Uncertainty in Competitive Selection Processes

Bert van den Berg, April 2021

Introduction

Competitive selection processes are widely used to allocate research funding and are also used in events like hackathons and staffing competitions. Just as for some sports (e.g., gymnastics and figure skating), selection competitions require experts to assign numbers to qualitative assessments. How accurately the experts can do this is central to the quality of the competition results. Well-structured scoring rubrics (or language ladders), training for those doing the scoring, as well as structured submission templates that encourage applicants to provide the relevant information all contribute to reducing scoring variability. For many selection competitions, scoring variability remains an important source of uncertainty in the final results.

This article explores scoring variability, its impact on the selection among submissions, and approaches to reduce or deal with this uncertainty[1].

Scoring Variability

First, let's look at criterion scoring. A language ladder can help with scoring accuracy by providing markers an objective structure for assigning numerical values to their qualitative assessments. Good language ladders describe in objective terms what is required for a particular score on a particular criterion. Ideally this should mean that well trained experts will assign virtually the same criterion score as another marker scoring the same submission.

In fields of endeavour that are dynamic, the selection criteria and the scoring rubric may struggle to fully describe scope of possible submissions, and the appropriate numerical values. Exacerbating the problem, those providing the scores tend not to be thoroughly trained in scoring for a particular competition, and may have limited experience scoring any kind of submission. As an example of what happens when markers do not receive substantial training, markers can be prone to ignore the language ladder and favour scoring in the top half of the scoring range, as shown in the attached figure. This compresses the resulting submission scoring distribution into a narrower range, making it harder to pick the 'winners'.

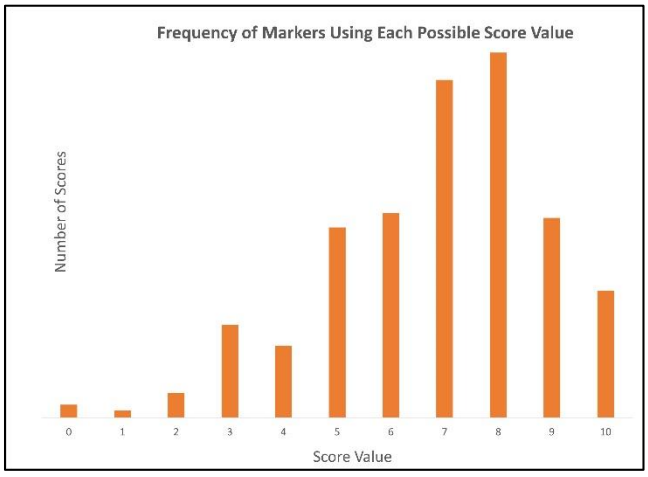

Figure 1: Frequency distribution of Criterion Scores

While heavily structured activities (such as figure skating and gymnastics) often see variability in scoring between judges in the scale of five percent of the

[1] This article is derived from a series of LinkedIn posts in April 2021.

scoring range, it is not uncommon for scoring variability in staffing processes or research funding to be in the range of one third of the scoring range.

Understanding the Uncertainty in Competition Results

The central activities in competitive selection processes are for markers to score submissions which are then compiled to arrive at a rank-ordered list of submissions. While there may be a tendency to accept the rank ordered list and move on to awarding the winners, it is worth analyzing the quality of the competition scoring to understand the confidence that should be placed in the results. This analysis starts with a plot of the rank-ordered scores.

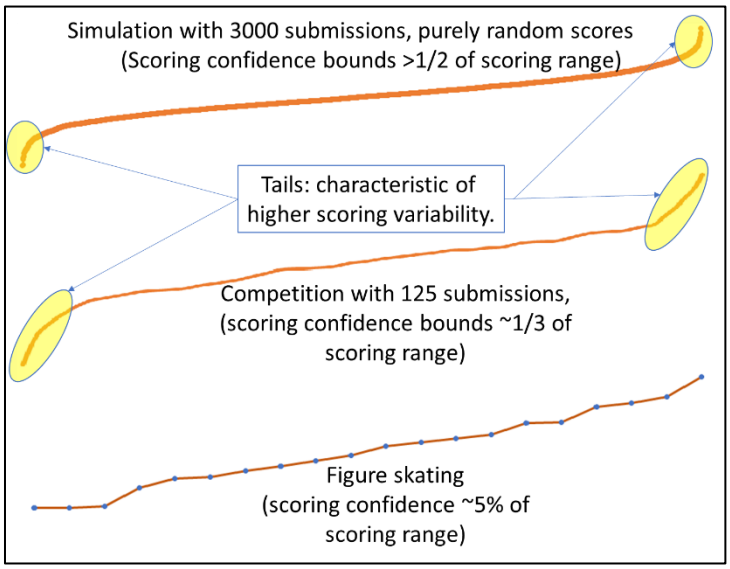

Figure 2: Three different Competition Results

Consider Figure 2 which shows three different competition results. The top trace presents the rank-ordered scores for a synthetic data set of 3000 projects, each with 20 marks.

The middle curve is for a competition with slightly more than 100 submissions, each with between two to four marks. The bottom data set is for a figure skating competition with 20 competing teams, each marked by the same nine judges[2].

What can we see from the rank ordered plot of the competition scores? First the slopes of each curve (as plotted) is generally shallow. If the boundary between winners and non-winners is in the flat zone, scoring uncertainty will be more important than if the boundary is in a part of the trace where the slope is steeper (i.e., the ends of the top two traces). This is demonstrated by the uncertainty analysis described further below.

The second feature of curves to note are the 'tails' at the start and end of the top and middle curve. Since the top curve is for randomly generated scores, we know that this is a feature of random data – essentially there are fewer submissions assigned a set of marks that given them substantially more (or less) than the middle range of the scores[3].

The second curve is from an actual competition. With regards to its 'tails', it is appealing to think that the curve represents accurate scores for a competition where there were a few submissions much stronger than the rest, as well as a few submissions substantially weaker than the rest. However, the similarity to the fully-random scores trace suggests a high level of scoring variability. The uncertainty

[2] One competition's technical scores from: https://github.com/mengyazhu96/figure-skating-analysis.
[3] If you create a scoring frequency chart from the submission scores, the top and bottom scores show up as low-frequency 'tails'.

analysis will help us understand whether these tails are a result of variability in the scores or a stratification in perceived quality of the submissions.

Two aspects for figure skating competitions are different from the more industrial applications of competitive selections: 1) figure skating uses a very structured scoring process, and 2) each judge marks each entry. This results in scoring with less variability, and thus less uncertainty in the competition results.

Assessing the Scoring Uncertainty

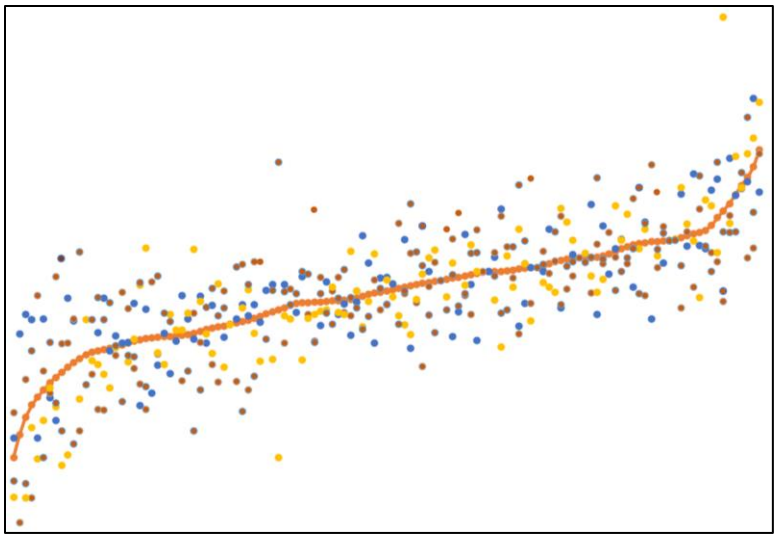

Figure 3: Scoring Variability

We can plot the markers' scores on the rank-ordered plot. This is shown in Figure 3. In the example data set we can see that the marks cover a significant fraction of the rank-ordered-traces' scoring span. This is a strong signal that there is significant scoring variation. This means that if a different group of markers were assigned to any particular submission, its score, and its position in the rank-order would likely change. We need to better understand how much it might change. However, the substantial scatter in the dots, makes it hard to further quantify the uncertainty in the scoring and rank-ordered submission list based on the markers' scores.

Another way to characterize the scoring variability is by calculating the variance of the scores. Since we generally have only a few scores for each submission, and markers repeat on other submissions, it makes sense to estimate the variance (or standard deviation) of the difference of marker scores from the relevant submission. Figure 4 shows the bounds corresponding to two standard deviations of these scoring differences. In other words, we can be confident that 95% of the markers' scores will fall between these two bounds, and consequently that the submission scores will also fall somewhere between these bounds. In this example the uncertainty bounds represent +/-20% of the scoring range, and are visually similar to the extent of most of the scores in figure 3.

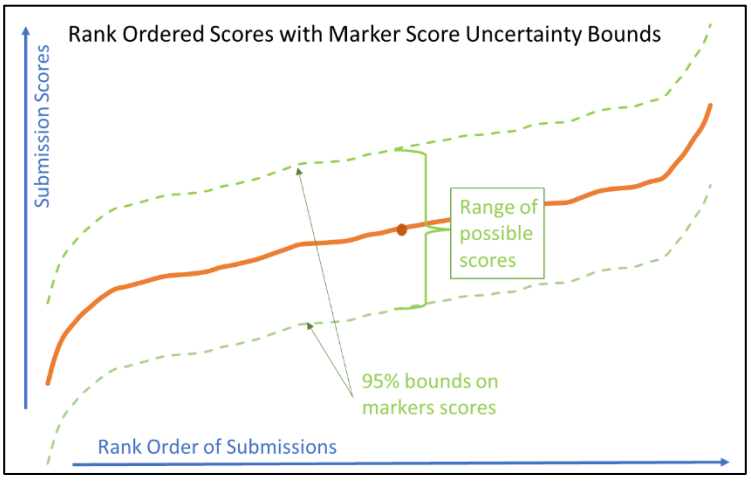

Figure 4: Scoring trace with 95% Marker Score Confidence Bounds

Assigning some, but not all the markers to score a submission is akin to a partial survey of a population of markers[4]. Consider a competition where four markers out of a pool of 12 markers are assigned to score each submission with a maxim score of 100 points. Based on the sample size calculator, we can be 95% confidence that markers' average score for that submission will be +/- 42 points of the scores that the submission would receive if all the markers scored it. Note that this estimate is based only on sampling theory and pays no attention to the effectiveness of a language ladder or the marking expertise of the people doing the scoring[5].

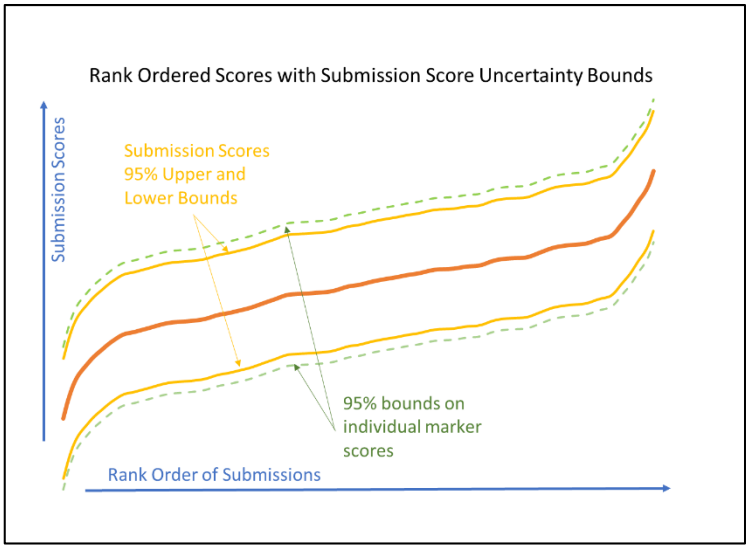

Figure 5: Submission Scoring Uncertainty Bounds

Since we have already created a bounding estimate for the markers' scores using the standard deviation of scoring differences (above) we can apply the sampling theory confidence bounds (+/-42%) to this span. Thus, we can be 95% confident that the submission scores will be +/- 17% (20% X 84%) of the scoring range of the observed submission score.

The Scoring Cut-off and the Uncertainty Zone

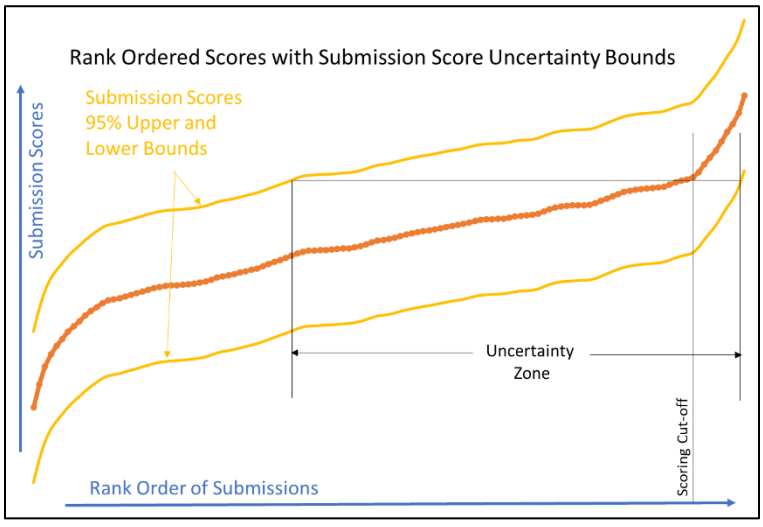

Figure 6: Uncertainty Zone

While it is interesting to understand the uncertainty in the scoring, it is particularly important to understand the impact of scoring uncertainty in the rank-ordered submission list. There is a simple way to look at this graphically. First draw a vertical line through the lowest-scoring submission that can be awarded (we'll call this the 'scoring cut-off'). All submissions to the right are nominally 'winners'.

Then draw a horizontal line through the submission at the cut-off. We will

[4] See this page for more information about sampling populations: https://www.surveysystem.com/sscalc.htm.
[5] The bounds provided by survey sample size calculations can useful for assessing the quality of the language ladder and marking expertise.

call the span of this horizontal line between its intersection with the uncertainty bounds the 'uncertainty zone'.

In this example the competition envisages selecting 12 submissions. We can see that one of the submissions are to the right of the uncertainty zone – we can be confident that its score is higher than the cut-off score. However, for the remaining 11 submissions at or above the cut-off, the uncertainty in means that they are indistinguishable from the 81 submissions in the uncertainty zone to the left of the cut-off.

Dealing with Uncertainty

The uncertainty analysis identifies which set of submissions have scores sufficient to be eligible as 'winners'. Given that there is uncertainty about the rankings in selection competitions, what should be done?

When designing competitions, the following can help reduce the uncertainty in submissions' rank-order:

1) Lower the scoring uncertainty by reducing marker scoring variability. This can be realized by training markers and interventions by staff to ensure markers are following the scoring rubrics.
2) Significantly increase (i.e., double) the number of markers per submission, which increased confidence in the submissions' score. For the example above, going from four of 12 markers scoring each submission to eight of 12 markers scoring each submission reduces sampling uncertainty (and thus the submission score uncertainty) bounds by half[6].

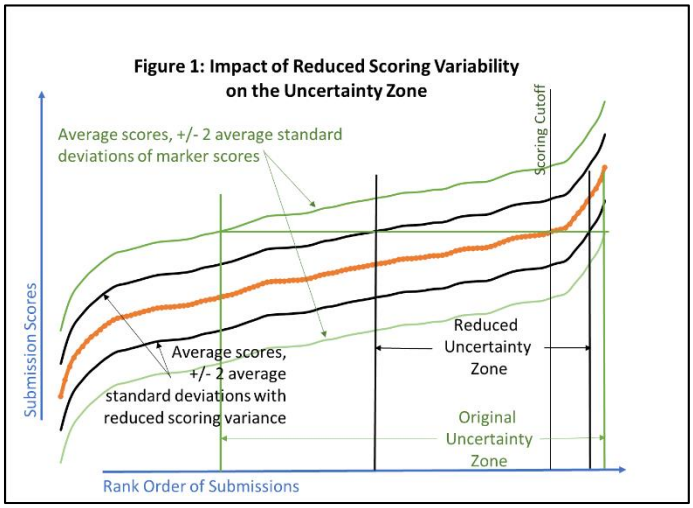

Figure 7: Impact of Reducing Scoring Uncertainty

[6] Unfortunately, if the pool of markers is large compared to the number scoring a submission, the impact of going from four to eight markers per submission is substantially smaller.

3) Increase the difference between scores particularly in the cut-off zone (Figure 2). This makes the rank-order trace steeper, which significantly narrows the uncertainty zone. To realize this, the scoring rubric needs to be revised to better spread scores, particularly for higher-scored submissions. Markers also will need training, and staff will need to be vigilant to ensure the revised rubric is being followed.

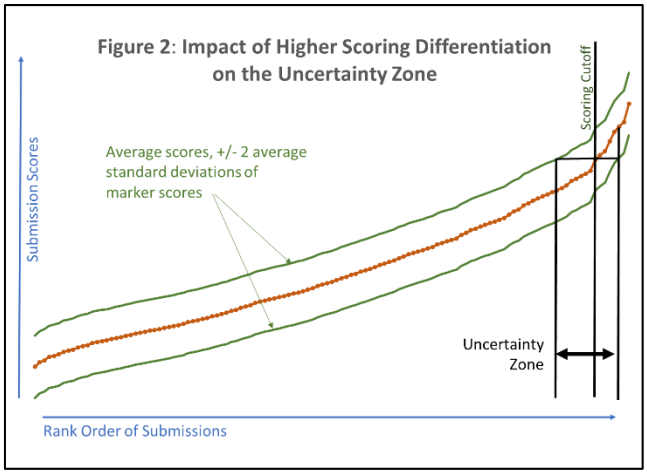

Figure 8: Impact of increased Scoring Trace Slope

4) If there are few prizes/awards, then submissions in the uncertainty zone can be ranked using the markers' tendency to have a favourite among the submissions they score. This document discusses the approach that exploits this tendency.

5) Implement some form of sharing for the submissions in the uncertainty zone (which requires a method for how the 'winnings' are to be shared).

In a context where the uncertainty zone encompasses a significant number of submissions, we need to accept that there are many equally 'good' submissions … and look for another way to choose the 'winners'[7] from among these candidates. Some selection competitions use the scoring phase to identify 'finalists' and then use a different process to make the final decisions (committee discussion, pitch to the judges, …). In this context all the submissions in the uncertainty zone should be invited as 'finalists' and considered equal at the start of the 'finalist' selection process.

Improving our Understanding of Selection Competitions

Selection competitions are often used as a means of allocating scarce resources. Fundamentally, the uncertainty zone demonstrates to the need to improve selection processes. What do we know about the link between the performance of a submission in the competition, and the subsequent impact of the resources allocated to winning submissions? Very little (as far as I know).

Selection criteria tend to be specified by the competition organizers in a highly qualitative development process, often with some input from those responsible for the resources to be allocated. Once the criteria are established their suitability in allocating resources is likely never studied. There is a need to explore the link between the selection criteria and the subsequent impact from resources allocated to 'winners'. Similarly, there is a need to study the effectiveness of eligibility criteria on the impact of the resources allocated to 'winners'.

Since the submissions in the uncertainty zone are indistinguishable, there is an opportunity to experiment in how submissions from within the uncertainty zone are selected. This can improve the

[7] Using a different approach to select the winners avoids the issues presented by rank-order uncertainty. However, the different/additional selection process may have its own selection-quality issues.

quality of future selection competitions by addressing questions like those posed in the previous paragraph[8].

Innovation Growth Labs encourages organizations to undertake structured selection experiments to improve the effectiveness of their programs.

If you are involved in a selection competition, consider asking the organizers about the selection uncertainty in the competition, and how it has been addressed.

[8] The timescale for such studies may span multiple competition cycles, but in a context where the performance of selection competitions may never have been studied this time scale is reasonable.