Embracing Peer Assessment

John Hamer

J.Hamer@cs.auckland.ac.nz

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

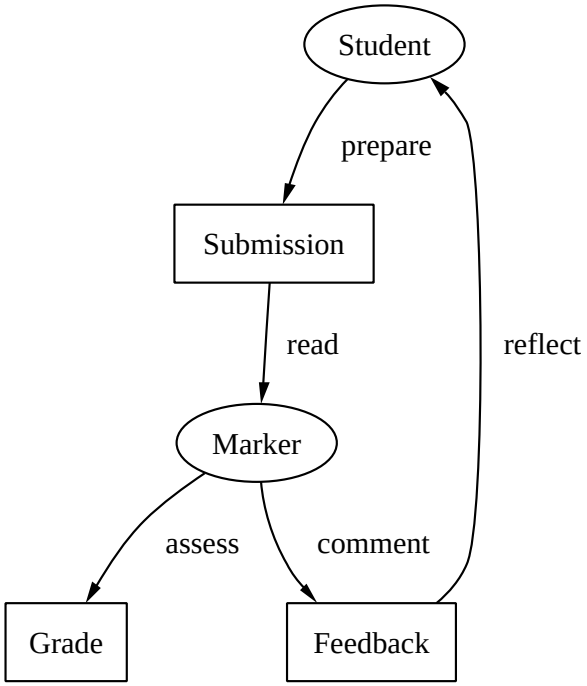
Today is St. Patrick's Day!

There are seven masters in Leamy's National School and they all have leather straps, canes and blackthorn sticks. They hit you with the sticks on the shoulder, the back, the legs, and, especially, the hands. If they hit you on the hand it's called a slap. They hit you if you're late, if you have a leaky nib on your pen, if you laugh, if you talk, and if you don't know things.
They hit you if you don't know why God made the world, if you don't know the patron saint of Limerick, if you can't recite the Apostle's Creed, if you can't add nineteen to forty-seven, if you can't subtract nineteen from forty-seven, if you don't know the chief towns and products of the thirty-two counties of Ireland, if you can't find Bulgaria on the wall map of the world that's blotted with spit, snot, and blobs of ink thrown by angry pupils expelled forever.

—Frank McCourt, "Angela's Ashes"

Context and motivation

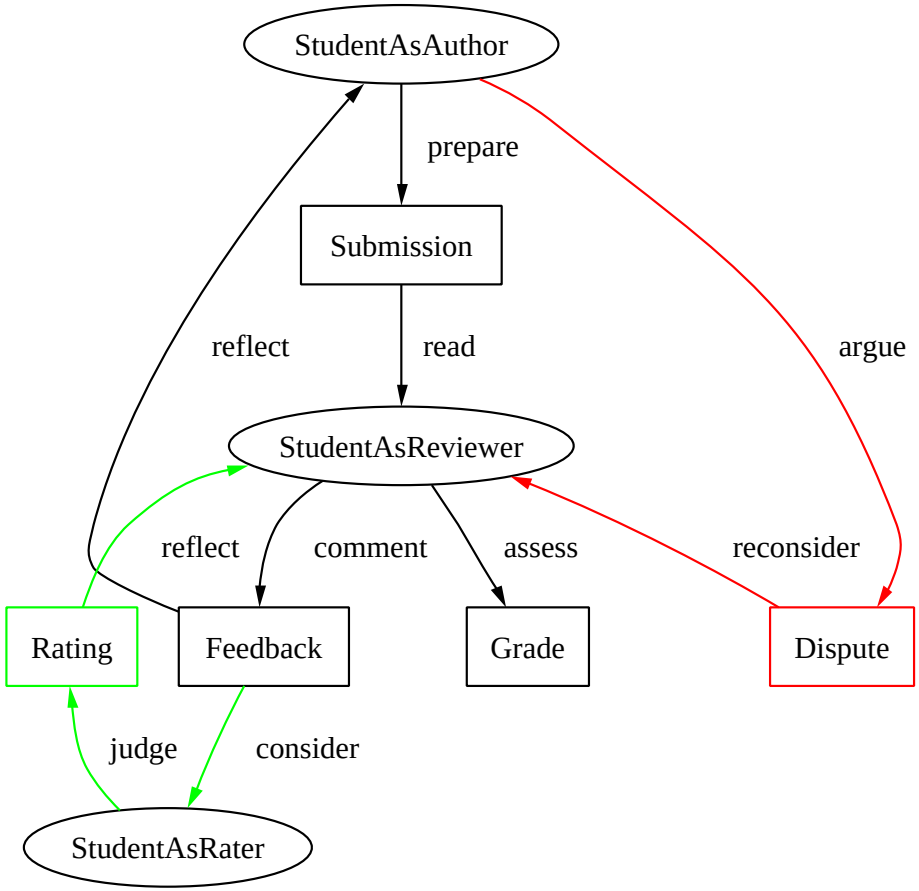
Traditional assignment



Where does the learning occur?

- Learning happens in the preparation of assignment submissions, and to a much lesser extent in reflection on feedback:
  - long marking time dilutes value of any feedback
  - markers have little incentive to produce quality feedback; monitoring marker performance is expensive (cheaper to wait for complaints. . .)
  - exhausting and repetitive marking workload leads to drudgery
- this type of assessment has a *summative orientation*, but suffers from problems of plagiarism and low quality marking

Peer-assessed assignment



Now where does the learning occur?

- *every link* involves some kind of learning
- time delay and drudgery eliminated
- includes performance incentives (ratings)
- primarily formative, unproven summative potential

The Dispute and Rating steps arise from a change in power relations. They are not a statement about the quality of the process. With peer assessment, questioning the reviewer is a *legitimate* activity.

What changes?

- Increased involvement by student (time on task, time engaged with task)
- Greater variety of tasks undertaken by student
- Reduced delay between authorship and feedback
- Increased volume and diversity of feedback
- More opportunities for reflection
- Raised awareness of own relative performance
- Change in power relations between author and reviewer, student and lecturer
- Greater social involvement
- Richer trace of student performance
- Assessment becomes a part of the learning process
- Department marking budget available for redistributing to remedial tutoring, etc.

Opportunities presented

- Broader range of assess-able assignment tasks become possible
  - essays and reports
  - user-interface designs
  - posters
  - seminar presentations
  - web sites
  - Problem-Based Learning
  - . . .
- (Qualitative) assessment becomes more reliable with multiple reviewers, allowing greater summative component
- Frequent (or continuous) formative feedback possible
- Combine peer assessment with group work (authorship and/or review)

Making peer assessment reliable

(Undergraduate) classes inevitably include a proportion of students who:

- do not participate;
- exhibit bias toward people they know;
- copy or plagiarise;
- mark erratically, or are incapable of making informed judgements;
- deliberately disrupt the assessment process.

These issues are addressed with double-blind reviewing, plus novel mechanisms to motivate participation and manage "rogue" markers.

Motivating participation

- Peer assessment *cannot* function effectively without a high participation rate (if noone marks a submission, you have no feedback, no grade).
- **Intrinsic motivation**. Sense of social responsibility, plus curiosity to see the work of their peers.
- **Extrinsic motivation**. Award a non-trivial portion of the assignment marks to the review and reviewer rating.

  A simple "participated" mark is not sufficient; marks must differentiate between good and poor performance.

Automatic Grade Calibration

Managing rogues

- Two issues:
  - detecting rogues (so you can punish them)
  - minimising the impact of rogues (so other students' marks are not distorted)
- Novel "Automatic Grade Calibration" method does this.
- Bear in mind: the objective is to motivate serious, considered participation. Influencing student perceptions is more important than absolute precision.
- Sustained student resistance (rebellion) will defeat this (or any) method. Effort still needs to be expended in promoting the positive benefits of peer assessment.

Detecting and deflecting rogues

Principles:

- The assessment of each assignment must be self-contained; should not pre-judge student performance in subsequent assignments (however much one might like to)
- Independent (tutor) marking should be optional. There is no cost saving otherwise. And anyway, it's like riding a bike without training wheels [without a crash helmet?].

The Method

- Incentivise students to take marking seriously. Each student receives a *grade* for their submission and a *weight* for the quality of their reviewing.
- Higher weights are given to "better" reviewers; i.e., those who mark close to the consensus.
- Grades are calculated using a weighted average of all reviewer marks.
- Requires finding the fix-point (an iterative algorithm rapidly converges).
- Dampening is used to prevent individual weightings becoming too high (there are many fix-points, and this avoids extreme cases)
- Weights typically range between 0.2 and 4.

Formally,

Let

- g_e^r be the grade assigned by reviewer r to essay e
- R_r be the set of essays allocated to reviewer r
- E_e be the set of reviewers allocated to essay e
- G_e, the current grade for essay e, is

$$G_e = \frac{\sum_{r \in E_e} g_e^r \times W_r}{\sum_{r \in E_e} W_r}$$

- Δ_r, the differences between assigned and awarded grades

$$\Delta_r = \frac{\sum_{e \in R_r} (G_e - g_e^r)^2}{|R_r|}$$

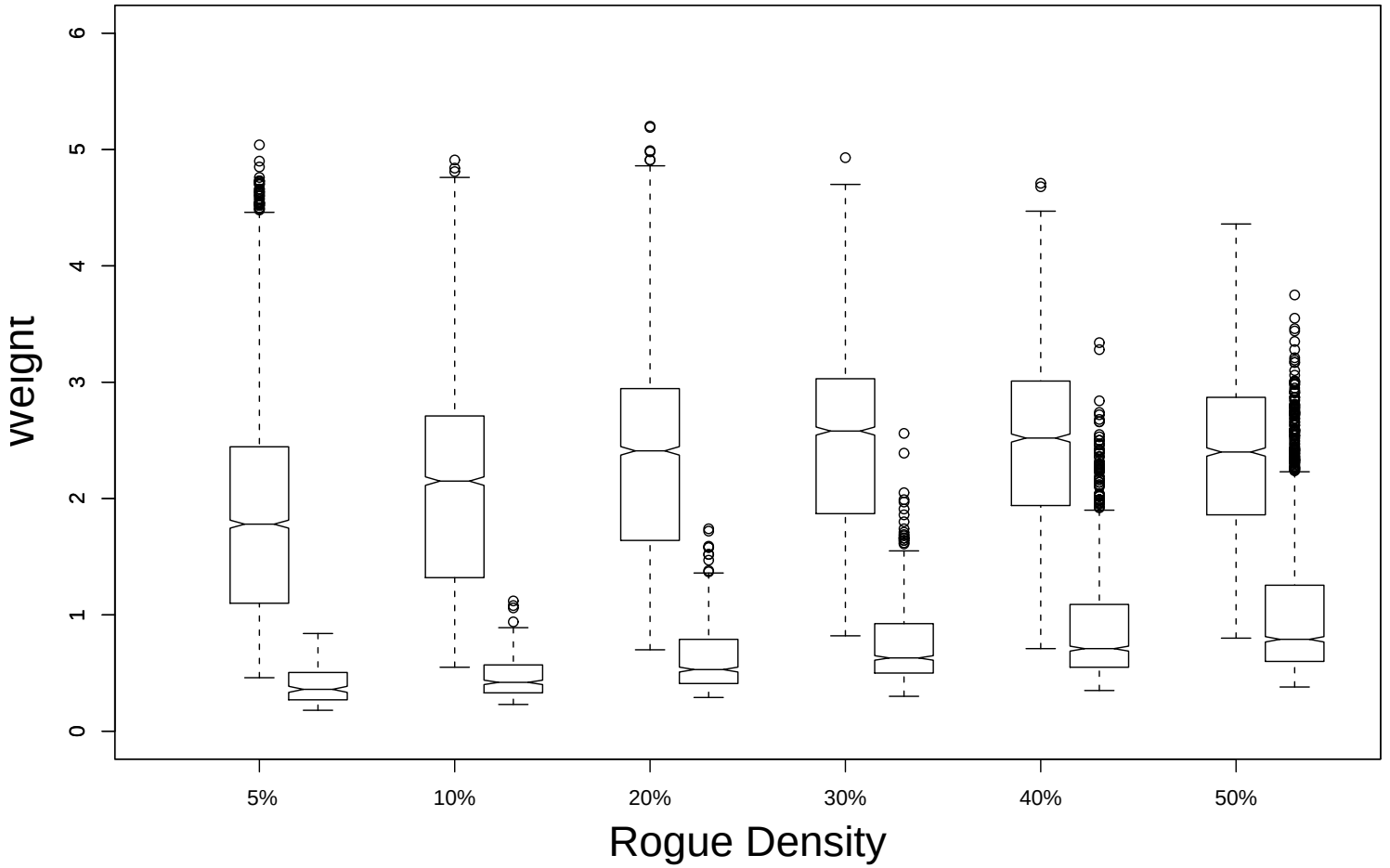
- Weights are assigned in inverse proportion to the difference:

$$W_r \propto^{-1} \Delta_r$$

Experimental simulation

- Decide on a model distribution of essay marks (e.g., 6% are worth 10/10, 10% are work 9/10, etc.)
- Decide on a distribution for marking accuracy (e.g., 20% of marks are accurate, 40% are ± 1, 25% are ± 2, 12% are ± 3 and 3% are ± 4)
- Populate the simulation with n reviewers, of whom x% behave as rogues (i.e., they generate grades without reference to the essay)
- Run the simulation with 100 students, each grading 10 essays, varying the rogue proportion from 5 to 50%

Simulation results



The box plot to the left of each pair is for the serious markers, the box plot on the right is for the rogue markers.

Grading rubrics

Grading rubrics

- Likert scales are often used, but problematic
  - does my "strongly agree" mean the same thing as your "strongly agree"?
  - exhibit a "centrist" tendency
  - require statistically significant sample sizes, not 5–10 reviews
- A *grading rubric* provides far greater precision, and is the key to providing reliability.
- A rubric is a scoring tool that lists criteria for a piece of work, and articulates gradations of quality for each criterion

An example rubric question

Descriptive variable names

Poor names chosen for at least two of the variables do not describe the information stored in the variables. If it is not possible to figure out what the variable is used to store simply by looking at the name, then the name is not adequate.

Good the names chosen for variables describe the information stored in the variable. However, there is one name which does not describe the information stored in the variable.

Perfect the names chosen for all variables clearly describe what the variables are used for.

Another example

Followed the Assignment's Directions

Inadequate The paper has no apparent relation to the directions of the assignment.

Needs Improvement Some of the paper follows the directions.

Adequate Most of the paper follows the directions.

Excellent The paper follows the directions precisely. (i.e. the sections are labeled, directions for finding the article are clear, all required information, etc.)

Rubrics (cont.)

- Rubrics can be very specific or very general
- You can trade off between precision and flexibility
- Focus on surface features or probe more deeply, together and in any combination
- The rubric can be designed with input from the students themselves. This promotes reflection, understanding and ownership
- Easy to combine with free-format, open-ended comments

Plagiarism

Detecting plagiarism

- Turnitin.com is all very well, but how much can you trust it?
- Peer-reviewers are in a good position to detect plagiarism
  - small number of submissions to cross-check
  - personally affected by other students cheating
  - familiar with Internet and other relevant resources
  - able to apply intelligent plagiarism detection
- Technical note: allocating about $\sqrt{n}$ assignments to each reviewer will ensure each possible pair is seen by at least one reviewer. E.g., with $n = 100$ assignments, 11 reviews each is enough, 400 assignments would need 20 reviews per student, etc.

Ongoing research

Focal questions

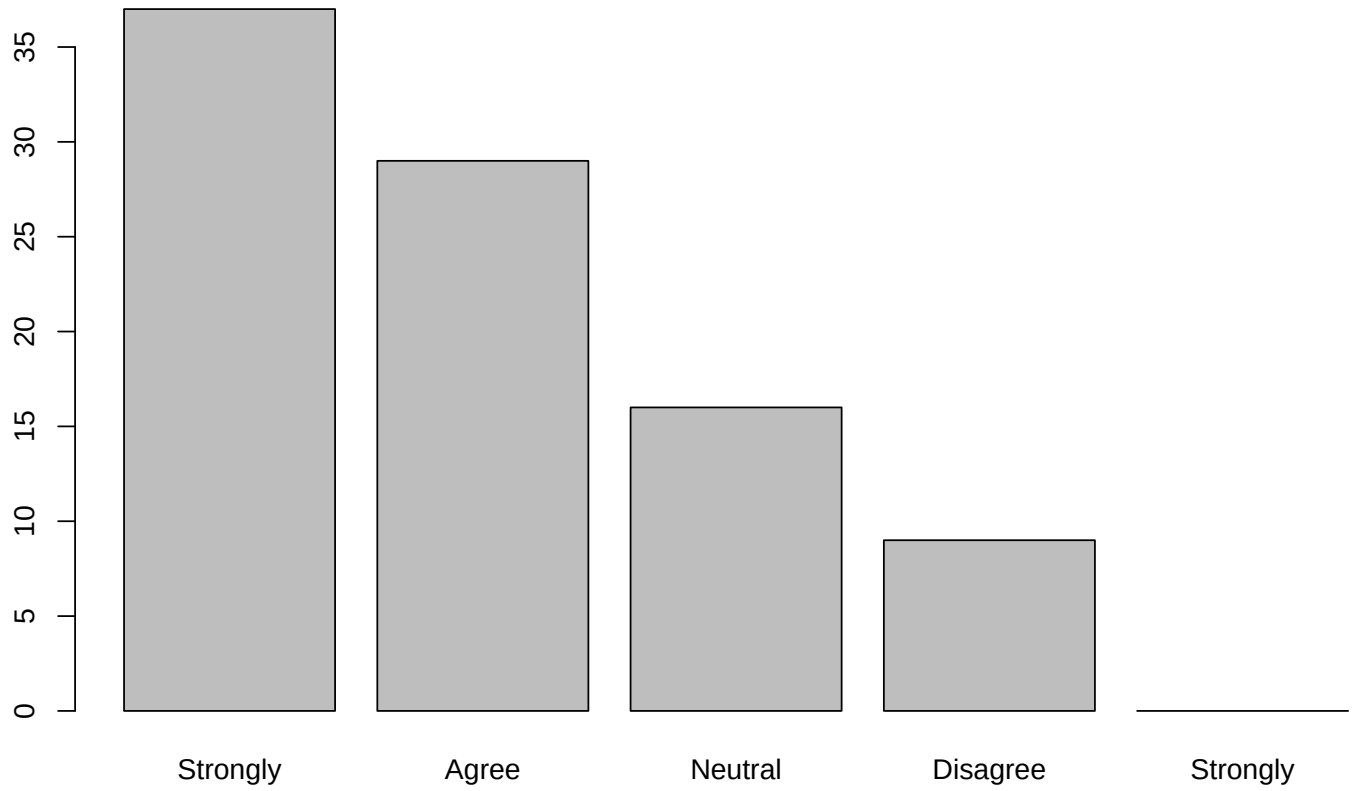
- Investigate disputes between grades assigned by independent markers and peer-assessed grades. Are disputes more often resolved in favour of one or the other?
- Measure correlations between grades and reviewer weights. We can expect the top students to be good at everything, but will some middling students turn out to be star reviewers?
- Do the automatic grade calibration weights mean anything? How stable are they, and do they correlate with author ratings?
- Do social changes occur (e.g., the kind of postings in electronic forum discussions) as a result of peer assessment activities?
- What do students think of peer assessment?

COMPSCI 101 SS Assignment 1

- 105 assignments
- Each reviewer marked seven assignments
- A team of three tutors independently marked each assignment
- Marks agreed $\pm 10\%$ for all but four
- Lecturer reviewed the four discrepant assignments and concluded in two cases the students were wrong, and in *the other two cases the tutors were wrong*

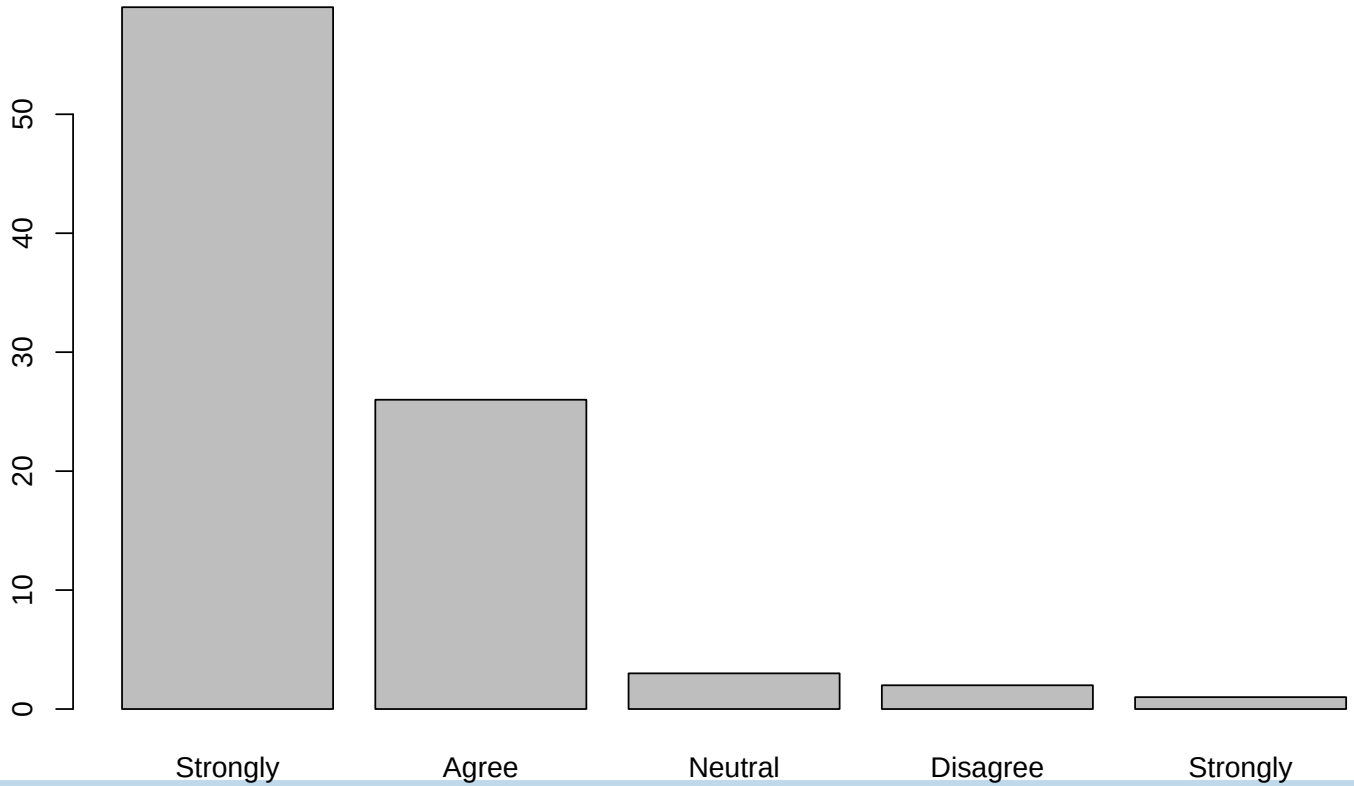
Ease of use

The web interface was easy to use



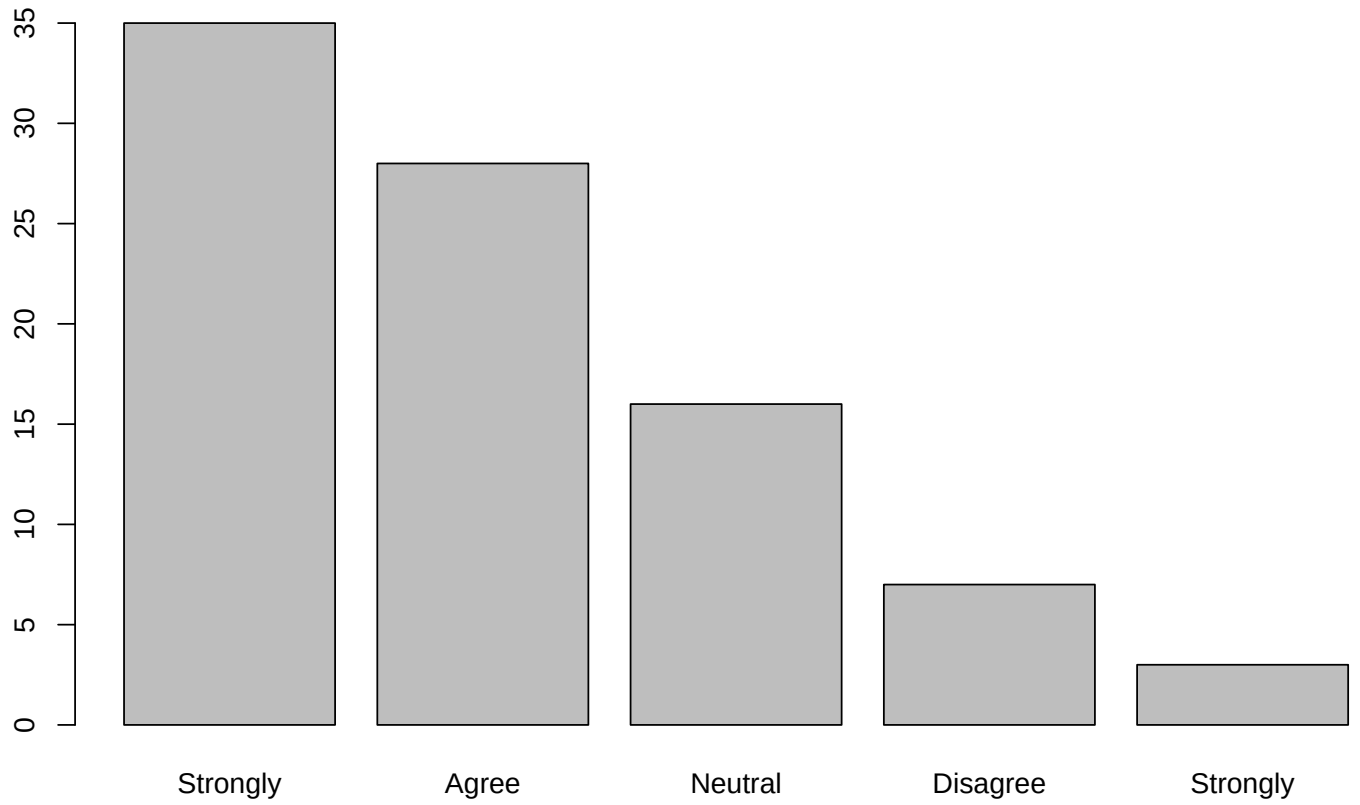
Response

The system responded quickly, and did not leave me waiting



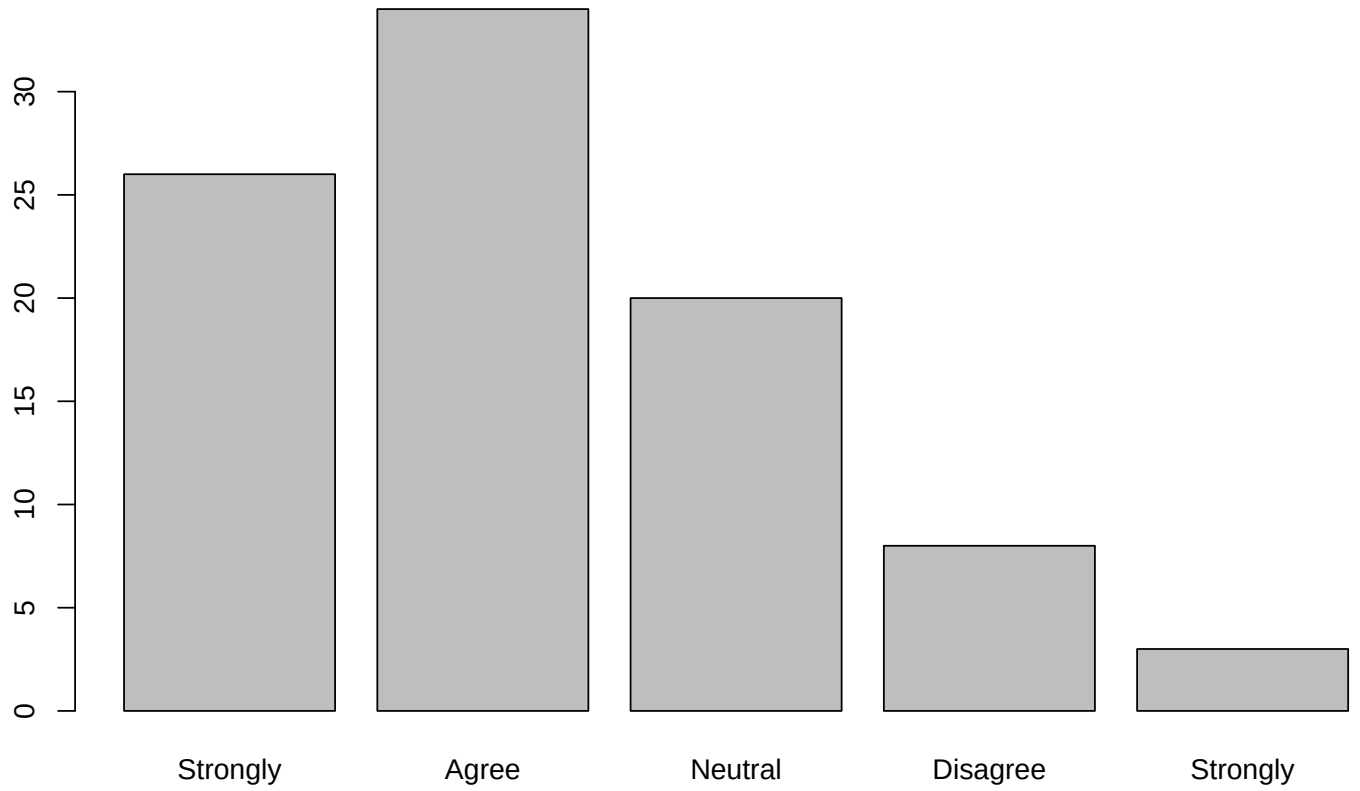
Reading

I feel comfortable with other students reading my work



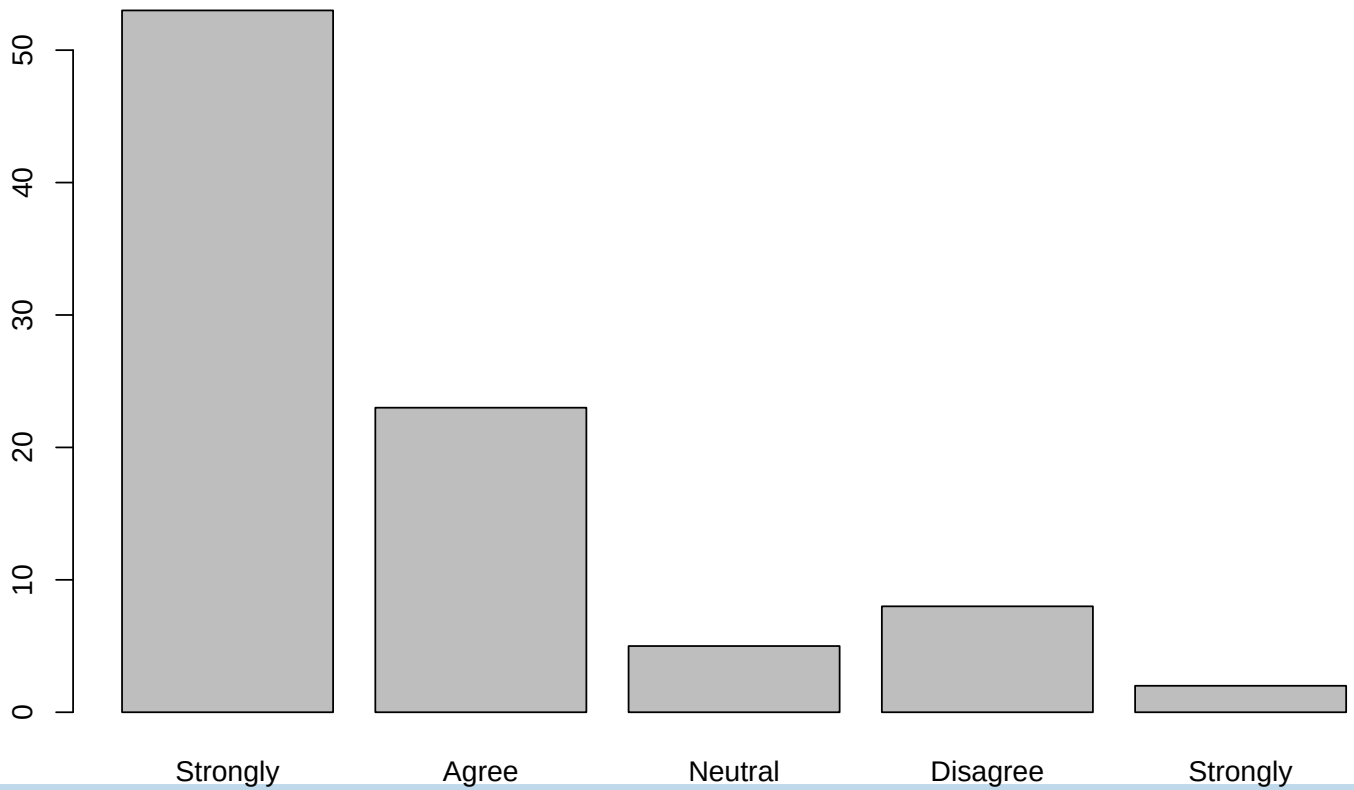
Marking

I feel comfortable assigning marks for other students



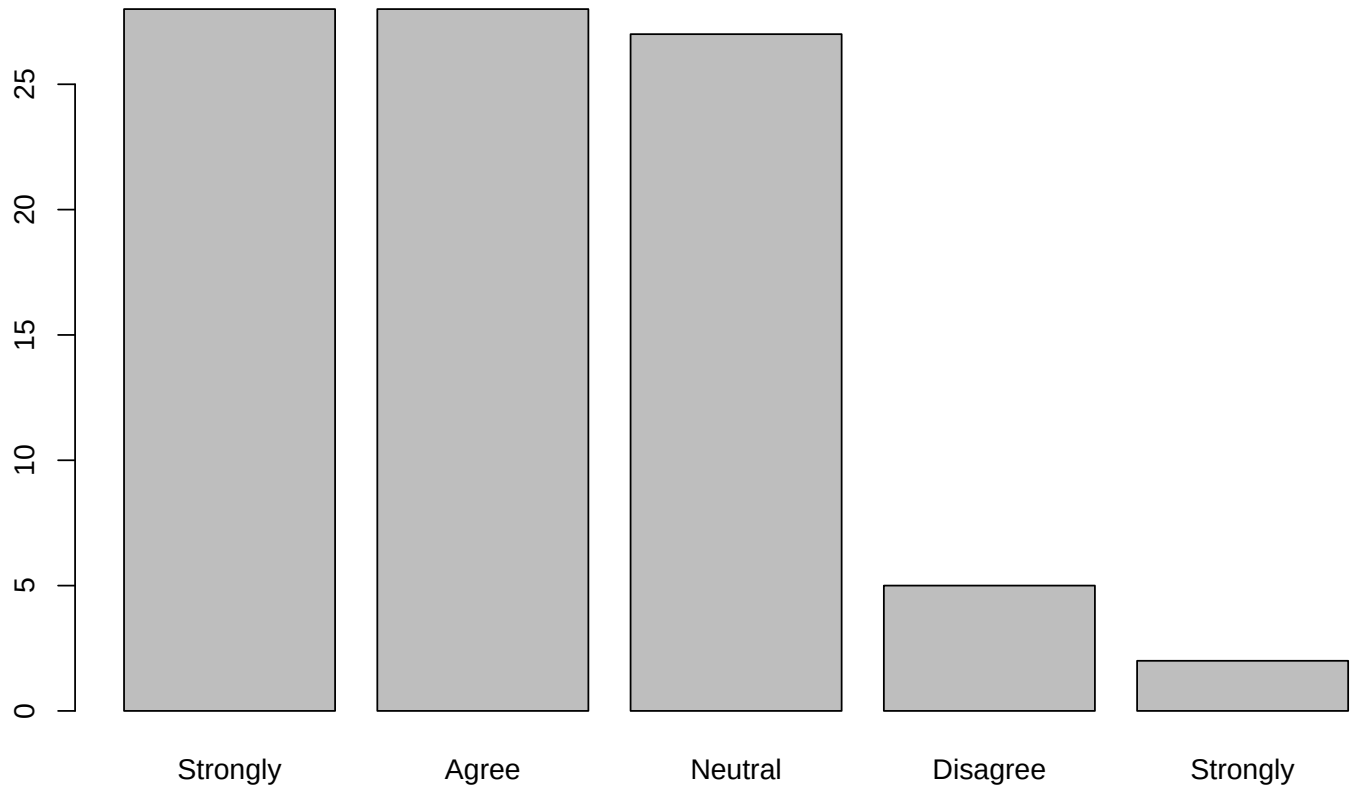
Mistakes

Marking other students work can help me spot mistakes in my own work



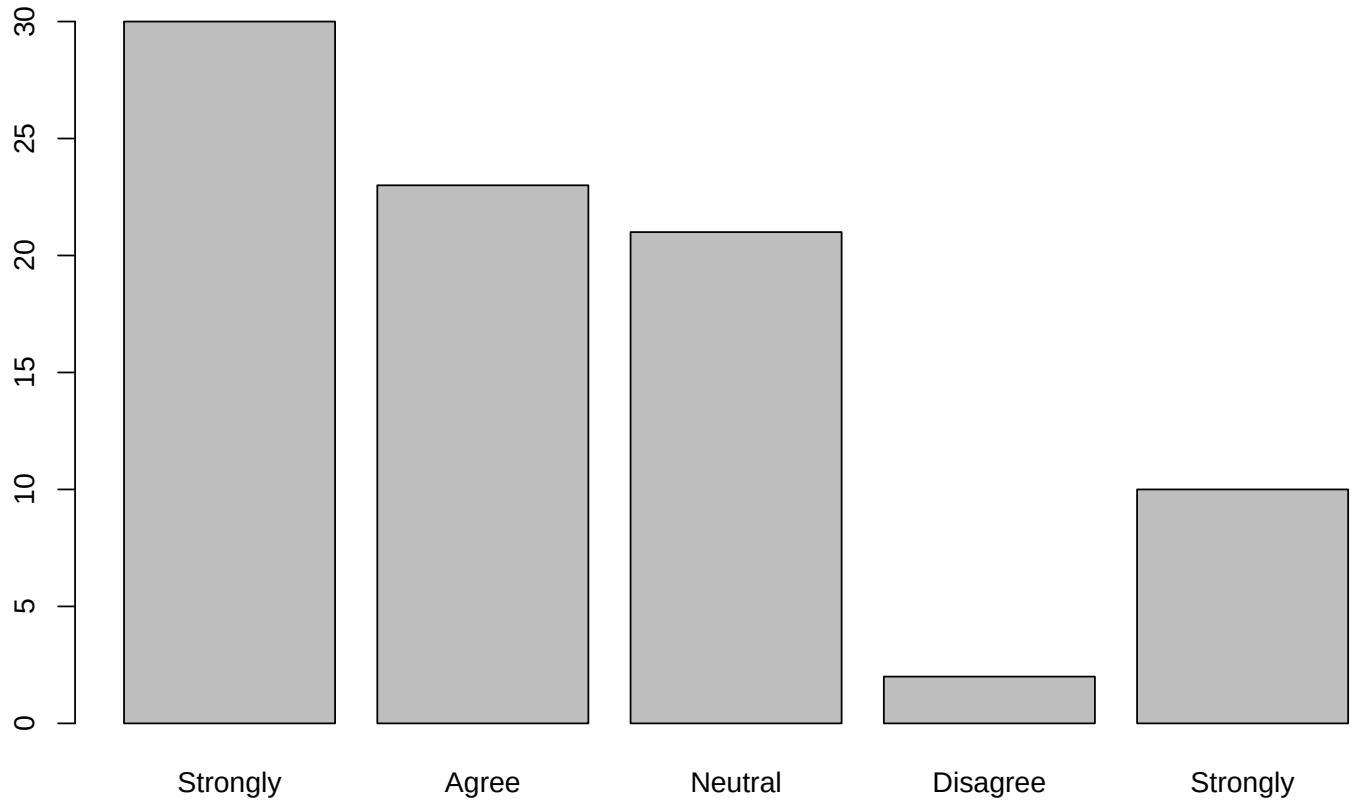
Comments

The comments from other students were helpful to me



More?

I would like more assignments to be peer marked



What did you like most?

- "I liked the way that reading other students work sometimes helped me realise the mistakes in my own work."
- "It was interesting and beneficial to see what others had written in their answers. Not only did it expand my knowledge of the subject matter but it gave me a better understanding of what makes a good answer"
- "I really enjoyed being able to see and comment on other students' work. It has given me a new perspective on the way I read my own work. I have a tendency to throw all my thoughts into an assignment and expect the marker to understand what I mean by wading through it. I think I am already trying to communicate more effectively by being more concise."

Dislikes

- "Students do not mark properly, some of them don't even read assignments properly I gathered that from comments I received."
- "Some people can have different point of views, some people might even have unique view (by thinking into details. . .while others are just ignoring some facts) and hence produce different marking results."
- "This process can be fairly time consuming and if, say, it was to be appended to every assignment, it would add significantly to workload, unless there was a corresponding reduction in asst scope."

Conclusions

Conclusions

- There are no longer any barriers to prevent the routine use of peer assessment in any (every) University course
- Low cost of adoption. You can use your usual assignments, or (gradually) adapt them to include more qualitative content
- Grading rubrics are effective at ensuring reliability. Grades can be as good or better than tutor marking. Feedback is far superior
- Student survey responses suggest the "right kind" of learning is happening