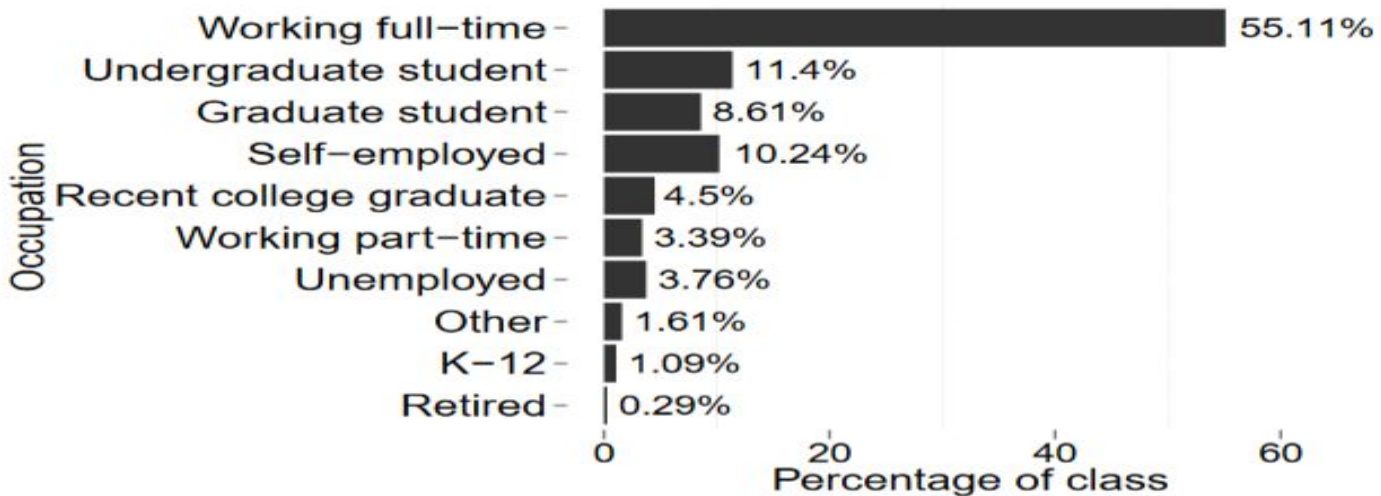
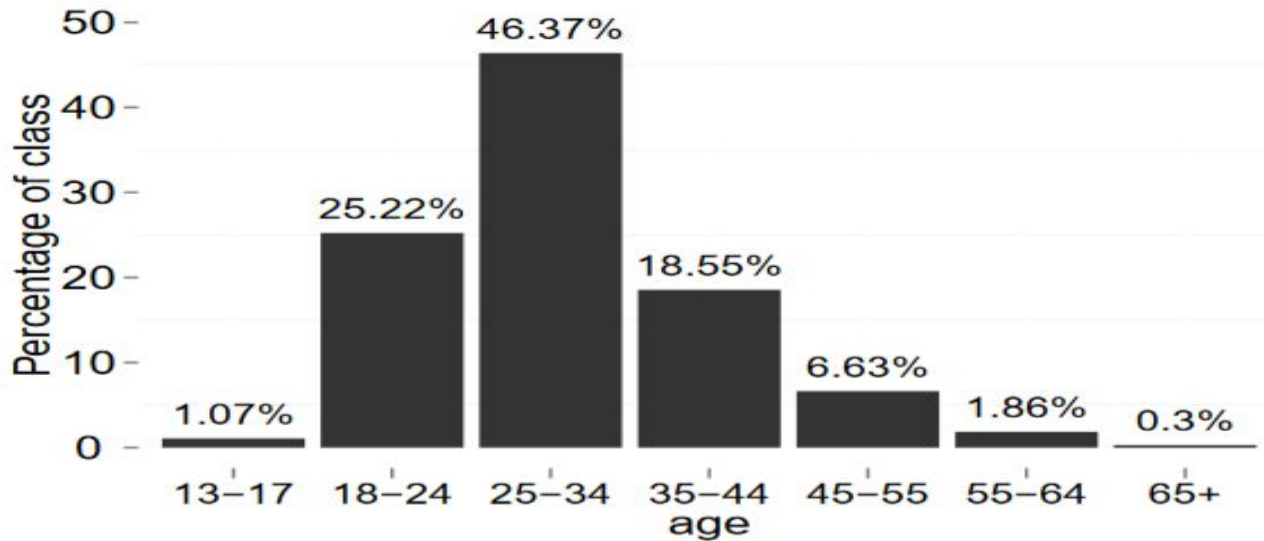
Eric Fakhourian

MOOC

Massive Open Online Course



User Stats





35,081

students who watched videos

2788

submissions of the first assignment

A background network diagram consisting of various sized nodes (circles) connected by thin lines, creating a complex web-like structure. The nodes are in shades of gray and blue, and the lines are thin and light gray.

1

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.
TOCHI 2013.

Learning Goals

- ① Understand peer and self assessment
- ① How the experiment was done
 - The accuracy Analysis
- ① Three approaches to improve accuracy

Peer Assessment

Viewing and critiquing
other's work plays a key
pedagogical role.





Problem#1

How to establish the rule of grading?

Rubric

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

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.

A decorative network diagram in the top-left corner, consisting of various sized circles (nodes) connected by thin lines (edges). Some nodes are solid grey, while others are hollow with a grey outline. The connections form a complex, interconnected web.

Problem#2

How to design the grading process?

Process

Staff evaluated
12 assignments



Student:
Calibrated Peer
Assessment

Calibrated Peer assessment



Discussion

(2 min, group of 2-3)

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

Peer assessment



Calibrated

Peer assessment



**How the score
of the assignment
is calculated?**



Problem#3

How to measure accuracy?

- with only several staff-graded assignment.

Accuracy



Method

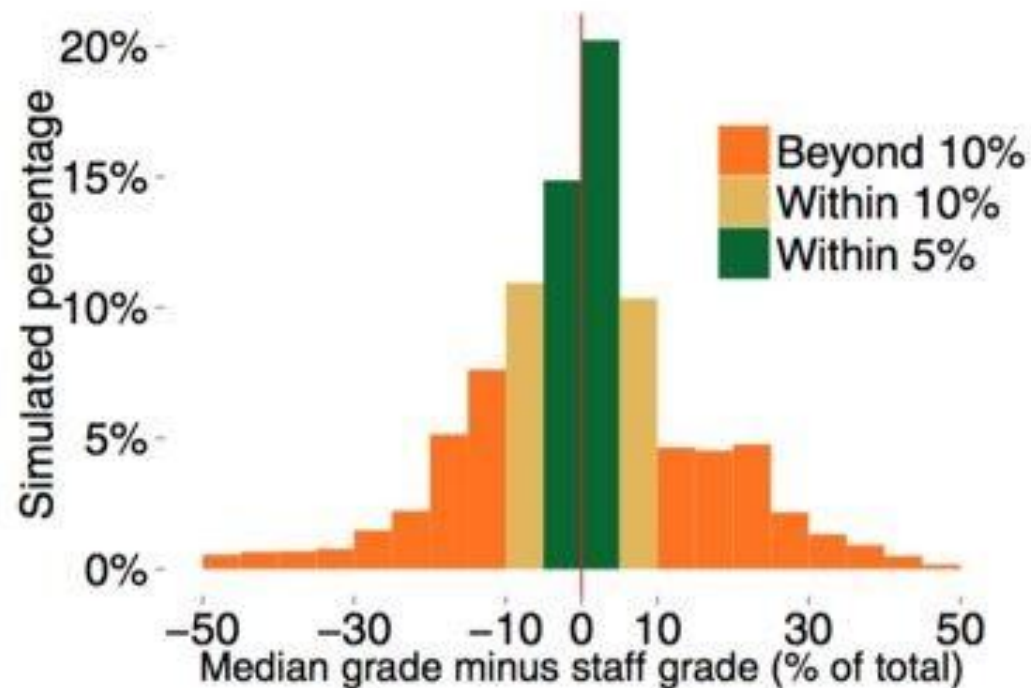
- Ground truth submissions
- Median-grade approach

Peer assessment



Accuracy

Result




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

A decorative network diagram in the top-left corner, consisting of various sized nodes (some solid, some hollow) connected by thin lines, forming a complex web structure.

Problem#4

Can we do better?



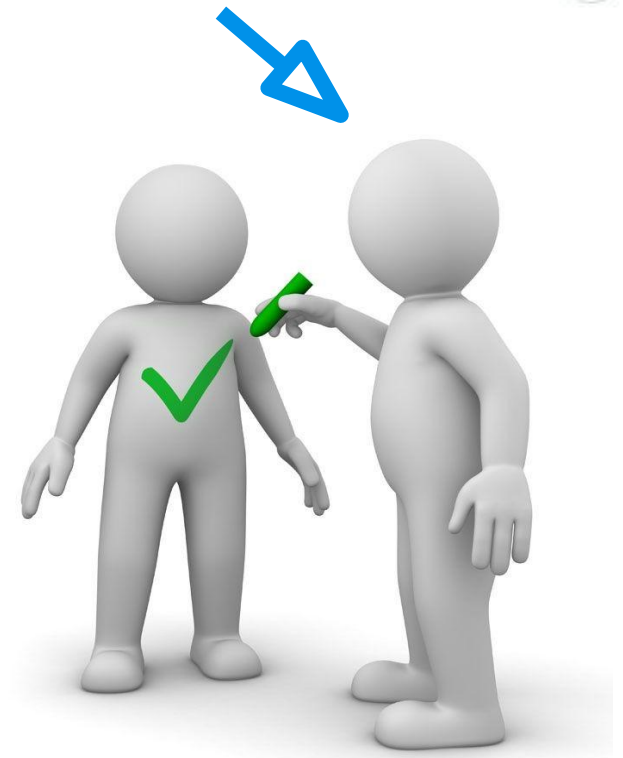
Improve Accuracy & Provide Qualitative Feedback

- Providing Feedback
- Fortune Cookies
- Data-driven Rubric Revisions



Feedback

- ◎ About 800 participants
- ◎ Two conditions
 - no-feedback control
 - feedback



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

Feedback

Word ONLINE
Human-Computer Interaction Associate Professor

Peer Assessments / HCI Assignment 4 - Ready for Testing

 You graded your peers' work accurately on Assignment 3! Keep it up!

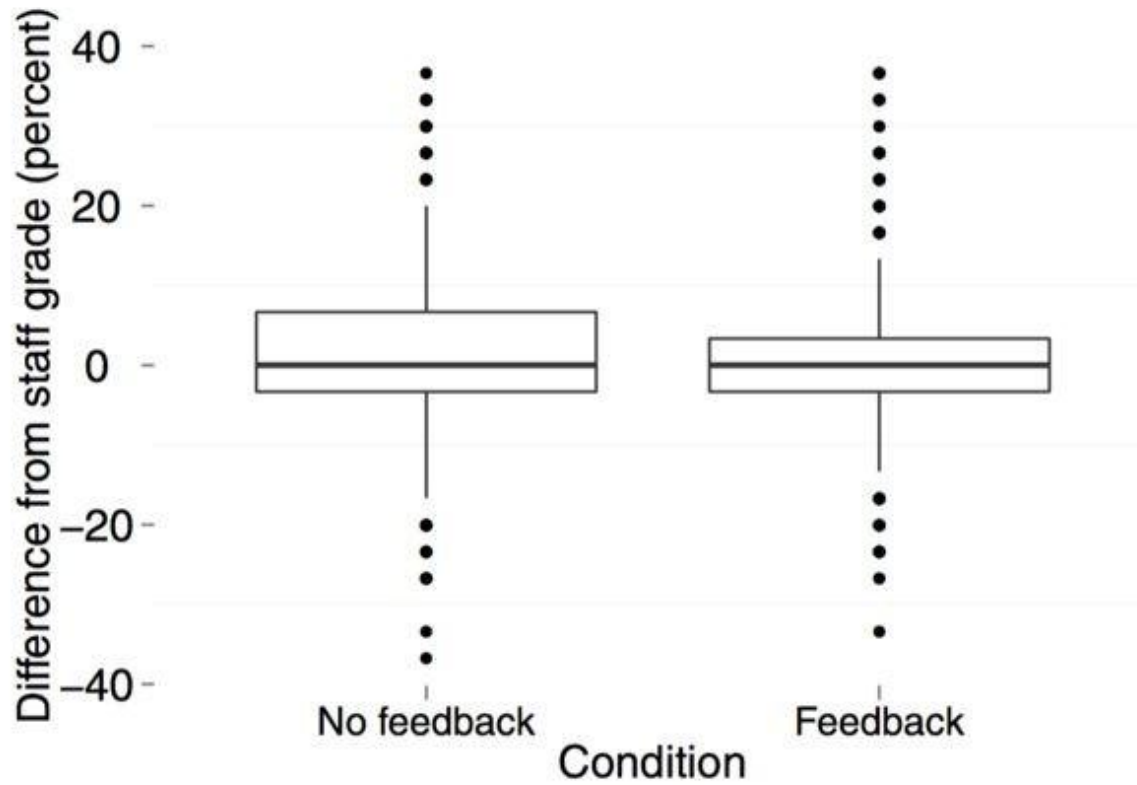
[What's this?](#) [Leave Feedback](#)

1. Do assignment ✓ 2. Learn to evaluate 3. Evaluate your classmates ⚙ 4. Evaluate yourself ⚙ 5. See results ⚙

[← Return to list](#)

[Save draft](#)

Feedback



Provide Qualitative Feedback



Peer ← Grader



Provide Qualitative Feedback

- ◎ Rubric Limitations
 - not clear exactly why did poorly on some topic
 - lack of pointing out how to improve

Fortune Cookie



... because

Fortune Cookie

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?

Fortune Cookie

- ⊙ $\frac{2}{3}$ contained fortune cookie
- ⊙ Do not encourage more students to leave feedback (36.2% v.s. 36.4%)

Fortune Cookie

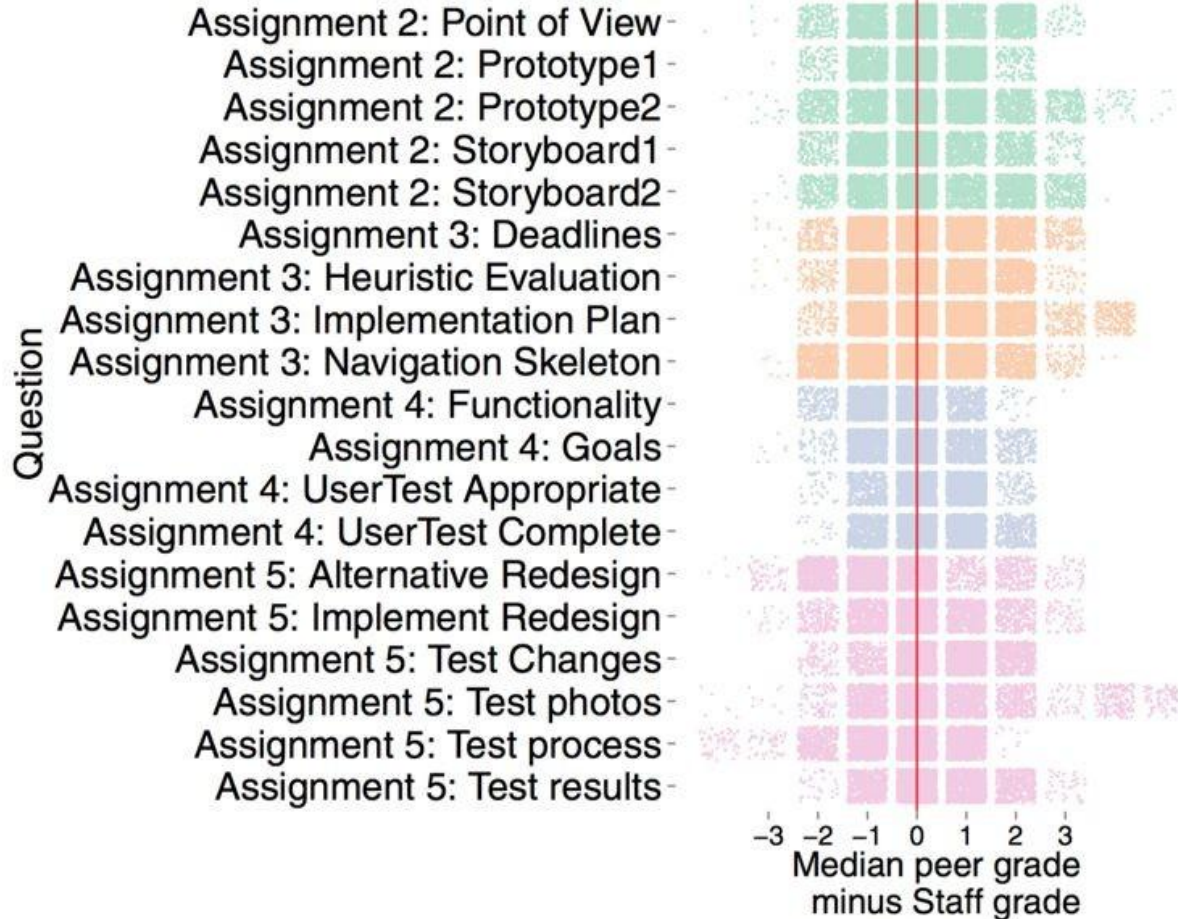
- ◎ However
 - Reduce feedback cost
 - Encourage brainstorming



Discussion (3 min, 2-3 group)

- 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 Rubric Revisions

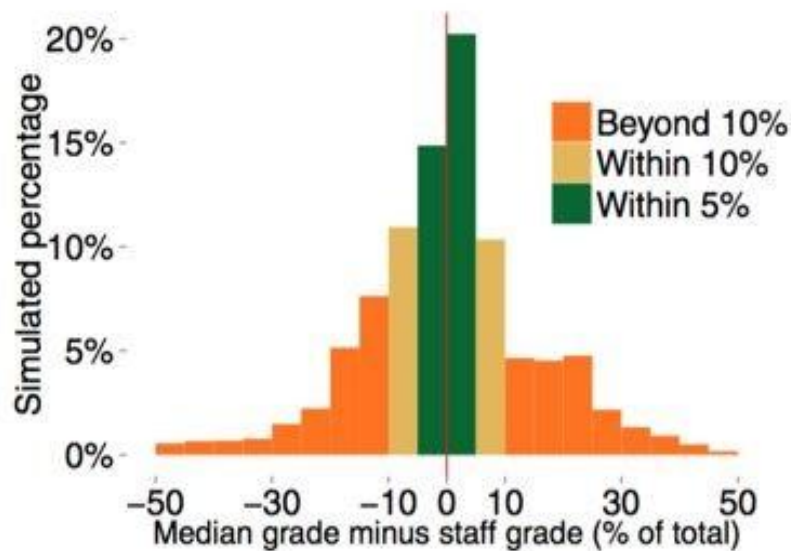


Data-driven Rubric Revisions

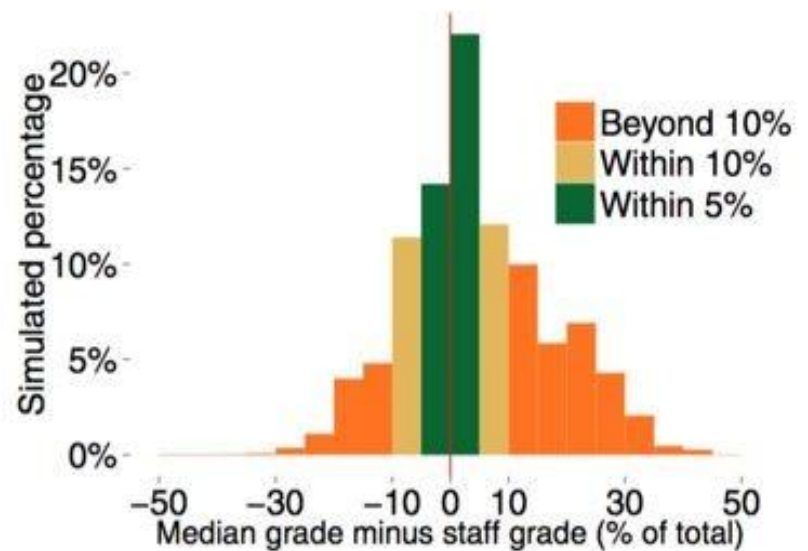
A decorative network diagram in the top right corner, consisting of various sized circles (nodes) connected by thin lines (edges). Some nodes are solid blue, while others are hollow white with a blue outline. The connections form a complex, interconnected web.

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

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



From Your Commentaries

- There are many uses of words like "many", "several", and "few" in the rubric which might differs from graders point of view. -Mohammad
- Any one else!



2

Methods for Ordinal Peer Grading

K. Raman, T. Joachims, ACM Learning at
Scale, 2015

Learning Goals

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

Evaluation at Scale is Challenging

Need to rethink conventional evaluation logistics:

- Small-scale classes (10-15 students) :
Instructors evaluate students themselves
- Medium-scale classes (20-200 students) :
TAs take over grading process.
- MOOCs (10000+ students) : ??

Peer Grading to the Rescue

Peer Grading: Students grade each other



Question?

Someone tell us what is ordinal and cardinal grading?

Ordinal & Cardinal

Ordinal Grading

- Project X is better than project Y

Cardinal Grading

- Project X is a B-

Ordinal v.s. Cardinal

cardinal	one	two	three	four
	1	2	3	4
ordinal	first	second	third	fourth
	1st	2nd	3rd	4th

Ordinal v.s. Cardinal

⊙ Ordinal

- Easier
- More reliable

⊙ Cardinal

- Different scale
- Difficult to provide non-linear



Discussion

(Discuss as a class)

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

Applying Grader Reliability to the Ordinal Bradley-Terry Model

- **GENERATIVE MODEL:**

- Decomposes as pairwise preferences using logistic distribution of (true) score differences.

$$P(\sigma^{(g)}|s) = \prod_{d_i \succ_{\sigma^{(g)}} d_j} \frac{1}{1 + e^{-(s_{d_i} - s_{d_j})}}$$

- **GRADER RELIABILITY:**

- Grader reliability estimation is the task of estimating the accuracy of the grader feedback.
- Grader reliability estimation can be applied to all the ordinal models presented in this paper by incorporating the grader reliability variable as done in this formula.

$$P(\sigma^{(g)}|s) = \prod_{d_i \succ_{\sigma^{(g)}} d_j} \frac{1}{1 + e^{-\eta_g(s_{d_i} - s_{d_j})}}$$

Paper's Approach to Ordinal Peer Grading

- Proposed/Adapted different rank-aggregation methods for the OPG problem:

- Mallows model (MAL).

- Score-weighted Mallows (MALS).

} Ordering-based distributions

- Bradley-Terry model (BT).

- Thurstone model (THUR).

} Pairwise-Preference based distributions

- Plackett-Luce model (PL).

} Extension of BT for orderings.

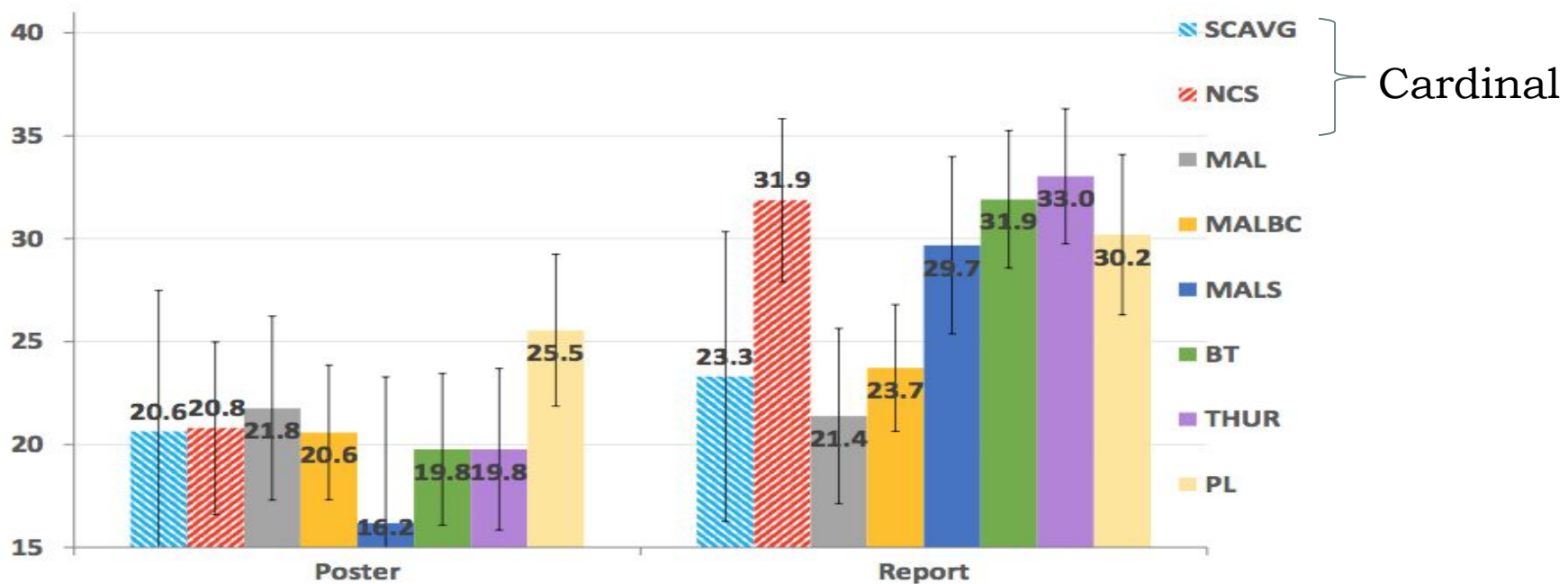


Experimental Validation: New Peer Grading Dataset

- Data collected in classroom during Fall 2013:
 - First real *large* evaluation of machine-learning based peer-grading techniques.
- Used two-stages: Project Posters (PO) and Final-Reports (FR)
 - Students provided cardinal grades (10-point scale): 10-Perfect, 8-Good, 5-Borderline, 3-Deficient
- Also performed conventional grading: **TA and instructor grading.**

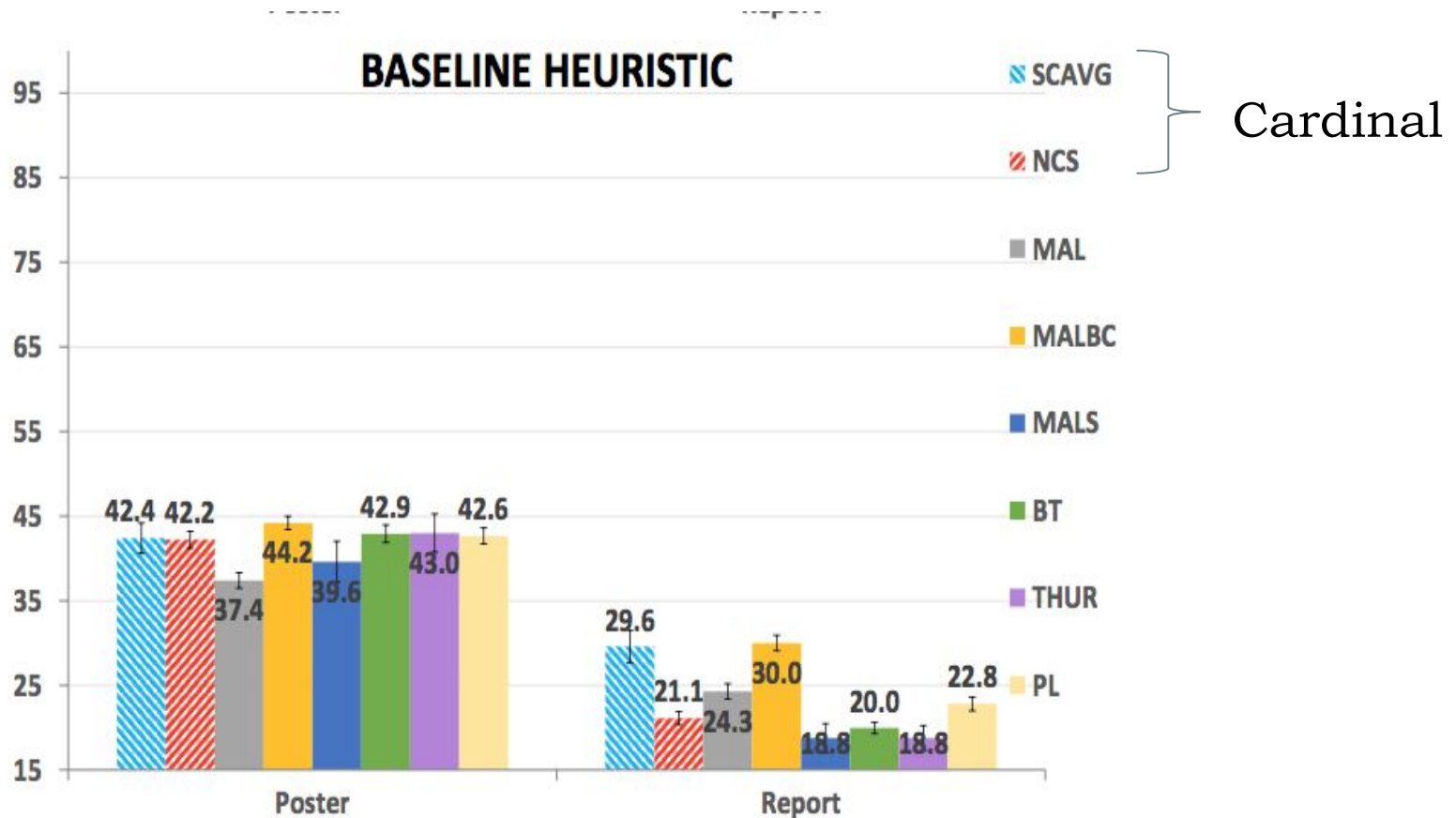
Statistic	Poster	Report
Number of Assignments	42	44
Number of Peer Reviewers	148	153
Total Peer Reviews	996	586
Number of TA Reviewers	7	9
Total TA Reviews	78	88

How well do OPG methods do w.r.t. Instructor Grades?



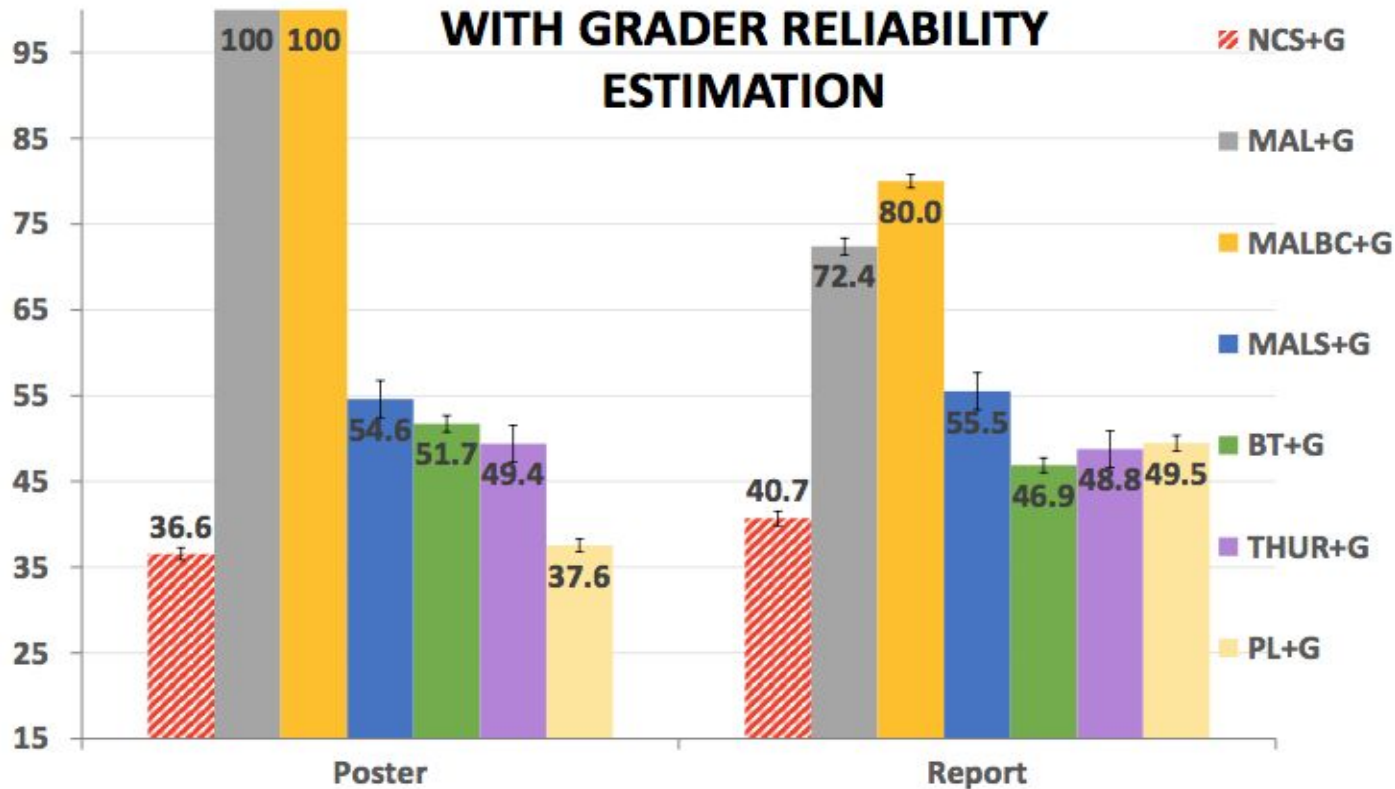
- Kendall-Tau error measure (lower is better).
- As good as cardinal methods (despite using less information).
- TAs had error of 22.0 ± 16.0 (Posters) and 22.2 ± 6.8 (Report).

Benefit of 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%)

Benefit of grader reliability



} Cardinal

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

Question

Why might ranking(ordinal) be better than scoring(cardinal)?



Discussion (2 min, group of 2-3)

- Should Coursera adopt this ordinal grading technique at scale?
- Discuss potential limitations of such peer assessment method.

Take Away

- Benefits of ordinal peer grading for large classes.
- Using data from an actual classroom, peer grading found to be a viable alternate to TA grading.
- Students found it helpful and valuable.



Thanks!