Concepts and Best Practices for

Selection Competitions

Bert van den Berg

November 2021

Overview

In some competitive selection processes, people sharing the work of scoring submissions against a fixed set of criteria. This includes many government-run funding programs, opportunities offered by charities and public competitions such as hackathons. Important processes for such competitions include:

- defining the selection criteria,
- recruiting experts to score the submissions,
- assigning experts to review specific submissions, and
- combining scores from experts to rank the submissions.
- Analyzing the competition results

This document explores the concepts important in competitive selection processes, including some best practices and issues. It provides an approach to analyzing the quality of the competition results (based on the scores). It also raises questions for further research that might provide further insight into competitive selection processes.

This document is intended as a companion to the CoSeT V7 Overview document. CoSeT is an open-source tool to support data-intensive aspects of competitive selection processes.

Contents

Table of Figures

Introduction to Competitive Selection Processes

Organizations offering opportunities that attract many submissions may engage a pool of experts[1] to share the workload of scoring the submissions[2] against a common set of criteria. The scores are then compiled to arrive at a ranked list of submissions. Such competitions often have multiple winners from among these submissions.

Examples of competitive selection processes include:

- The grant selection processes used by various programs in government departments and agencies, including those used to award academic research funding.
- Hackathons that attract large numbers of submissions from informal teams.
- Opportunities offered by charities and other public organizations.
- Staffing processes aimed at building pools of qualified candidates from large numbers of applicants.

This document introduces important concepts for competitive selection processes and is intended as a companion for CoSeT[3], an open-source tool which supports data management, assignment, scoring and analysis for competitive selection processes.

Concepts for Selection Processes

When selection processes use experts to score submissions against criteria, the general sequence involves:

- defining the selection criteria and launching the competition,
- recruiting experts to score the submissions,
- assigning submissions for each expert to score,
- gathering scores from the experts,
- compiling the resulting scores,
- analyzing the quality of the competition results, and
- distributing feedback on the submissions to the applicants.

In the remainder of this document, these processes are discussed, along with potential issues and best practices. The document also includes the charts produced by CoSeT, and a discussion of the kind of insights that these charts may facilitate.

---

[1] In this document, the people who score submissions will be referred to as markers, reviewers, judges, or experts interchangeably. CoSeT internally uses the term 'Marker'.

[2] In this document, the item submitted to the competition will be interchangeably referred to as application, proposal, project, or submission. CoSeT internally uses the term 'Project'. In some competitions a person (candidate), group or team may be competing, notionally via the submission(s) they have provided.

[3] CoSeT is a Microsoft Excel workbook with tables and macros that help with the management and flow of information for competitive selection processes, particularly assigning markers to projects, and compiling (normalized) scores for projects. It interfaces with online sheets and generates charts for data analysis.

Defining the Selection Criteria

The selection criteria provide a skeleton for consistently categorizing the strengths and weaknesses of the submissions. Language ladders help structure scoring by offering specific (and typically objective) descriptions of what corresponds to a particular rating for a criterion. Well-designed language ladders are essential for increasing the consistency of experts' scoring[4,5].

The selection criteria should be few (three to seven), orthogonal (non-overlapping), and readily understood by both those creating submissions[6] and the experts doing the scoring.

To understand the challenges involved in developing scoring criteria, consider the design of criteria for a research funding program. Typical criteria would address the qualities of the proposed research team (their expertise, and scope relative to the research proposed), one or more criteria addressing the research planned (research novelty, management of research activities), perhaps a criterion related to the student training opportunities, and a criterion related to how research results will be shared with those who can potentially use them. However, evaluation of the research plan can too readily overlap with the student training plan and the plan for communicating research results. For example, if the research plan is poorly organized, or the researchers have a past track record of excellent training, or if the research partners are not interacting with the students, it may affect the scores that markers give for the training criterion, the scores for the research plan and the research transfer criteria. This overlap among the criteria will impact markers differently and will increase score variability.

A variety of approaches are employed to improve the design and effectiveness of selection criteria and language ladder. One approach is to use focus groups from the applicant community, competition domain experts, as well as people who understand criteria design and evaluation processes. Their combined guidance can help create criteria that effectively implement the competition's objectives[7].

Recruiting Experts

Competitive selection processes require people who have expert knowledge of the general topics covered by the submissions and are willing to contribute time to a result they may not directly benefit from. By acting as a judge for a competition, experts may gain:

- insight into the people, organizations or topics being proposed,
- recognition from their community as leaders, and
- acknowledgements for their (often voluntary) contributions.

Beyond having appropriate expertise, it is important that those recruited reflect the priorities of the organization hosting the competition. For example, competitions that will impact particular groups should have those groups significantly represented among the scoring experts. It is also generally seen as important to have a mixture of new and returning experts. The experienced raters tend to have a

---

[4] ProGrid offers a full scoring system that includes customized language ladders for specific programs (and other tools to support competitions).

[5] For an example of a language ladder, see NSERC's Discovery Grants Merit Indicators table: https://www.nserc-crsng.gc.ca/_doc/Professors-Professeurs/DG_Merit_Indicators_eng.pdf.

[6] In some types of competitions (staffing pools come to mind) the scoring rubric may not be available to applicants.

[7] For further advice on designing selection criteria see: https://facultyaffairs.gwu.edu/guidelines-developing-selection-criteria.

deeper understanding of the process and can help train the new people. However, experienced raters may also perpetuate unhelpful habits, and be less current on the topics of the competition.

For competitions where submissions are to be expected on a variety of topics, the expertise of the judges also needs to span these topics. The competition organizers can start the recruitment of judges with anticipation of the number of submissions expected under each topic, and then increase the number of judges for topics that receive more than the expected number of submissions.

Other Concepts Important for Competitive Selection Processes

Equity Diversity and Inclusion (EDI)[8], and anonymizing are both important considerations when developing competitions and selecting experts to judge the submissions. EDI may be reflected in specific aspects of the scoring criteria and may also be a blanket policy that submissions are expected to address.

It is important that those scoring the submissions do not apply their own biases to scoring the submissions they have been assigned. Competition organizers should consider whether they can reduce the influence of unconscious bias from those involved in the selection process. Some organizations ask people scoring submissions to take (online) training in unconscious bias or EDI before completing the evaluation of the submissions they have been assigned[9]. Research shows that evaluators are influenced by the identity-linked traits of applicants (gender, ethnicity). This has led to orchestra applicants performing behind a screen, and some scholarship and HR-screening processes being anonymized.

Assigning Experts to Submissions

Both the submissions to competitions, and the experts recruited, generally cover a spectrum of knowledge and experience. The quality of the submission scoring can be enhanced by ensuring each submission has the best quality of expertise assigned to it, and/or asking experts to rate submissions aligned with their expertise[10]. However, experts must not be involved in the scoring of proposals where they are in a conflict of interest[11] – which often will exclude those with the most expertise relevant to a submission.

Both the number of scoring assignments for each expert, and the number of experts scoring each submission is important. Statistically, as the number of assignments per expert increases, the probability increases that the expert scores a set of submissions that is a representative sample of the competition. Increasing the number of experts scoring each submission reduces the uncertainty in the competition results. Research has shown that more than 10 experts scoring each submission is required to ensure that the average of their scores closely estimate the score that would result if all the available experts scored the proposal[12].

---

[8] Not to be confused with Electronic Data Interchange 😉

[9] For example: https://www.chairs-chaires.gc.ca/program-programme/equity-equite/bias/module-eng.aspx?pedisable=false

[10] CoSeT provides a tool to automate the assignment of markers to submissions.

[11] Experts are also excluded from consideration as markers on individual submissions for reasons beyond conflict of interest.

[12] See the paper by Richard Snell: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120838

If markers judge a large fraction of the competition submissions they will have a good sample for self-calibration, and the chances increase that the subset of the competition they judge are representative of the overall competition[13].

Having each expert judge all the submissions would be (statistically) ideal if the marker's expertise is appropriate (which is the approach often taken for competitive activities like Gymnastics, Diving, and Figure Skating). However, this approach tends not impractical when competitions receive many qualified submissions, leading to the need to each expert judge being assigned a subset of the submissions.

On Assigning Experts to Submissions

Various strategies can be considered when assigning experts to score submissions. A first order approach is to assign any submission to any marker (i.e., ~randomly).

An improvement on the first-order strategy is to ensure that markers are not asked to score submissions where they there is a reason to exclude them (including conflicts of interest)[14]. Sources of conflicts of interest might include:

- Personal involvement in the submission, or a competing submission,
- Role in advising on the submission.
- Personal relationship with the proponent(s) of the submission.
- Family relation to the proponent(s).
- A work relationship with the proponent(s) (often this is time-limited, i.e., in the last N years).

A better assignment strategy is to ask the experts to score submissions that align with their expertise (while not being in a conflict of interest). There are several approaches to matching experts with submissions better aligned with their expertise:

1. **Direct:** In the direct approach each marker is asked to rate their confidence in scoring each submission based on a small amount of information about the submission. This might include the title and a short precis.

2. **Indirect via Keywords:** The indirect approach requires the experts to indicate their expertise in a way that is relevant to the submissions in the competition. This requires the competition organizers to develop the list of keywords pertinent to the anticipated (or received) set of submissions.

3. **Indirect via Topics:** A variant of the keywords approach is to create a set of (sub)topics for the competition, and then group both the submissions and the markers by these topics. Some competitions specify themes or topics in their call-for-submissions, so this approach can build nicely from the call-for-submissions.

As an example, consider a competition focused on climate change. In the direct approach, each marker would be given the titles and public summaries of each of the submissions and asked to signal their anticipated confidence reviewing each submission (i.e., High, Medium, or Low).

---

[13] This is particularly important when considering single normalization (discussed below).

[14] Other reasons to exclude an expert from the scoring of a particular proposal include having previously scored a previous version of the submission, or not being comfortable in the language that the submission is written in. Also, some competitions have experts (potential markers) formally or informally advising applicants.

In the indirect-via-keywords approach, each submission is asked to specify keywords relevant to their submission. These keywords are then compiled for the competition. Each marker is then asked to signal their expertise for each of the keyword areas (i.e., **H**igh, **M**edium, or **L**ow). The markers' self-ratings would then be combined with the submissions' keywords to arrive at estimates of the likely confidence of the markers to rate a particular submission.

Competitions often have a number of eligible topics that help applicants better understand the scope of the competition. The applicants typically must choose one or more topics relevant to their submission. For the indirect-via-topics approach the competition organizers can: ask the markers to select topics relevant to their expertise, allocate experts to topics based on information they have available about the markers, or ask the markers to choose their expertise in the topic areas.

The direct method is expected to produce more relevant expertise information but requires the experts to provide confidence assessments for all the submissions and thus requires more effort from them. The indirect methods require more work (and possibly domain knowledge) from the competition organizers and can be expected to produce less specific expertise utilization than the direct method[15]. When we discuss scoring variability and selection uncertainty the importance of maximizing the use of marker expertise will become clearer.

Making Scoring Assignments

Once the submissions have been received, and scoring experts identified, the competition organizers can assign the experts to mark a particular set of submissions. This process generally seeks to give each expert approximately the same level of workload, each submission the same number of markers (readers), and to optimally match the expertise available with the submissions. The marker exclusions also constrain the possible assignments.

When assigning experts to submissions, different approaches can be taken. For example, the competition organizers can seek to maximize the level of expertise each submission receives[16], or maximize the use of the experts' expertise.

In summary, the following general concepts apply to assigning markers:

- Some markers will need to be excluded from review of individual submissions due to conflicts of interest, language of the submission, etc.
- Information about the experts' expertise with regards to the subject(s) of the competition (or its submissions) is needed to match between markers expertise and the submissions they score.
- The more markers scoring projects and the more projects that markers score, the better the statistically validity of the competition results.

[15] CoSeT supports the direct, keyword-indirect, and topic-indirect methods for classifying marker expertise.
[16] CoSeT attempts to give each submission a similar level of expertise among those assigned, and where possible to maximize the overall use of marker expertise.

Scoring Submissions

In most competitions, there are a limited number of possible 'winners'. This might be based on the number of positions available, the number of awards available, or the total budget available for 'winners'.

The scores provided by each marker (reader) assigned to a project are averaged to produce a score for each submission. These submissions can then be rank ordered based on the scores to see which submissions nominally would receive one of the available awards/positions. Figure 1 shows the rank ordered scores for three competitions:

1. An Ice Dance competition[17] with 20 submissions and nine expert judges who scored all the competitors.
2. A random simulation of 3000 projects, with scores randomly generated.
3. A competition with about 100 submissions and 50 markers.

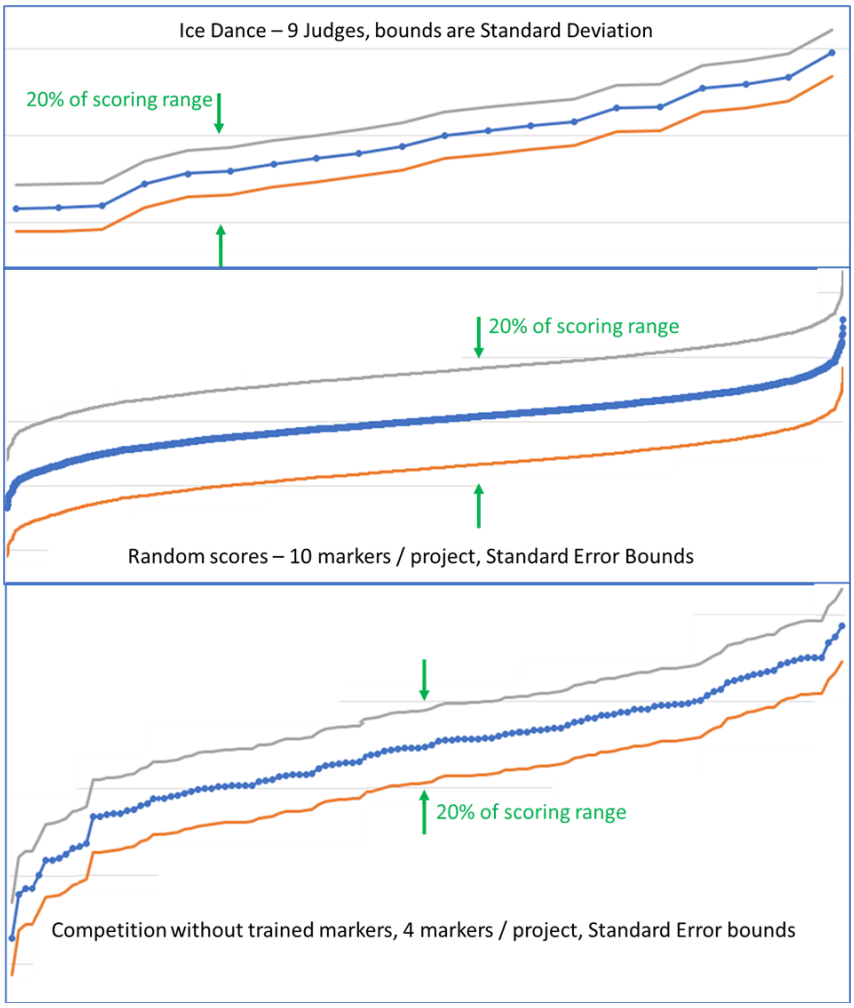

Figure 1: Three competition results

[17] The Ice Dance scoring data is for a single competition's data taken from "Figure Skating Scores: Prediction and Assessing Bias", by Jessica Zhu (2018 Bachelor's thesis). For the purposes of this comparison, the judges scores for the five program component criteria were used (the judges score all competitors on these criteria, while other scoring depends on which elements the competitors execute).

In the case of Ice Dance competitions, we know the top three performers typically get medals. In this case there were 20 teams competing, so 15% of the competition would be 'winners'. For the second and third competitions, the analysis was conducted assuming 15% submissions were to be selected.

Historically, once the scores had been compiled, this selection competition results were decided. However, as discussed in the next sections, it is important to look at the variability in the scores, and determine what parts of the competition are concluded, and where/if further work is required to determine the 'winners'.

Score Variability

We can see in Figure 1 that the variation in scoring is different between the three competitions, with the figure-skating competition having little score variation, and the random simulation having the largest variation. The curious observer might ask: *what impact does score variability have on the competition results*? As it turns out score variability is very important.

In an ideal world, where each person has deep expert knowledge, and applies the same (effective) processes to score the proposals, all scores on a submission would be the same (i.e., the variance between scores on a proposal would be zero). The variability observed in scores from different experts is an important source of uncertainty as to the 'proper' score for submissions and reduces confidence in the competition results[18].

For selection competitions there are a variety of sources of scoring variability:

- The mental processes of people differ leading us give different weight to the same items of information.
- People have perceptual biases that lead us to give difference weight to the same information[19],
- Experts may also tend to score submissions more harshly or generously than their peers.
- Experts may be more generous for submissions with aspects new or exciting to them.

Knowing that there is the potential for variability in the scores, how can we reduce score variability, thus improving the quality of the competition results?

Techniques to Improve the Quality of Scoring

Given the importance of realizing objective and statistically valid scores from the assigned experts, a variety of techniques are applied to improve scoring quality and consistency. These include:

- Educating the experts about the expectations for the competition – what are emerging trends among submissions, which aspects may have been less well understood by applicants or judges, for which aspects of submissions were judges on past competitions most likely to diverge in their scores.
- Using a language ladder (noted above).
- Training the experts on how to score, including the use of the language ladder for the scoring. This may involve running sessions with the experts to score sample applications together, having

[18] Some competition processes follow the scoring phase with a qualitative discussion and ranking process. However, research appears to indicate these do not improve the quality of the competition results.

[19] Well-structured competition processes anticipate common perceptual biases. Training staff to recognize such biases, and to challenge experts to address them can improve the quality of selections.

prospective experts shadow-score submissions before they act as judges for a competition, and even requiring judges to get accreditation.
- Making experts aware of cognitive biases.
- Encouraging experts to self-calibrate their scores across the submissions they score.
- Normalizing scores between experts.

Of the above methods for reducing scoring variability, only normalization is supported by CoSeT.

Measuring Scoring Variability

Before further discussing score normalization let's first discuss measures of score variability.

In a selection competition, the organizers have access to criteria scores from each marker, which can be combined into submission scores for the markers, and overall submission scores. It is useful to look at the variability of scores:

- For each criterion
- Between criteria by a marker
- Between markers on a submission
- Between submissions by a marker

Two measures of numerical variability are relevant: standard deviation and standard error.

> *"Standard deviation describes variability within a single sample, while standard error describes variability across multiple samples of a population. Standard deviation is a descriptive statistic that can be calculated from sample data, while standard error is an inferential statistic that can only be estimated.*[20]*"*

Since each marker scores each submission against all the criteria, standard deviation is useful for looking at the variability in a single marker's scores. To calculate standard deviation(σ) of a marker's scores (X) on n criteria:

$$\sigma = \sqrt{\frac{\sum(X - Xi)\text{^}2}{n}}$$

At the competition level, since a subset of markers typically scores each submission, standard error is useful for looking at the variability of submission scores. The Standard Error (*SE*) of a set of n scores for a submission is straightforward to calculate:

$$SE = \frac{\sigma}{\sqrt{n}}$$

We can see that the Standard Error decreases with the square root of the number of measurements (while Standard Deviation is nominally independent of the number of scores). Thus, doubling the number of markers on each submission should decrease the scoring confidence bounds by $\sqrt{2}$ (since Standard Error is the appropriate metric). The figure below shows the difference between the Standard

[20] https://careerfoundry.com/en/blog/data-analytics/standard-error-vs-standard-deviation/

Deviation confidence bounds and the (narrower) Standard Error confidence bounds for a competition with four markers per submission.

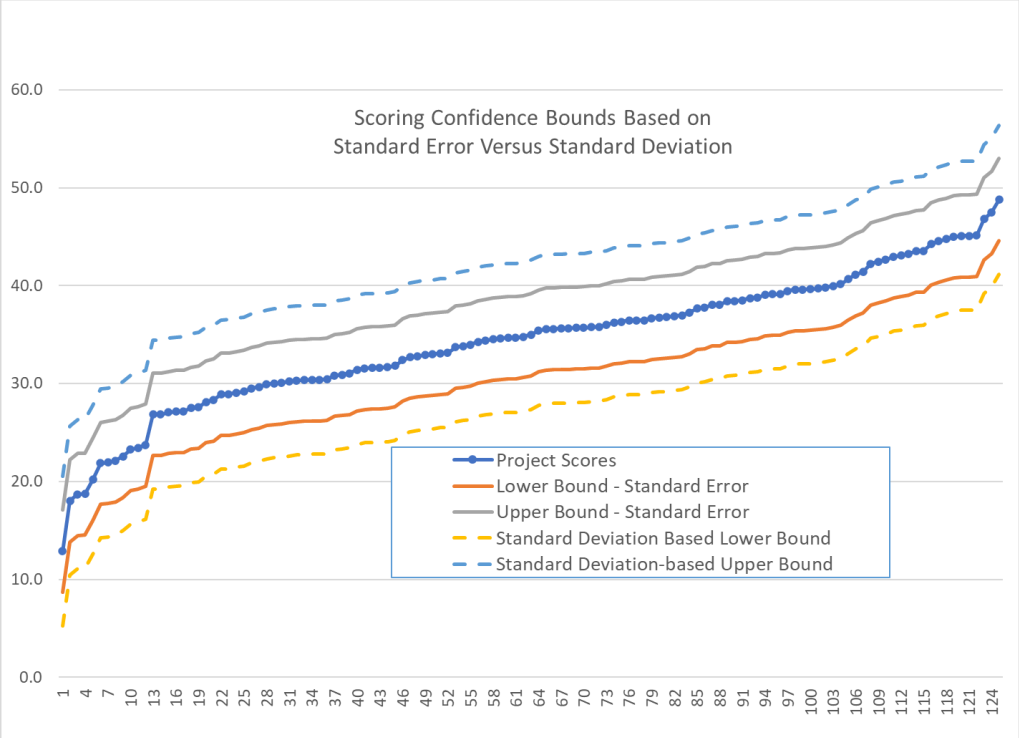

Figure 2: Standard Deviation versus Standard Error Example

Normalizing Scores

In this document normalization refers to numerical processing of marker scores with the objective of reducing the overall scoring variability, and thus the scoring uncertainty. Normalizing can only compensate (partially) for trends observed in the scores. The most obvious approach to normalization compensates for experts' tendency to be generous or harsh in their scoring. This implies calculating a global factor for each marker that reduces their observed generosity or harshness. By reducing this source of scoring variability (between experts) the competition's scoring confidence bounds are reduced.

Two methods of score normalization are supported in CoSeT, and are discussed below. These normalizations apply a consistent approach to all markers, and all their scores.

Single Normalization

If the number of submissions each expert is asked to rate is a significant sample of the competition submissions, then statistically, the average of the submission scores from each expert should be the same. Single normalization works with this idea to produce a consistent average for the total scores provided by each expert.

For single normalization of scores, the competition organizer should first decide on a target average score (K_t) for markers (say 50-65% of the scoring range). Then to calculate the single normalization for each marker factor (K_i), we need to find the average score they give submissions.

$$K_i = \frac{\frac{\sum X_j}{n}}{K_t}$$

The ratio between a marker's average scores and the target average score is the normalization ratio. Divide each of the marker's criteria scores by their normalization ratio to make their average score equal to all the other markers (and thus compensate for harsh or generous bias in some markers).

Double Normalization

In selection competitions, each marker may not be assigned a (statistically) representative sample of the submission to the competition. To compensate for this, double normalization first calculates a relative score (K_{ri}) for each marker on each submission as the ratio of the marker's submission score (X_j) to the average of all the (n_s) markers' submissions scores for that submission.

$$K_{ri} = \frac{X_i}{\frac{\sum Xj}{n_s}}$$

This relative score provides a metric of how the marker scored a project relative to the other experts also assigned to the submission. The subsequent average of a marker's (m) relative scores is their double normalization factor:

$$K_i = \frac{\sum K_{ri}}{m}$$

Both single and double normalization are observed to reduce the standard error of competition scores, and thus enhance the certainty in the (normalized) competition results. It would appear that double normalization is less disruptive than single normalization in terms of changes to the (raw) rank order of the submission scores (the author interprets this as better identifying markers that are more harsh or generous).

Note that single normalization compares the average of a marker's assignment scores to the average of other markers assignment scores (so is largely dependent on the number of assignments each marker completes). In contrast, double normalization compares each of the marker's assignment scores with the scores of other markers assigned to that project, and then compares the average of their relative scores to that of the other markers. Thus, double normalization has the benefit of involving a larger fraction of the competition scores it its calculations, which likely helps to reduce (but not remove) the influence of sampling errors.

The impact of Scoring Variability

Score variability means that the rank-ordered competition scores are generally not definitive. Why? Because a different set of equally qualified experts would likely award different scores to submissions, potentially changing the rank order of the projects. In other words, although the competition may produce a rank-ordered list of submission, the score variability means that for a significant number of

submissions with similar scores, it is often not possible to say definitively whether they are 'winners' or 'losers'[21].

Scoring Confidence Bounds and Uncertainty Zone

Selection competitions are tools for choosing among submissions. The submission with the lowest score that is nominally 'selected' defines the nominal 'cut-off' score. The selection competition's uncertainty zone encompasses all submissions whose scores are within the scoring confidence bounds of the cut-off score.

We can set the scoring confidence bounds from the measures of variability discussed above.

- For the Ice Dance competition, all (nine) judges scored all competitors[22] so Standard Deviation is the appropriate measure.
- For the competition of about 100 submissions, and for the synthetic competition with 3000 submissions, the experts only mark a sample of the submissions (4 for the smaller competition, and 10 for the synthetic competition), so Standard Error is the appropriate measure.

For the figures in this document, the scoring confidence bounds are taken as +/-2 of the appropriate measure of variability[23]. As seen below, the scoring confidence bounds in the Ice Dance competition are narrow (Figure 2), while for the random competition (Figure 3), and the competition with the untrained markers (Figure 4) the scoring confidence bounds are wider – even though the latter two competitions benefit from the Standard Error calculation.

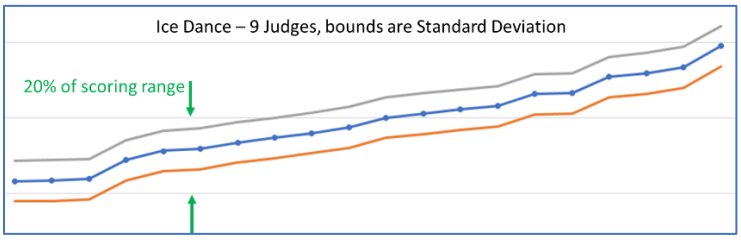

Figure 3: Score Confidence bounds for Ice-Dance Competition

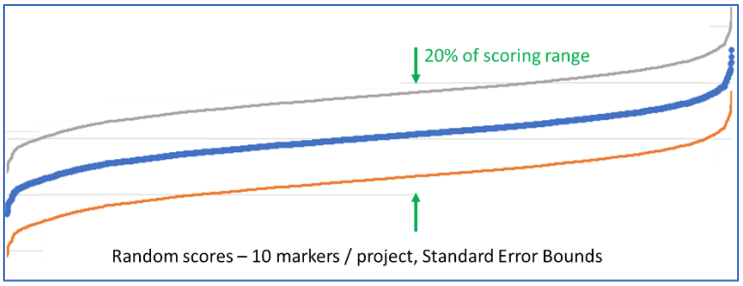

Figure 4: Score Confidence Bounds for a Large Competition with Random Scores

[21] From a statistical perspective, all the scores within the uncertainty zone are statistically indistinguishable, and a different process is needed to define which among these submissions will be awarded.

[22] Recall that for the purposes of this document, the scoring data from the Ice Dance competition **only** includes part of the Ice Dance scoring system, namely the scoring on the five components that are scored in a fashion amenable to CoSeT.

[23] Recalling that +/-2 standard deviations encompass 95% of a normal distribution.

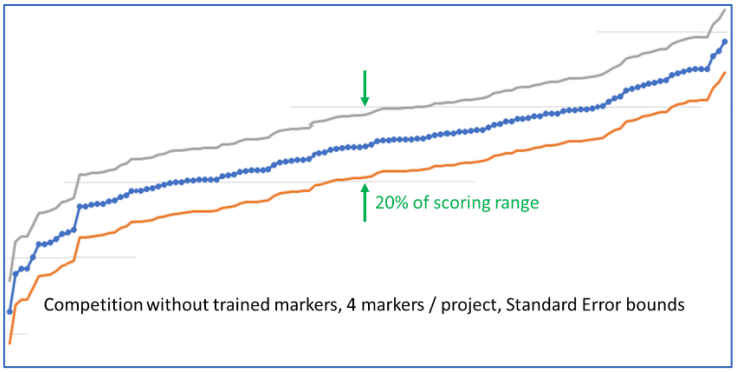

Figure d: Score Confidence Bounds for a Competition with Untrained Markers

Given a particular set of competition scores, scoring cut-off, and score confidence bounds, we can readily identify the span of submissions whose scores are statistically indistinguishable from the cut-off score. The cut-off, and span of indistinguishable submissions for each of the three competitions is shown in Figures 5 to 7.

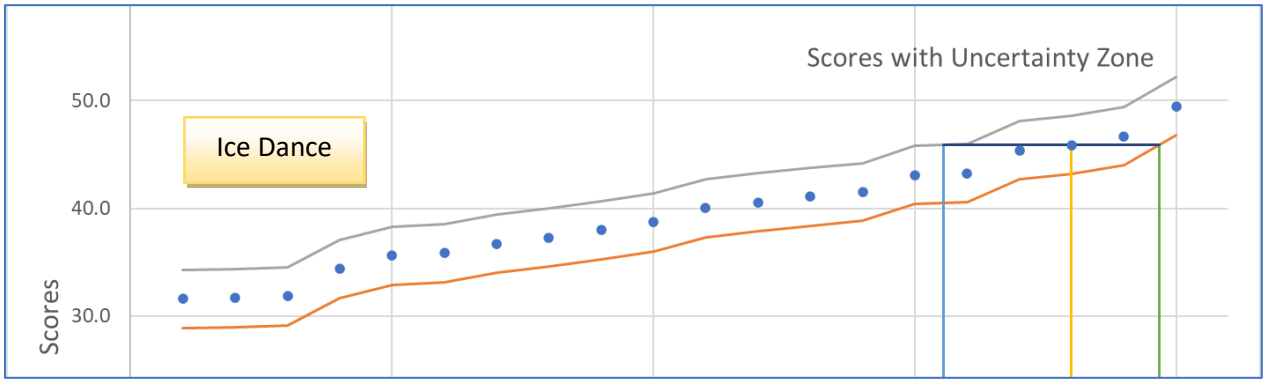

Figure 5: Uncertainty Zone for Scores from Ice Dance

The narrow uncertainty zone for the Ice Dance competition means (only) three competitors having scores that are (by this analysis of part of the scoring system) indistinguishable from the third-place competitors. [As an aside, the double normalization factors for the Ice Dance judges range from 0.96 to 1.03, indicating that individual judges do not generally demonstrate harshness or generosity across their scores. Consequently, normalization has no effect on the competitor ranking.]

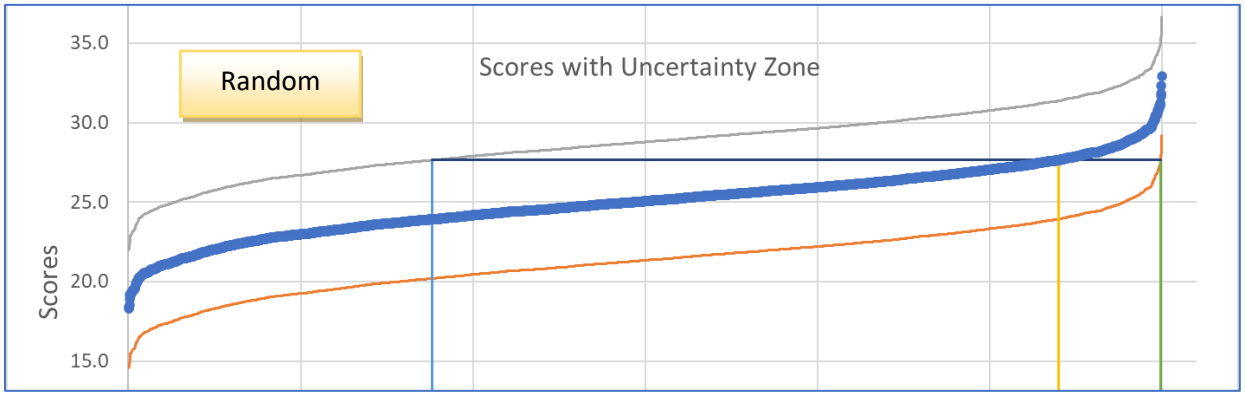

Figure 6: Uncertainty Zone for Large Competition with Random Scores

For the synthetic competition with 3000 randomly scored submissions (10 markers per submission) we can see in Figure 6 that the uncertainty zone is large, extending across more than half of the submissions. This shows the challenge of using a selection competition when there are many submissions. For this data, the double normalization factors range from .97 to 1.03.

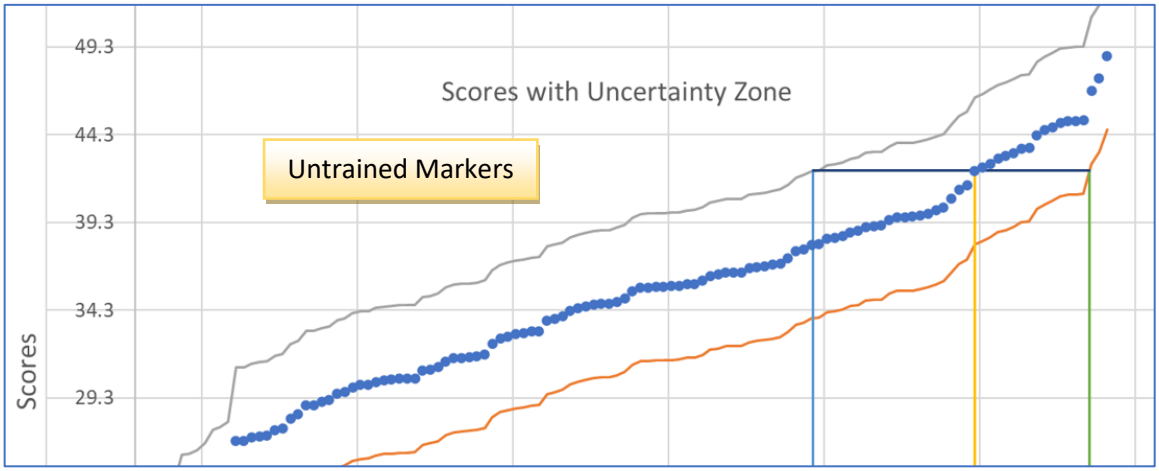

Figure 7: Uncertainty Zone for Competition with Untrained Markers

The competition with about 100 submissions and untrained markers has an uncertainty zone that approximately doubles the number of submissions that must be considered as possible winners[24]. For these latter two competitions we can see that the selection process was only effective in reducing the number of submissions under consideration by about 65%. For competition data, the double normalization factors range from .75 to 1.2.

These figures emphasize the importance of efforts to reduce scoring variability (i.e., training), as highly trained markers are likely to produce competition scores with much narrower standard error.

As an aside, while we would hope that highly trained judges would also make fuller use of the scoring range, we can see that the scoring histogram for the Ice Dance competition is like the scoring histogram for the competition with untrained markers – both groups avoid using the lower half of the scoring range (see Figure 8 below).

[24] Note that the position of the cut-off generally has an important impact on the size of the uncertainty zone. Scoring cut-offs near the top of the competition benefit from the higher slope of the rank-ordered scores curve, producing a narrower uncertainty zone. In contrast, if the competition will make 'winners' of a large fraction of the submissions (i.e., 30% to 70%) the uncertainty zone spans more submissions.

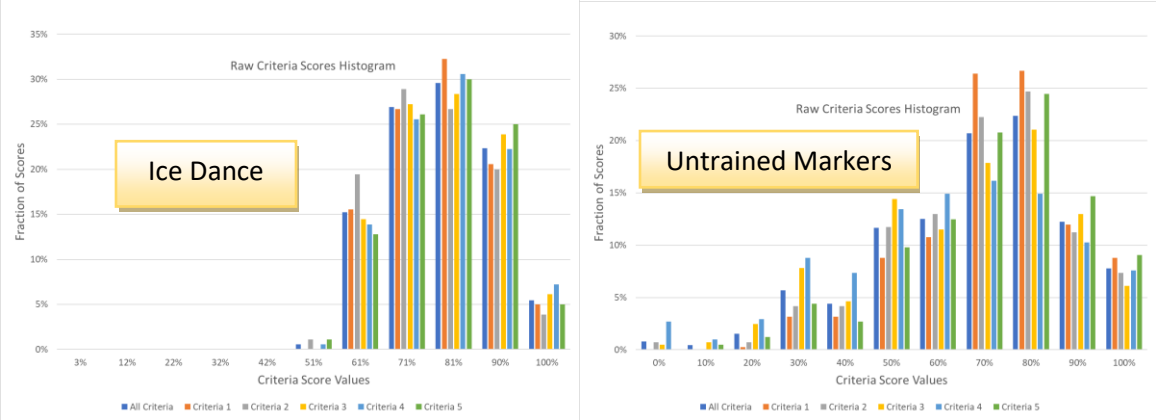

Figure 8: Criteria Scoring Distribution for Ice Dance and Untrained Marker Competitions

The scoring uncertainty and consequential expansion of the number submissions that must be (statistically) considered equivalent creates a quandary for competition organizers. The competition already integrates the information provided by the experts' scores, so new/additional information is required to selection among the submissions in the uncertainty zone. This is discussed in the companion document "Selecting Submissions from the Uncertainty Zone – A Proposal".

Tools for Understanding Competition Scoring Quality

By analyzing the competition results, the organizers can better understand the level of confidence they should place in the results, and which submissions need to be further assessed. Given the impact of score variability on competition results, it is also important to look for ways to reduce the competition uncertainty (while retaining the insights provided by expert judges). This document has already discussed the importance of criteria orthogonality, language ladders, expertise, training and normalization in reducing competition uncertainty. A subsequent section of the document identifies some topics for further research about selection competitions.

The information quantitative available in selection competitions includes:

- Markers' confidence with regards to submissions, keywords or topics
- Criteria level scores from markers on each submission
- Derived quantities like normalization factors, submission scores, standard deviation and standard error.

From these data we can perform analyses that may address questions like:

- Do the markers use the full range of the language ladder when scoring each criterion?
- Which criteria tend to have different scores than other criteria?
- Is there a relationship between the confidence a marker express for a project and the scores it receives?
- How are the project scores distributed (in comparison to the possible scoring range)?
- How wide is the band of possible scores for projects?
- What fraction of the projects are in the uncertainty zone?
- Do markers with a weaker or stronger set of projects assigned get higher or lower normalization factors?

- Are there many projects in the uncertainty zone?
- Is the uncertainty zone in the 'flattest' part of the rank ordered scores?
- Which projects move in/out of the uncertainty zone due to normalization?
- How much do projects change ranking because of uncertainty?
- Is there a relationship between ranking position and the amount a project moves?
- Which markers tend to use a narrow band when scoring a project, which ones make more full use of the scoring range?
- Do markers tend to give a submission the same score for each criterion
- Is there a correlation between the project scores and normalization applied?
- Is there a correlation between the two normalization factors?
- Do higher or lower scores correspond to a trend in the double normalization factor calculated?
- Is there a relationship between the confidence a marker expressed for the set of projects they were assigned and their double normalization factor?

CoSeT currently generates 14 charts from the competition data that allow competition organizers to better understand their competition results. These charts and the associated tables help organizers interpret results and look for trends related to the scoring distribution, scoring variability, confidence of markers, and normalization of scores. Appendix A presents and discusses the analysis charts produced by CoSeT.

Concluding the Competition

The final set of experts' scores are generally combined in a master scoresheet as average scores for each submission. As discussed in this document, scoring variability means that an uncertainty analysis should be conducted ***before*** choosing which submissions are 'winners'. It may also require an additional selection process to choose from among the submission in the uncertainty zone.

Funding competitions run by public agencies tend to involve committee meetings to score and discuss the submissions before the final selections are made. Such meetings may rely on the committees to 'fine-tune' the final selections. However, while research indicates that these discussions among experts improves the understanding of strengths and weaknesses of submissions and does see experts revise some scores, the research indicates that such discussions do not meaningfully reduce the variance between expert scores and have minimal impact on which submissions are accepted/funded[25].

Ideally, the additional processes to select from among the uncertainty zone would include new of information, better justifying the final selections.

Feedback to Applicants

Beyond being informed of their standing in the competition, applicants appreciate feedback on the strengths and weakness of their application. One approach to crafting this feedback is to assign a marker with the responsibility of assembling a coherent set of feedback for the submission (this role may be

[25] This is also consistent with my years of experience with selection processes.

called the 'first-reader')[26]. However, it is typically the competition organizers' responsibility to ensure the feedback is consistent with the competition results and the objectives of the organization.

Information management

The starting information in a selection competition includes the submissions, selection criteria, markers, and their expertise. Further information created during the competition includes the assignments, scores, comments from experts and the final list of ranked submission scores.

Historically, this information was managed on paper, with information sent to and from experts by mail. This has evolved into information managed on office productivity tools and/or customized databases with files shared by email or via online sheets. These approaches have often been supported by templates and customized spreadsheets. More recently, some organizations have invested in bespoke information management systems that include secure extranet websites for sharing information with experts. For some organizations, the use of online sheets and online forms have proven to be useful for managing information and sharing it with the scoring experts.

Topics for Further Research

As discussed in this document, scoring uncertainty and other sources of error can have significant impact on the quality of selection competition results. While organizers can use methods that are likely to improve the quality of selection competition results, there are also significant opportunities to analyze the competition data to better understand the quality of selection results.

Although competitive selection processes have been used for decades, there is surprisingly little research available about these processes. The advent of electronic tools to manage the information for selection competitions means new avenues for research to understand and improve selection processes is viable. Below are some possible research questions for consideration.

Defining the selection criteria

: How does the wording used in competition call-for-submissions, posters and related descriptions influence the decision to participate, and the qualities of the submissions (for example from different EDI groups[27])?

Recruiting experts to score the submissions

: Expert participation is generally voluntary, and people who might apply to the same competition or organization may be more interested in volunteering. Do volunteers who also apply to the same types of competitions trend towards certain scoring patterns? Do the experts who participate have different personal characteristics that bias against some of the types of submissions that the competition organizers may desire (depending on their reasons for participating as judges)? Some people may be paid to compensate for their efforts as a judge. How does this impact the characteristics of the judges? Are there demographic factors that align with particular types of judging biases?

[26] CoSeT support gathering the scores and comments from markers and can compile comments for the applicants. However subsequent editing is required to ensure the comments from different markers provides coherent and relevant feedback to the applicants that aligns with the competition results.

[27] For example, there is research signaling the impact of language on the participation of women in competitions.

Assigning submissions to the experts to score: What other information about the submissions, experts, competition domain and criteria leads to assignments that improve the overall impact of the set of submissions selected? What other attributes of people (beyond self-declared expertise) align with high quality scoring of the submissions? How do different expertise assignment methods affect competition results?

Gathering scores from the experts: Does the time delay between experts receiving their assignments and when the experts' assessments are returned align with the experts professed level of expertise, or the characteristics of the scores returned. What are the impacts of: 1) EDI training, 2) unconscious bias training, 3) training on using the language ladder, and 4) paying judges to participate on the results of competitive selection processes?

Compiling the competition's results: Given the significant variance to be expected among the submissions ranked near the win/lose cut-off zone, there are opportunities to explore how other factors correlate with the success of applicants/applications in the competition. For example, when choosing among submissions in the 'uncertainty zone' criteria different than those the experts scored can be used. Subsequently tracking of the results of these experiments could provide insight as to other criteria that improve the impact of the competitions. Following this line of thinking, submissions could be selected from those in the uncertainty zone:

- randomly … to see if subsequent impact (of 'winners' and 'losers') correlates with their nominal ranking, or
- randomly … and tracking a wide variety of demographics associated with the submission to see if any are aligned with better subsequent impact, or
- based on a characteristic of the submission that might align with future success (i.e., young researchers being more innovative).
- Based on efficiency (lowest level of requested resources)

The research to date on competitive selection processes largely focused on the individual process steps, with little research linking the results of competitions to the competition methods used. Consequently, competition organizers often rely on experience and organizational preferences when designing selection competitions. In this era of 'big data' there are also opportunities to use research to strengthen the performance of selection processes by linking downstream impacts with competition participation.

Feedback from experts on the submissions: does the quantity or existence of feedback from experts correlate to the level or variation in scoring, and does the existence of feedback (or the feedback's characteristics) improve the success of subsequent applications?

Acknowledgements

While the author is responsible for any errors or omissions, I would like to thank Dawn Davidson for the encouragement to develop a better normalization technique, and Guillaume Beaulieu-Houle for suggesting Standard Error.

Appendix A: Charts produced by CoSeT

The collection of scoring data and related information in a database (spreadsheet), provides an opportunity for analysis that can provide insight into the selection competition. CoSeT collects (in numerical form):

- scores for each criterion for each reading assignment for each marker/project
- confidence of the markers about submissions, as collected directly or calculated indirectly

From this data, currently CoSeT creates 14 charts:

Chart #	Chart Description
1	Project scores histogram with distribution of project scores across YES, NO, & uncertain zones
2	Project scores in rank order, with offset traces
3	Chart showing the projects affected by scoring uncertainty
4	Projects that are moved in or out of the uncertainty zone by normalization
5	Change in rank position of all projects by either type of normalization
6	Histogram of the criteria scores
7	Scaled Standard Deviation of criteria vs scores for assignments
8	Histogram of the standard deviation of criteria scores for each assignment
9	XY plot of the marker's Double Normalization factor versus standard deviation of the marker's assignment scores
10	XY plot of each markers Double Normalization fact versus the average of their raw scores
11	XY plot of the marker's confidence versus the corresponding project's score
12	XY plot of markers' confidence and their Double Normalization factor
13	XY plot of the Double Normalization factor versus Single Normalization factor for each project
14	XY plot of the Double Normalization factor versus the resulting project score

In this appendix, examples of these charts are presented, along with some discussion of what the data might infer about the competition. Unless otherwise indicated, the plots are from a single competition with about 100 submissions and untrained markers.

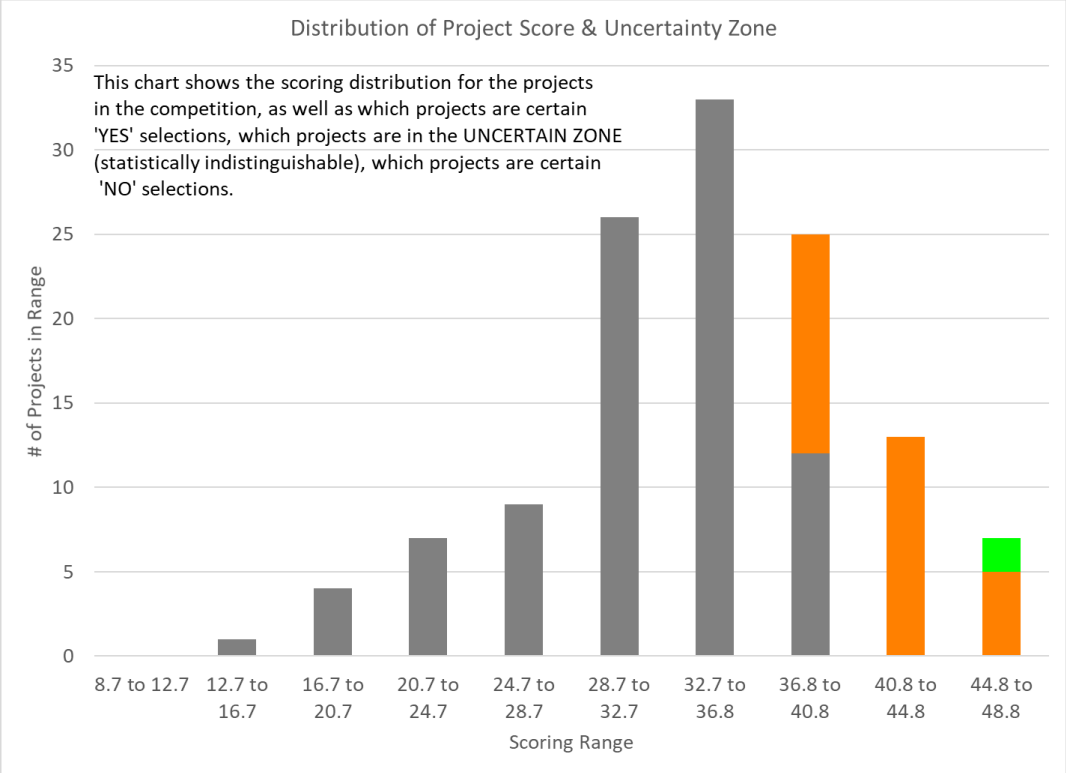

Figure 9: Submission Score Histogram (Chart 1)

This chart shows how the submission scores are distributed across the scoring range.

Discussion: 80% of the project scores for this competition are clustered in the top half of the scoring range. If the language ladder could be designed to better distinguish among the stronger submissions, then the scores could better distinguish among the submissions, and reduce the uncertainty zone.

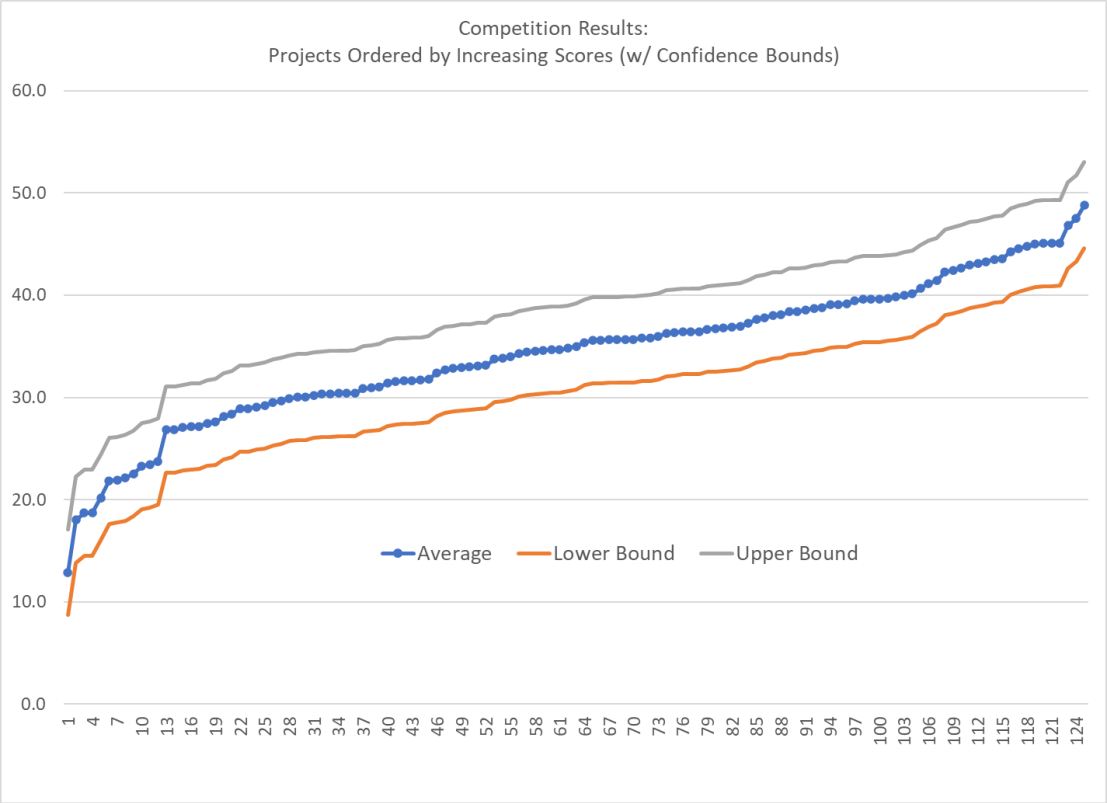

Figure 10: Submission Scores with Confidence Bounds (Chart 2)

This chart presents the 'Average' score of the assigned markers for each submission. The competition scores are rank ordered. The upper and lower bounds for this competition are +/-2 Standard Errors of the average of the scores from markers for each submission.

Discussion:

The slope of the project scores in this figure is steady over much of the competition range, with a slight increase in slope for the upper ~15 projects. This means it slightly easier to distinguish among the top 15 submissions, as compared to the submissions ranked 15 to 105.

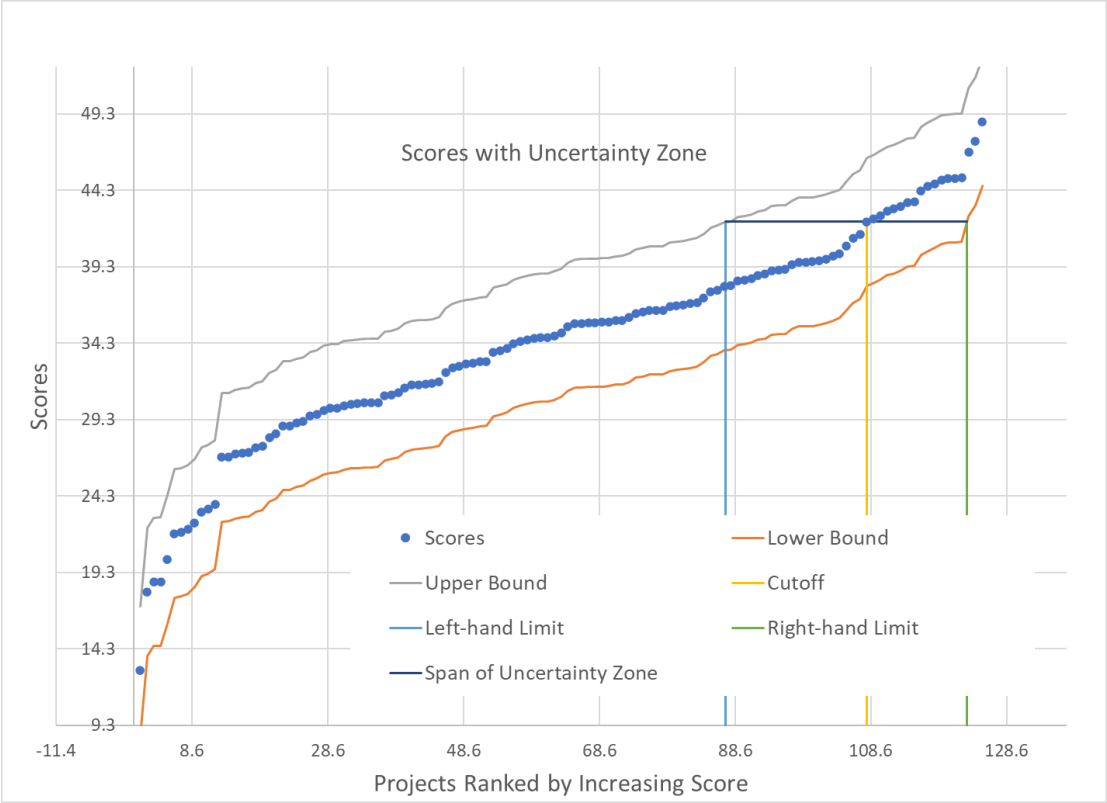

Figure 11: Uncertainty Zone Chart (Chart 3)

This chart shows the uncertainty zone – that is the upper limit and lower limit submissions whose scores are statistically indistinguishable from the nominal cut-off.

Discussion: In this competition 125 submissions were scored, and the cut-off is set to select 18 projects (as indicated by the orange line). The three submissions to the right of the upper limit (green line) are outside the uncertainty zone, and thus are confirmed. Since the uncertainty zone encompasses 42 submissions the competition organizers will need another method to select from this zone the remaining 15 ‘winners’.

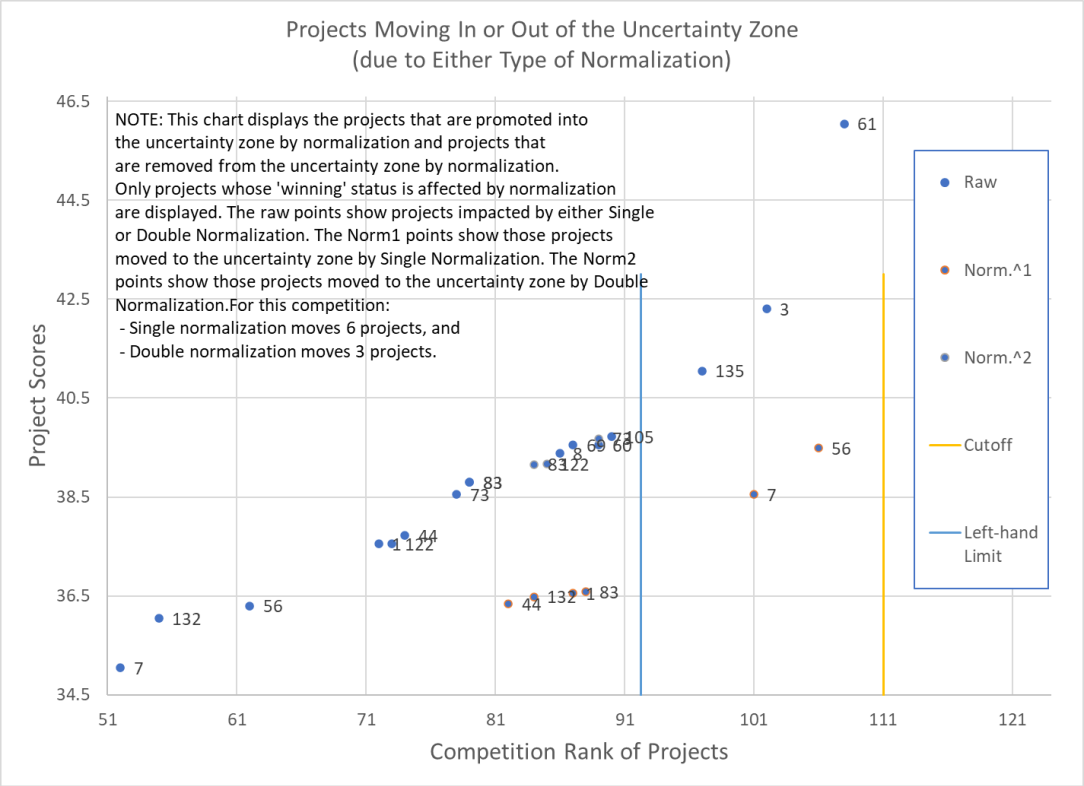

Figure 12: Projects Moved In/Out of Uncertainty Zone by Normalization (Chart 4)

This chart shows the project numbers that move in or out of the uncertainty zone depending on which type of normalization is applied.

Discussion: We can see that single normalization promotes the largest number of projects, while double normalization promotes fewer submissions, most of which are near the lower limit of the uncertainty zone. Note, the difference in vertical position of the raw, single normalized, and double normalized of project scores reflects the impact on scoring due to normalization.

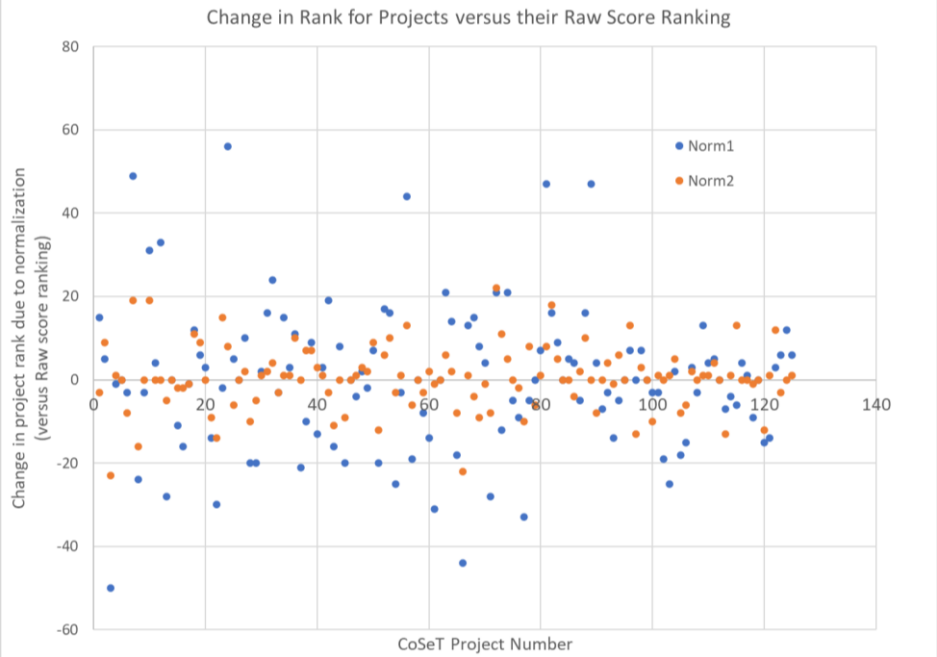

Figure 13: Change in Rank by Either Kind of Normalization (Chart 5)

This figure shows the number of ranking positions each submission moves due to either type of normalization.

Discussion: We can see that double normalization tends to change ranking positions projects less than single normalization. The smaller perturbation of the ranking (from double normalization) can be useful when the experts have influence over the competition processes.

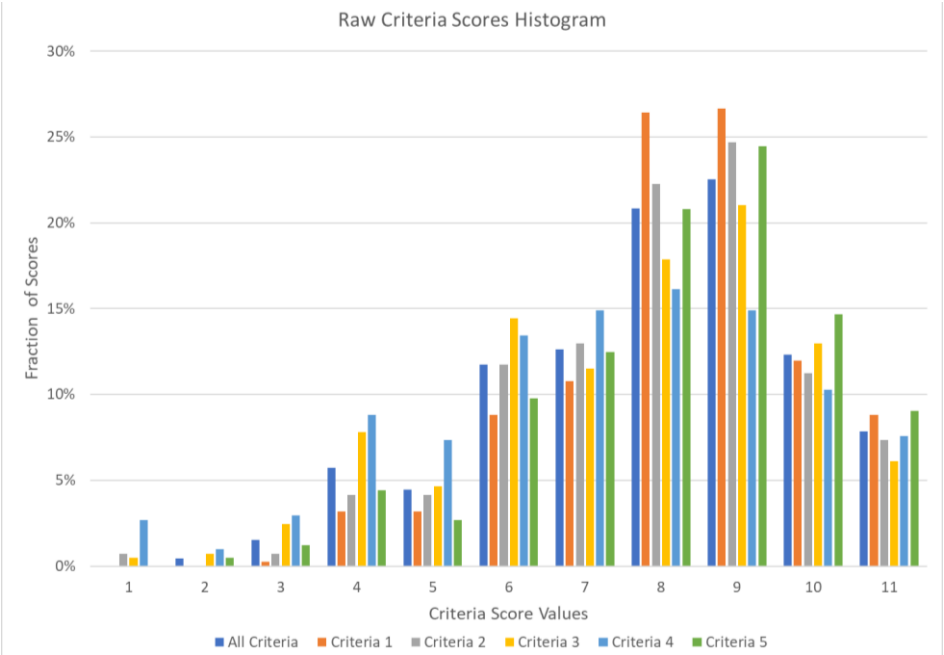

Figure 14: Criterion Level Scoring Histogram (Chart 6)

This chart shows the distribution of criterion scores for each assignment.

Discussion: Like the submission scoring histogram the criterion scores are clustered in upper half of the scoring range. Note also that – for a given criteria score value – there are similar numbers of scores for all the criteria (i.e., the number of projects that scored '5' on Criterion 1 is similar to the number of projects that scored '5' for the other four criteria.

Perhaps the projects in the competition are homogenous in their strengths against the criteria, or perhaps the markers tend to apply the same general scoring level across a submission's scores. This chart may also indicate that the criteria are not orthogonal (i.e., one aspect of a submission may tend to influence the score across more than one criterion). This chart suggests redesign of the scoring criteria, or better training of the markers might help differentiate better between the submissions.

Figure 15: Standard Deviation of Criteria Scores versus Reader's Scores (Chart 7)

This chart further explores whether there is a relationship between the scores a reader gives for each assignment, and their tendency to spread criteria scores across the available range.

Discussion: The left-hand figure is for the competition with untrained markers, while the right-hand figure is from the large random data set. For the competition with untrained markers, it appears that there is a slight tendency for readers to use a broader range of criteria scores for submissions which they perceive as weaker, while readers tended to use a narrower set of criteria scores for submissions they perceive as stronger.

Note the dome shape of the large random data set. The standard deviations of criteria scores (by assignment) are smaller for the lowest and highest submission scores, while broader in the middle. If a real competition data set exhibits this general shape it may suggest that randomness in the scoring (an alternative interpretation is that the markers tend to agree on the best and worst submissions).

In response to the untrained markers' chart, more training for the markers on how to apply the language ladder may be appropriate.

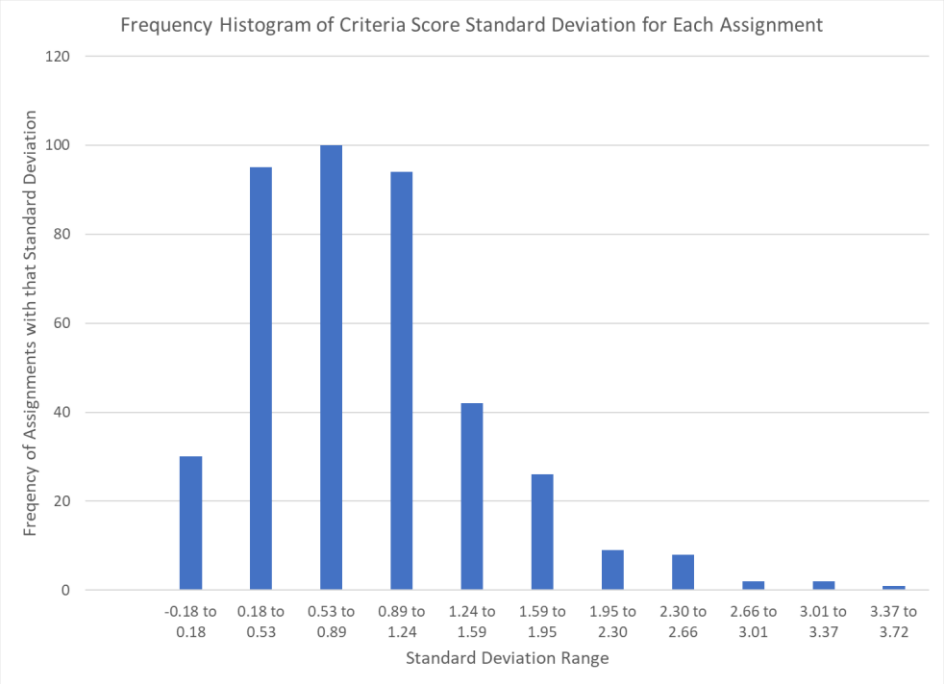

Figure 16: Standard Deviation of Criterion Level Scores (Chart 8)

This chart provides information about the spread of criteria scores which readers use for their marking assignments.

Discussion: In this case we can see that the criteria scores for most assignments are less than 1.25 standard deviations from the average of the marker's average score for that assignment (of the 10-point scoring range). In other words, for each criterion score of an assignment, the markers each tend to give scores +/- 1 of the assignment average score.

Similar to the two previous charts (above) this chart raises questions as to whether the criteria are orthogonal, and whether the markers' training could be improved.

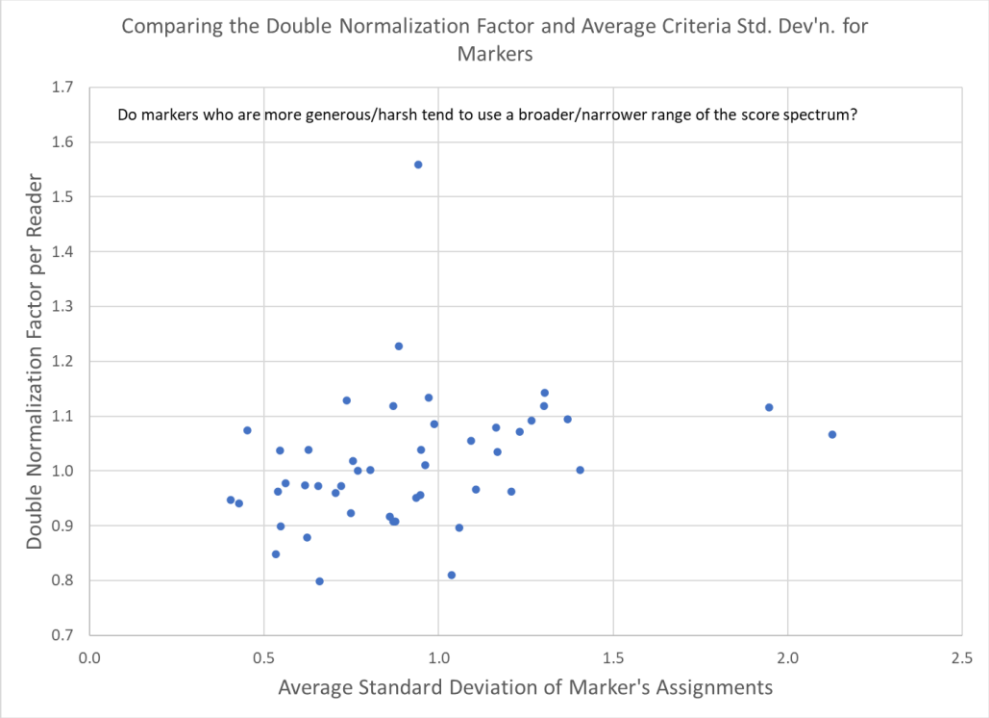

Figure 17: Normalization Factor versus Markers' Scoring Standard Deviation (Chart 9)

This chart shows the double normalization factor against the average submission score standard deviation (for each marker).

Discussion: As the above chart asks, do we see a trend between markers who are more generous (higher normalization factors) and their tendency to score using a wider or narrower spectrum of the scoring range? The above chart does **not** signal a particular relationship.

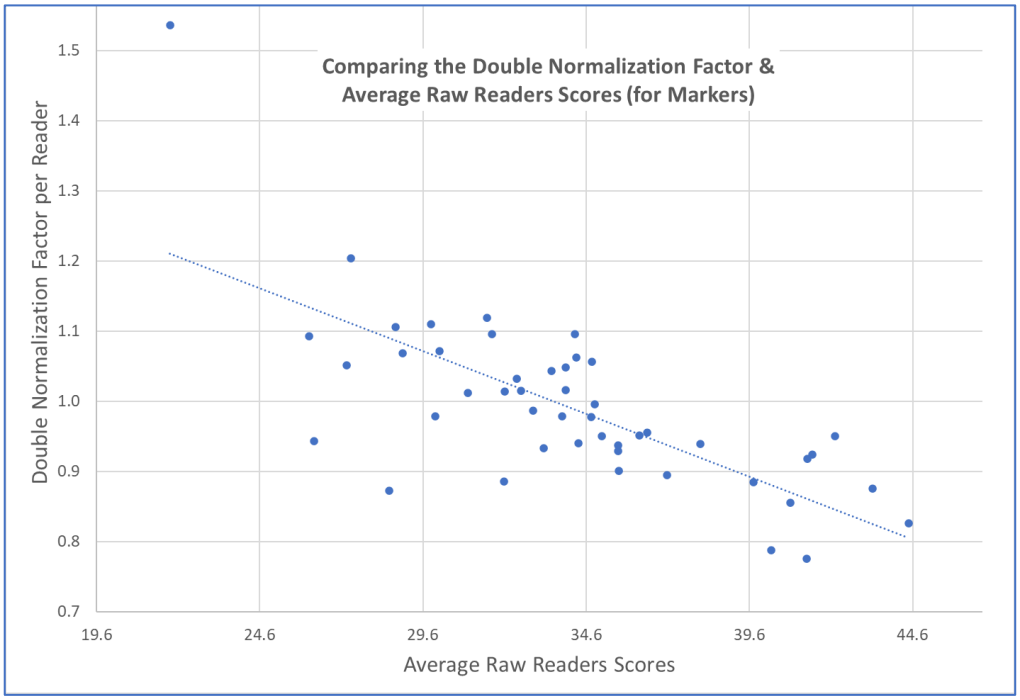

Figure 18: Comparing Double Normalization Factors and Reader Scores (Chart 10)

This chart plots the average double normalization factor against the average submission score (for markers).

Discussion: Normalization is supposed to increase the scores of harsh markers and decrease the scores of generous markers. Consistent with these expectations, (and as we can see in this chart) the markers with a low scoring average (raw) tend to have higher normalization factors while markers with higher scoring averages (raw) tend to be given lower normalization factors.

The chart has plenty of outliers – markers with low scoring averages who are given low normalization factors. This can occur if a harsh marker is assigned to projects where there is a tendency for other 'harsh' markers to be assigned. If the markers each received a similar set of submissions to score, then the variability in this chart would appear to be due to variability in how markers score the different submissions they were assigned.

Figure 19: Project Scores and Marker Confidence (Chart 11)

The above chart is intended to explore whether markers tend to give higher or lower scores to submissions based on their perceived level of expertise/confidence in marking the submission.

Discussion: Anecdotally, the author has heard experts suggest they are more generous to submissions outside their areas of expertise.

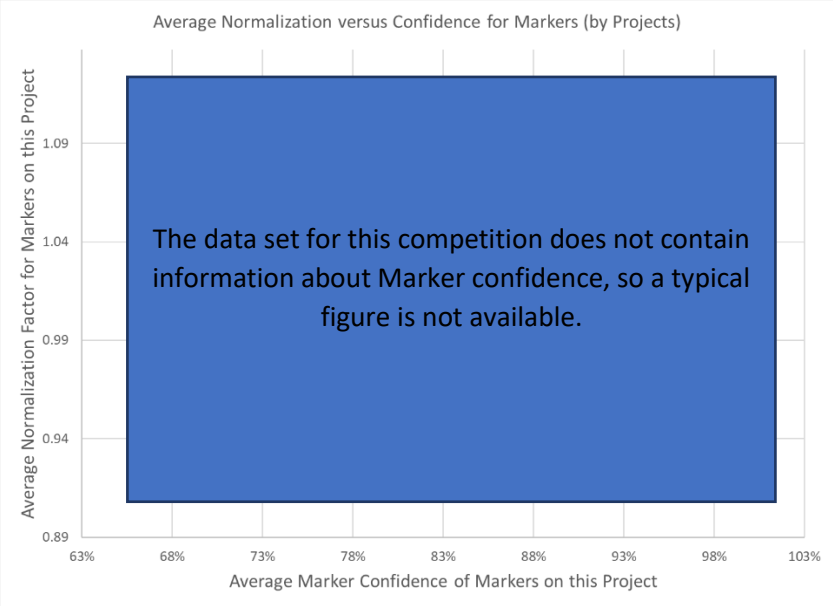

Figure 20: Average Normalization Versus Confidence for Markers (by Projects (Chart 12)

This chart plots the average confidence of readers on a project against the average normalization factor for the readers.

Discussion: If markers scores tend to vary with their level of confidence for the project they are scoring, then there could be a relationship between the average normalization factor for markers and the average confidence for markers (on a project-by-project basis).

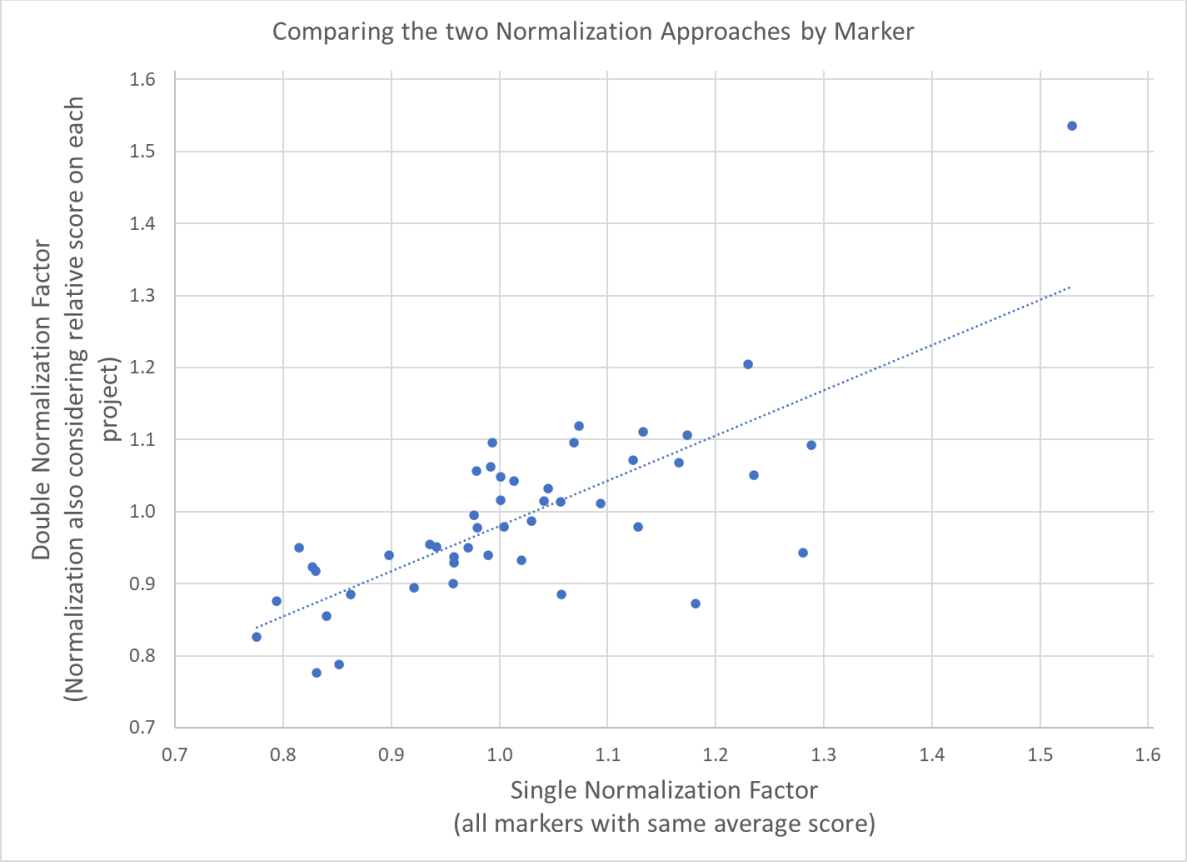

Figure 21: Double Normalization versus Single Normalization factors for Markers (Chart 13)

This chart compares the two normalization factors for markers.

Discussion: This chart should help confirm whether the single normalization factors tend to agree with the double normalization factors, a trend we see in the above chart. Note also that the double normalization factors tend to be smaller than for single normalization, although the correlation is weak. This implies that the double normalization may tend to make smaller (more surgical) adjustments to the scoring.

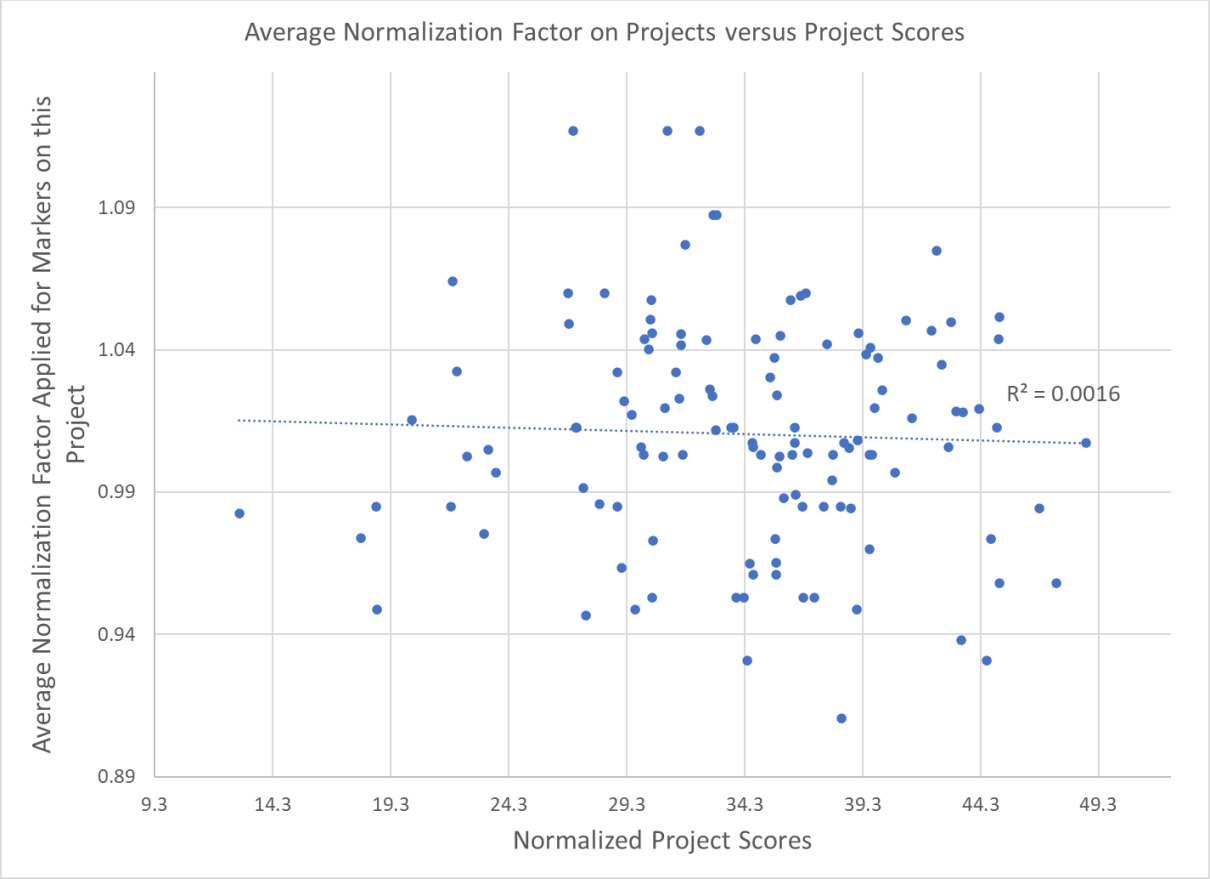

Figure 22: Normalization Versus Scores for Projects (Chart 14)

This chart plots the average normalization factor against the average of the readers score for a project.

Discussion: We do not expect the level of normalization to be correlated with the ranked position of the projects. This chart allows the competition organizers to assess whether there is such a relationship.

In the chart above there does not appear to be a correlation between the normalization applied to projects and the ranking score they achieved, increasing confidence in the use of double normalization for this competition.