ABC Judging explained

Craig

2025-10-01

Welcome to ABC judging, explained!

Every year, there are questions around how the judging for the ABCs *really* works. Hopefully, this description both demystifies the process and provides an opportunity for community feedback, so that anything that has been overlooked can be flushed out into the open and addressed.

In an ideal world, we might have a specific set of judges with equal knowledge across all subjects that could watch all talks without fatiguing and judge every talk in a uniform way. In that case, we would tabulate scores by simply adding them up for each participant and the highest score would win.

But, we don't live in that ideal world. No one can be in more than one place at a time. We have different expertise and appreciate different aspects of talks we see. Some people are very critical and others are less so. Energy levels fluctuate across the day. In addition, conscious and unconscious bias impacts our score-giving whether we like to admit it or not. The judging procedure is therefore built around a few key "rules". Judges are not micromanaged, but instead go to talks that they would like to go to. They judge as many talks as they would like. We do try to make sure that at least a couple of judges are at any given talk. To make all of this work, the more judges we have, the better.

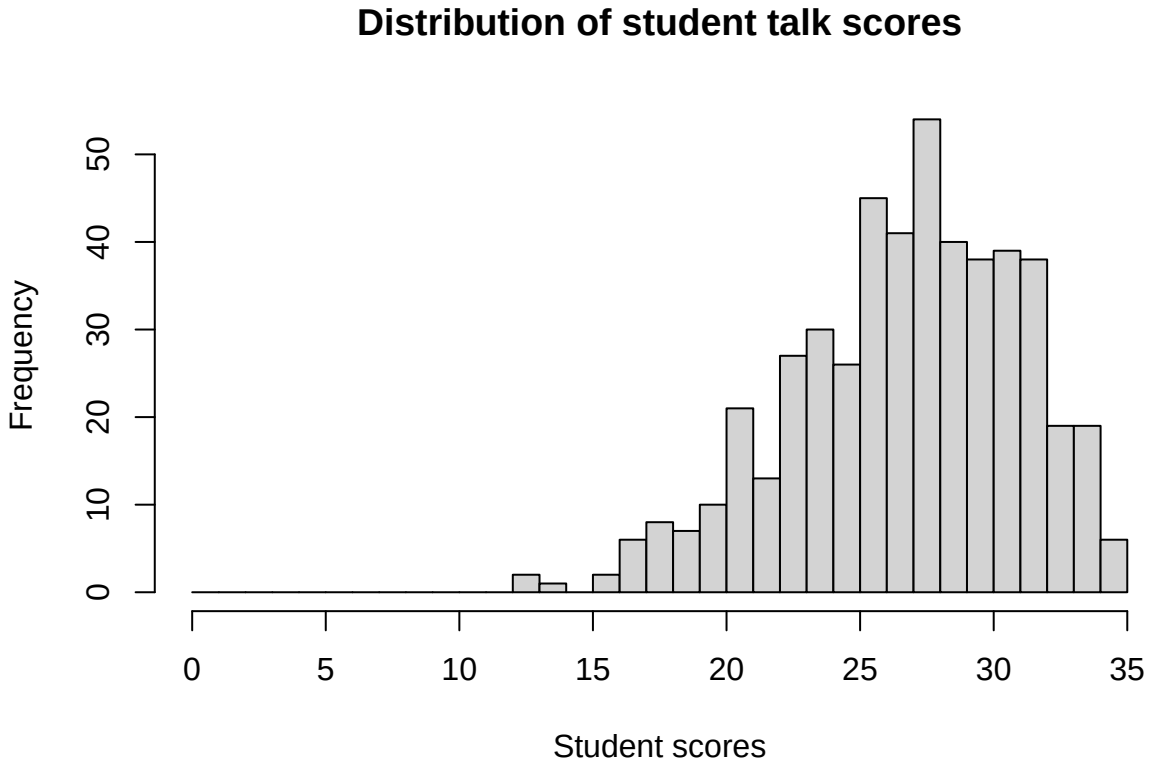
In the judging form, we ask judges for a few pieces of information.

1. Judge code - this is an anonymised judge code. The judge remains anonymous, but we can track how a judge scores across different talks
2. Student name - this is from a drop-down menu so that it is error-free
3. Scores
    - Bicultural confidence and competence (1-5 points)
    - Delivery (1-10 points)
    - Science within context of storytelling or traditional presentation (1-10 points)
    - Engagement (1-10 points)

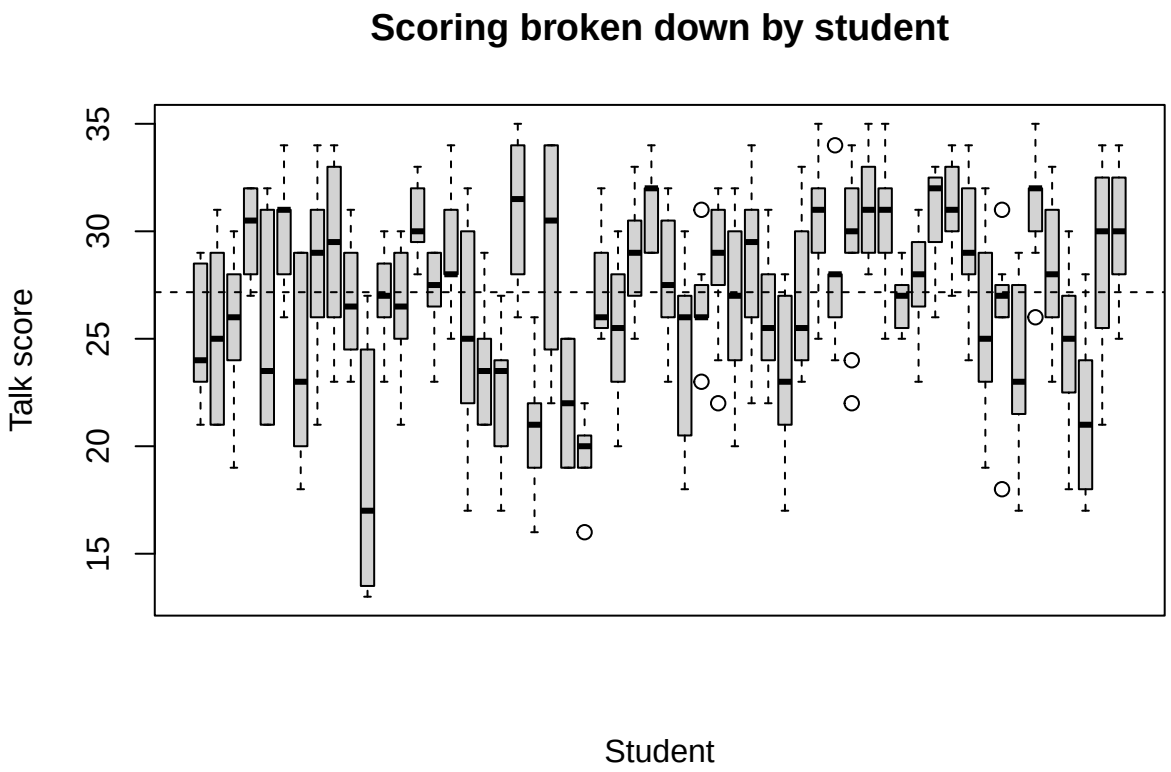
Scores are collected through a Google form, which automatically populates a spreadsheet that is downloaded for the analysis, which is done in R. This enables us to avoid manual mistakes, and to ensure consistency across the years. Scores are tabulated for each student, by each judge as sum of the four scoring categories together.

```
# Calculate the scores as a sum of the four given scores
scores$Total<-scores$`Bicultural confidence and competence`+
  scores$`Science (within context of storytelling or traditional presentation)`+
  scores$Engagement+
  scores$Delivery
```

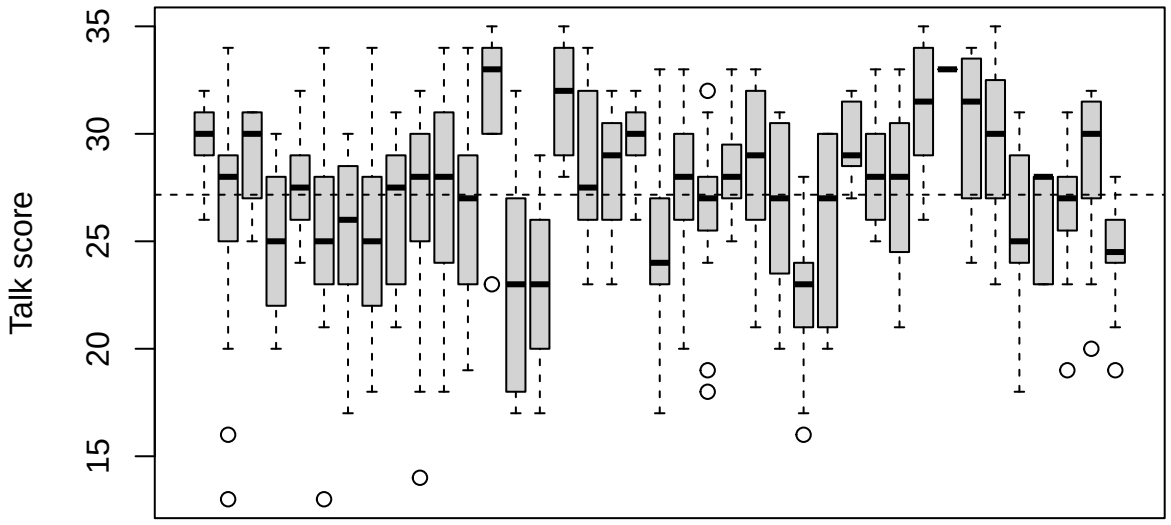
First, examine the distribution of all scores for all students from all judges:



While the median score was around 28, we can see a wide range of scores. We can look at the distribution of scores broken down by student:



The dashed line shows the median score. We can see clearly that some students did much better than the others. It should be easy to decide who the winner is, correct? Let's now look at the same data, this time broken down by judge:



Hmmm, it appears that some judges seem to score much higher than other judges. Also note that different judges have more or less variation in their scores. I guess it is just bad luck if you get harder graders and good luck if you get easier graders? Maybe it doesn't have to be...

To overcome this inherent difference between judges, Several folks at SBS (Daniel Stouffer, Matthias Dehling, and Sarah Flanagan) developed a way of normalising this data to account for this variance using a linear mixed effects model. In it, the total measured score is modeled as a dependent variable of both judge and student. But the judge is considered a fixed effect, meaning that the model predicts student-to-student variation that is not explained by the judge. The score that is not attributed to the judge is then taken as the effect the student had on their score. In this way, scores can be predicted that would reflect a more perfect world, where the variation introduced by different judges is (mostly) removed from the equation.

Even with the use of such a nifty model, it can be somewhat unclear which person should get the prize. To account for this, the model can be used to simulate the results using the uncertainty in the model. These simulations can be thought of as "different day - same rules". On a different day, all other things being equal, how often would the same person win?

We answer this question by examining the rank of the students across the simulations and determine who came in first place more often.

```
##                     Student Degree Discipline  Wins Seconds Thirds
## 28 student  34  - phd - bel      p        BEL 10517    4425   2645
## 45        student  5  - bsc      b        BSc  5228    5137   3602
## 14 student  21  - msc - bsa      m        BSA  2902    3549   3346
## 46 student  50  - phd - bel      p        BEL  2655    3925   3699
## 51 student  55  - phd - sdc      p        SDC  1168    1996   2427
## 38 student  43  - phd - bsa      p        BSA   875    1880   2473
```

In this case, student **student 34** came in first place in **10517** simulations while student **student 5** came in first place in **5228** simulations. While on another day student **student 5** may have gotten the prize, on average we estimate that **student 34** would come out ahead more often.

The final step is to collect winners. We choose the overall winners by degree type (BSc, MSc, PhD) first, and remove them from the pool. Then we pick out a MSc and a PhD for each category

```
## [1] "Summarising the results"
```

```
##         award                   winner
## 1         BSc        student  5  - bsc
## 2 Overall_MSc student  21  - msc - bsa
## 3 Overall_PhD student  34  - phd - bel
## 4     MSc_BEL student  16  - msc - bel
## 5     MSc_BSA  student  9  - msc - bsa
## 6     MSc_SDC  student  8  - msc - sdc
## 7     PhD_BEL student  50  - phd - bel
## 8     PhD_BSA student  43  - phd - bsa
## 9     PhD_SDC student  55  - phd - sdc
```