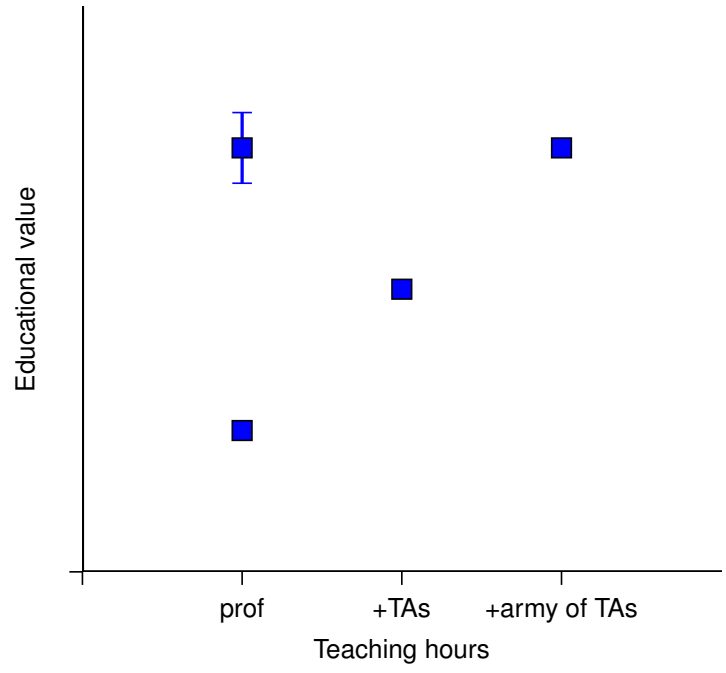
# Mechanical TA:
# Partially Automated High-Stakes Peer Grading

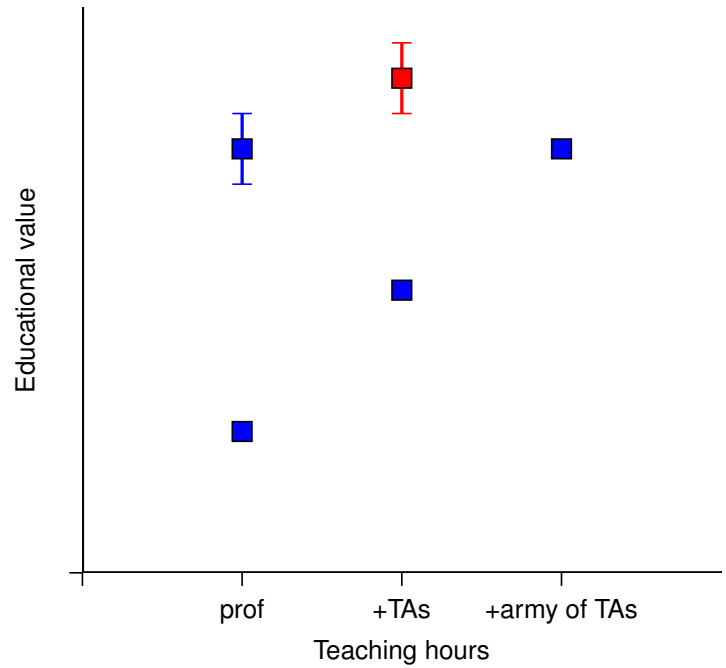
James R. Wright, Chris Thornton, Kevin Leyton-Brown

# Peer Grading




- With just an instructor, maybe an exam and 1 assignment.
- With an instructor and TAs, exams and several assignments.
- With peer grading, students grade the assignments.

# Peer Grading Drawbacks

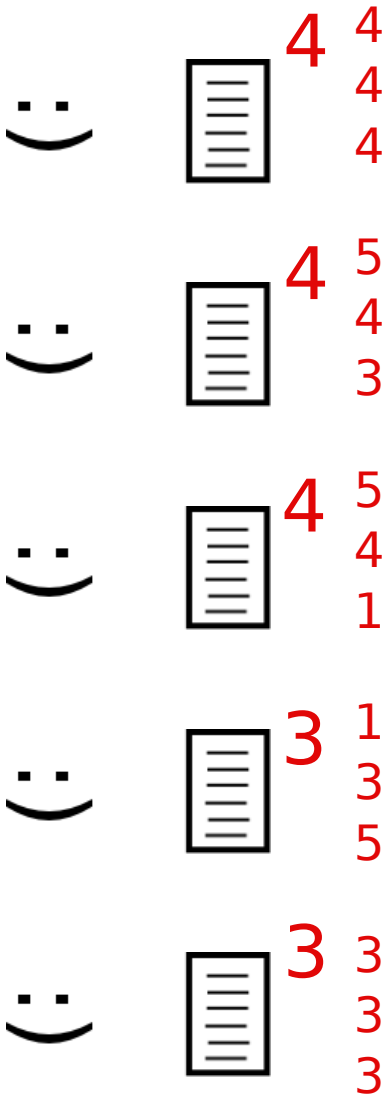



- Only works if we trust students to give meaningful feedback!
  - Students may not have the ability to give high-quality, accurate grades and feedback.
  - Even if they are able, students may not put in the effort.
- Mechanical TA leverages TA time to solve these problems.

# Motivating Example(s)

- CPSC 430 — "Computers and Society"
  - Fourth-year undergraduate course (70–100 students).
  - Reasoning critically about implications of technology.
  - Crucial element: weekly essays.
    - Excellent tool for practicing (and assessing) clear thinking.
    - Encourages engagement with the material.
    - Major component of the students' grades (35%).
- Many of the same issues apply to programming as critical writing:
  - Practice is an important part of learning to program.
  - Subjective feedback is extremely valuable.

# Peer Grading

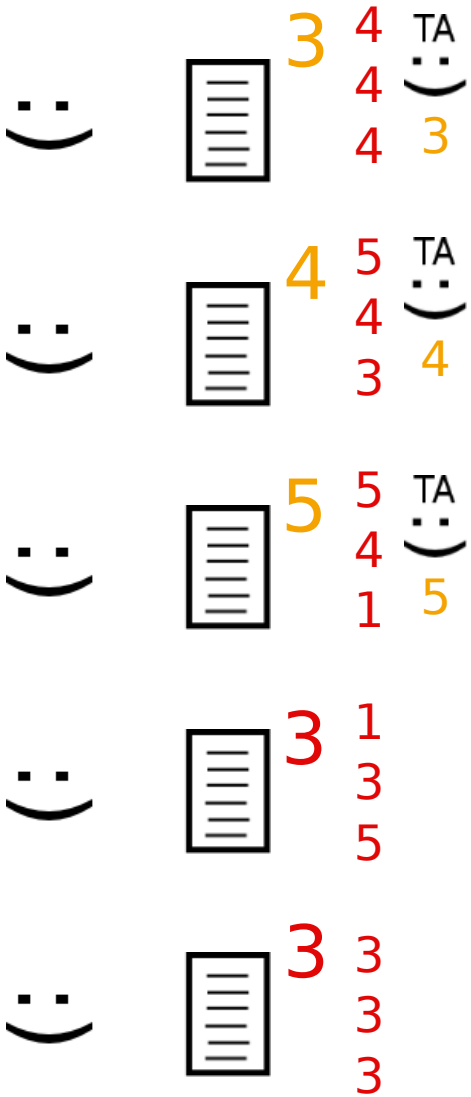



- Students grade each others' submissions.
- Every submission gets multiple student reviews.
- Aggregate reviews to get the submission's "true" grade.

Related work:

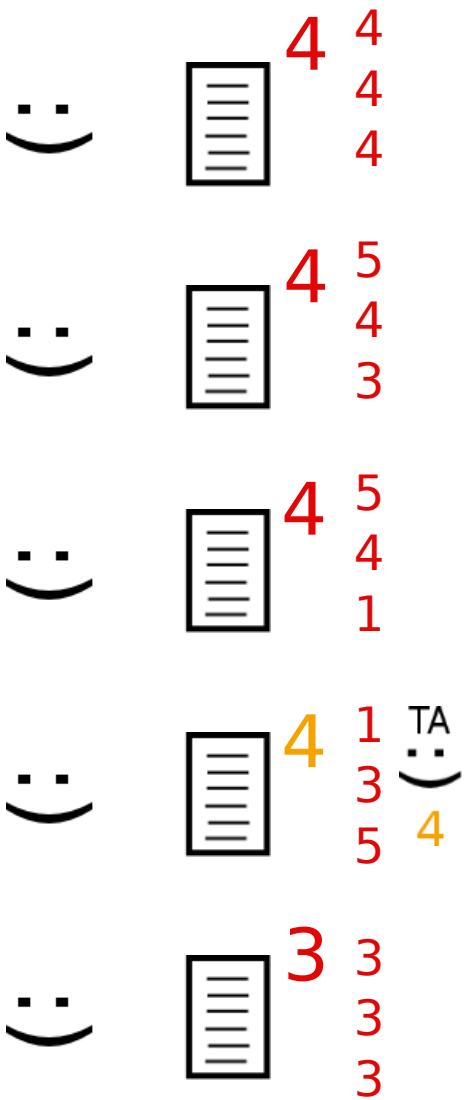
- Calibrated Peer Review [Chapman 2001] tests students for reviewing competence before each assignment.
- Aropa [Hamer et al. 2005] re-weights reviews by consistency with the "consensus" grade.

# Supervision




- Initially, students may not have the ability to give good reviews.
- Supervised students: TAs mark both essay *and* the reviews themselves.
- Each student becomes independent (trusted) after his/her reviews meet a quality threshold.
- Once a student is independent, they stay independent (unless demoted).

# Spot Checks




- Independent students have demonstrated ability to review competently.
- We randomly spot check to ensure that they are motivated as well.
- Large fraction of students' final grade is from reviewing:
  - Supervised reviews are marked by TAs.
  - Spot-checked reviews are marked by TAs.
  - All other reviews get 10/10.
- If a spot checked review is below the quality threshold, student may be demoted to supervised again.

# Automated Review Practice/Assessment

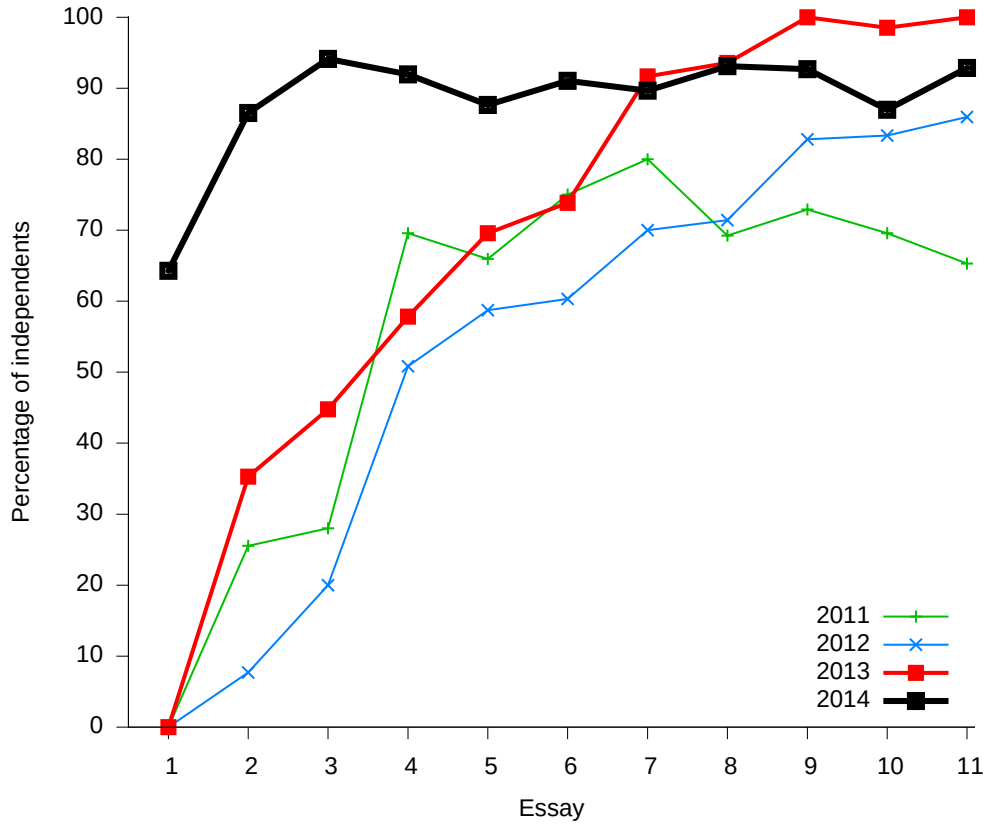
In 2011 and 2012:

- Every student starts out supervised.
- Promoted to independent when review marks pass threshold.
- TAs have to mark every submission of the first assignment!

Starting in 2013:

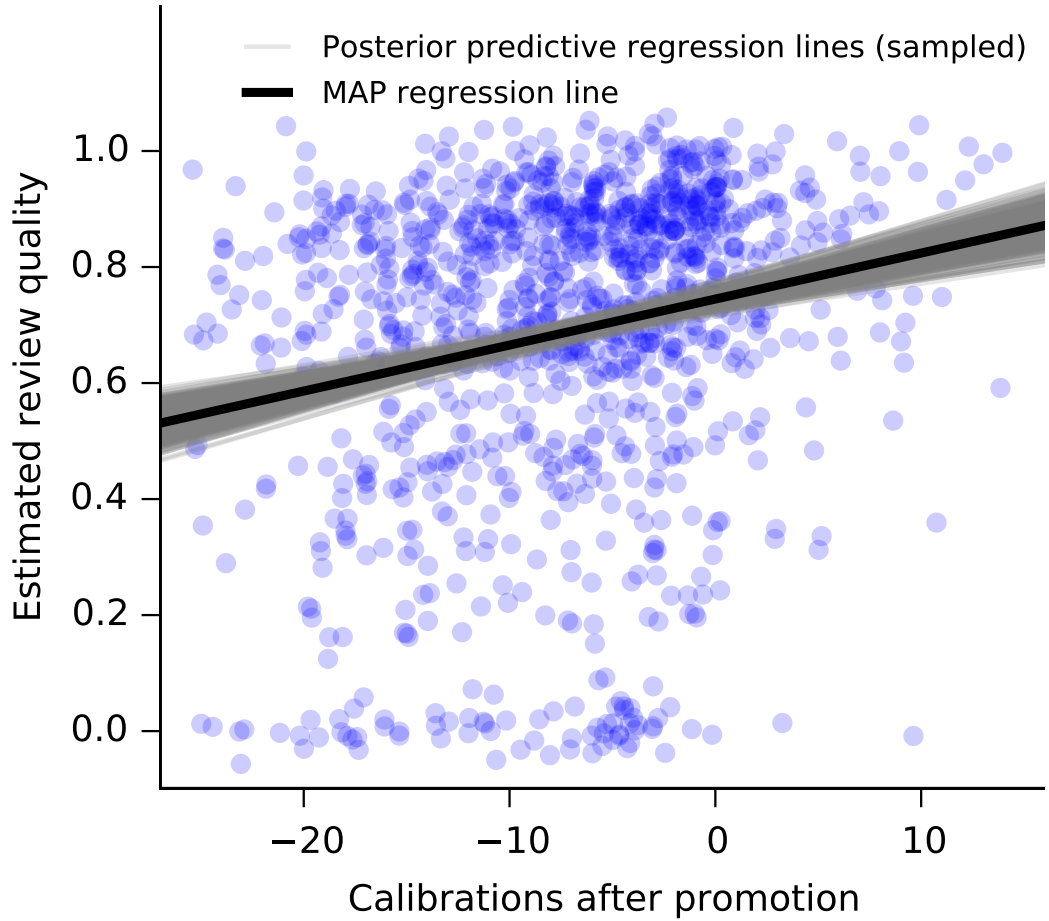
- Students optionally review "gold standard" essays.
- Immediate feedback.
- Promoted automatically if they match answer key closely enough.

# 1. Independent Reviewers



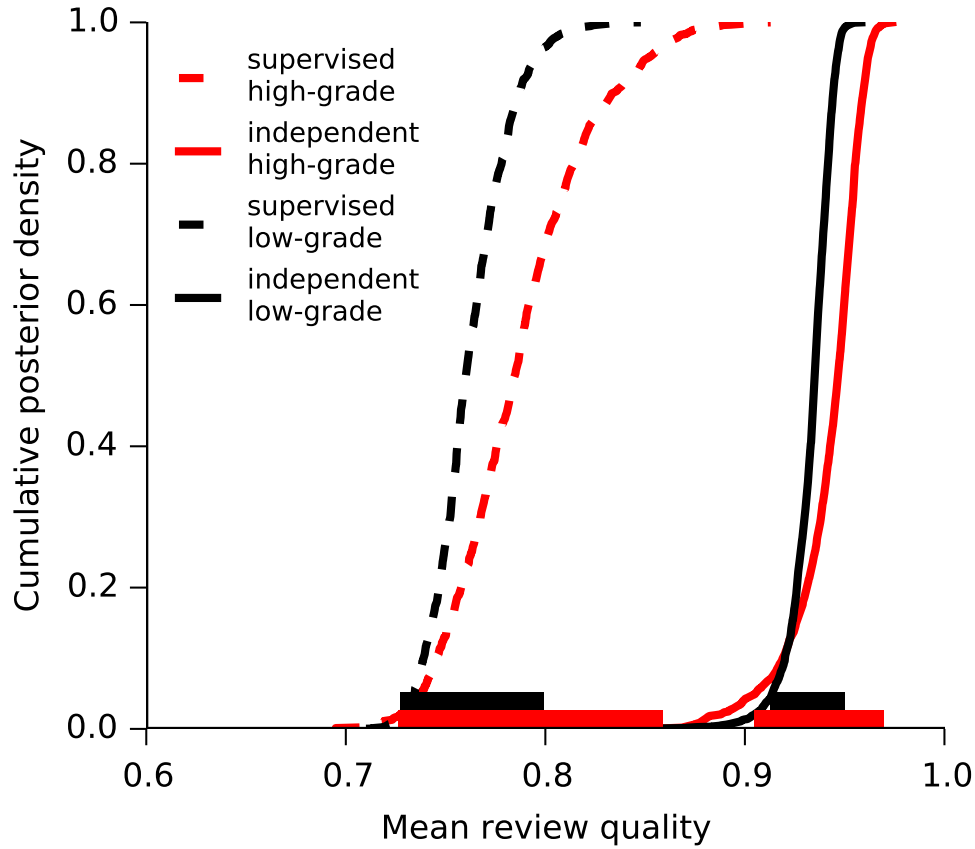

- 2011: Promotion threshold was too easy.
- 2012: Promotion took longer but tended to stick.
- 2013: No automatic promotions, but faster promotion.

# 2. Automatic Review Practice




- Different starting abilities, so normalize by promotion time.
- Students' reviewing ability improves with automated practice.

# 3. Independent/Supervised Review Quality

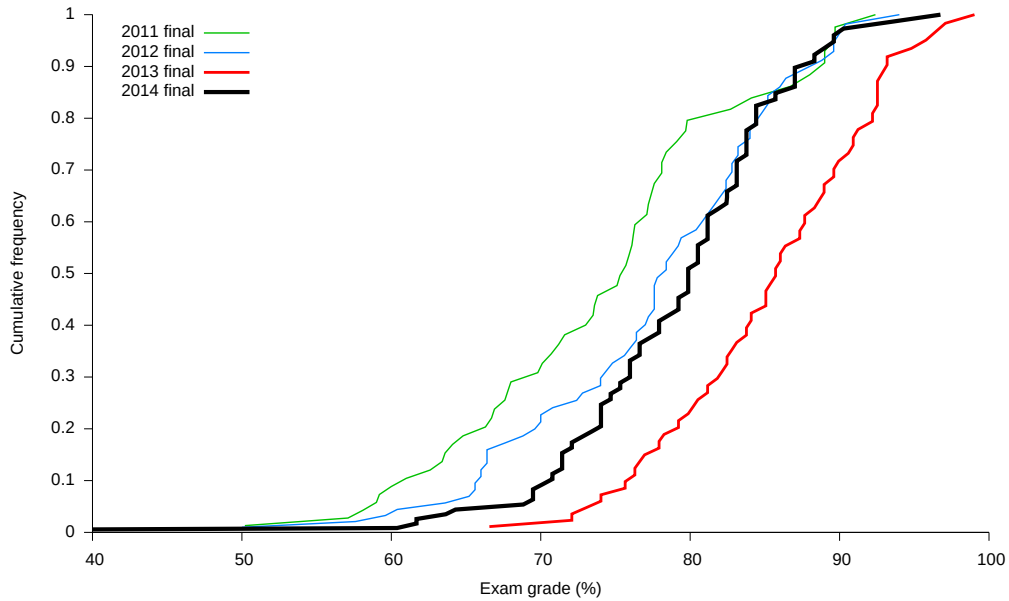



- Supervised/independent distinction is key to our design.
- But do independent reviewers actually do a better job?

# Summary

- Peer grading allows frequent, rich assignments to scale up but brings new problems:
  - Unverified reviewer ability
  - Unverified reviewer honesty
- Mechanical TA leverages TA resources to solve these problems.
  - Allowed us to run an essay-based course at a scale that would otherwise be impossible.
  - Peer review has benefits of its own.
- You can use it too!
  - Download available at `www.cs.ubc.ca/~jrwright/mta/`.
  - UBC CS IT maintains an instance at `www.cs.ubc.ca/mta/`.

# 4. Exam grades




- Assignment grades incomparable between years due to drastic rubric changes.
- Final exams were roughly comparable between years.
- 2013 class did better on final exam than earlier two years.
- 2014 did better too but not as strikingly.

# Improved Calibration

Two main improvements:

1. "Squared-deviation" performance measurement.
   - Reviewers grade $0 - -5$ on 4 dimensions.
   - Originally: Students who were within 1 on
   - Original calibration had maximum difference
2. Data-driven quality threshold.