Analysis of Results from The 2015 Artisans Cup

Dan Yamins

The 2015 Artisans cup was a landmark event in American bonsai — it brought together more than 70 of the finest trees in the US from artists/owners all over the country, selected from a pool of more than 300 submissions. The competition portion of the event sought to determine the most exciting and effective trees from within this already highly-selective grouping. Five judges, each eminent practitioners of the bonsai art, were invited to score each tree based based on a variety of considerations (see reference [1] at the end). After the competition was over and the winning trees announced, scores for each from each judge were released by the Artisans cup organizers (see reference [2]).

To help better understand these results, I performed some simple data analyses of the score data. In doing these analyses, I uncovered two findings that I believed would be useful to share with competition organizers and the larger bonsai community. Specifically, I looked at:

- A question of how scores were *rounded* to produce final results. The result of this investigation is that I believe an additional 3rd-place prize would be deserved.[1]
- The issue of how scores could be better *normalized* to compare between judges more fairly. The result of this investigation is a recommendation for how such normalization could be implemented in future competitions.

Below, I will go into both of these issues in greater detail.

Rounding Policy

As I understand the way the scoring system was described, for each tree, the highest and lowest scores were dropped and the remaining three scores were averaged to produce a final result. This procedure is called *trimmed averaging*, where the trimming percentage is 20% — one out of five datapoints each on the high and low sides. Perhaps the most prominent usage of trimmed averaging is in Olympic figure skating scores, where the top and bottom scores for each participant are dropped. The reason for using trimmed averaging in the Olympics is to prevent the undue influence of judges who might be specially aligned with one of the participants because of prior relationships, e.g. the judge from one country might be biased towards the participants from his or her country. I believe that the choice of trimmed averaging is well-suited to a bonsai competition, because a similar relationship exists in the bonsai world: that of tree owner and tree designer. After all, competition organizers would be hard-pressed to find a panel of suitably

[1] I was very pleased to see that Artisans cup organizers recently announced that they would indeed award this tree, Konnor Jensen's Japanese White Pine, an additional 3rd-place prize.

high-powered judges none of whom had ever performed work on any of the trees in a contest such as the Artisans Cup.

Just as a basic sanity check, I loaded the data using the `Python` programming language and used simple analysis tools to the compute this trimmed average score for each tree. I then compared the results to the listed official average scores. The results were almost completely consistent between my computations and the official scores (see Fig. 1). That's great — the numbers basically check out. However, there are couple of differences, as you can see by the fact that the points in Fig. 1 aren't *completely* on the diagonal. Specifically, the key difference is that the official scores have been *rounded* to the nearest whole number. For example, if you consider Tree 1, the official score is 41 — but the actual numerical average of the three scores (excluding the highest and lowest score) is actually closer to 40.66667. In this case, the score has been rounded up (from less-than-41 to 41). That is, the rounding slightly increased the score of this tree. For other trees, the rounding slightly decreased the score — for example, if you look at tree 32, the official score is 51, but the actual average of the judges' scores is 51.33334.

The discrepancy between the rounded official scores and the actual numerical average is small. However, the difference is enough to have an effect on the winners. In the official scores, the first and second place trees were a tie (Trees 32 and 52 both had 51 points), and the third- and fourth-place trees were a tie (trees 27 and 4 both had 50 points). However, in the actual numerical averages, these trees were *not* tied. Tree 32 had 51.33334 points, while Tree 52 had 51.0 points. Tree 27 had 50.33334 points while Tree 4 had 49.66667 points. If the actual numerical averages were used, the winners would have been:

1. First Place: Tree 32 (51.33334 pts), Randy Knight's RMJ — no need for a tie-breaker.

2. Second Place: Tree 52 (51.0 pts), Tim Priest's sierra juniper.

3. Third Place: Tree 27 (50.33334 pts), Konnor Jensen's JWP — *not* Tree 4 (Amy Blanton's RMJ).

Thus, one advantage of using the actual average is it makes ties less likely — Randy's tree actually won outright. Ultimately, I'm not sure whether the rounded scores or the actual average is the "right thing" to use from an artistic standpoint, as I think you can make a case for doing it either way. Using the rounded scores increases the likelihood of ties very significantly. If two trees are so close that they differ in score by less than a point, perhaps it would be more artistically coherent to allow a panel of tie-breakers to award the decision. On the other hand, having more tie-breakers increases the possibility for bias to be injected when the ties are broken. Regardless, it's probably important to be sure about the implications of this issue ahead of time, because it can lead to different results.

Normalizing the Judges

Perhaps a more important issue is that of comparing scores between judges. There was definitely a lot of healthy divergence between the opinions of the judges. For example, Colin and David's scores were so different that they were actually *anticorrelated* with each other. (See Fig. 2, left panel.) It's interesting to look at the pattern of which judges were correlated with which other ones (Fig. 2, right panel.)

In and of itself, judge divergence is fine — it's the sign of the fact that, among trees all of which are high quality, there are bound to be artistic differences in judgement. However, more problematically, different judges' scores ended up having very different influence on the final score. The way we can see this is by, for each judge, looking at the correlation between that judge's individual scores and the final average score. (See Fig. 3, left panel.) You can see that Walter had the most influence (influence factor = 0.77), while Boon had the least influence (influence factor = 0.37), with other judges somewhere in between.

Why did this happen? The reason is because of inconsistencies in the way the judges did their scoring. Boon's scores were on average lower than the other judges — he didn't use the range of possible scores as fully as the other judges (see Fig. 4). As a result, Boon's scores were dropped much more often than the other judges (Fig. 3, right panel). This problem was not confined to Boon — all the judges used somewhat different ranges — it just happened to affect his scores the most.

This is pretty problematic because it means that what should really be an irrelevant factor — the *absolute* score values — ends up having a strong influence on the final ranks. For example, a Boon score of 40 is more like another judge's score of 50, making it really hard to compare the scores in a fair way. To make this clearer by example: consider tree 32 (Randy's winning tree). Boon's score for this tree was 42 — and as the lowest score it was dropped. However: for Boon, 42 was one of his highest scores. In fact, Randy's tree was one of Boon's top 5 picks. Two of the other judges actually ranked it lower than Boon did, relative to their other scores, but it was Boon's score that was dropped for being low. As another example, Boon's highest-ranked tree was tree 56 (Eric Schikowski's hemlock), but because his effective range was different, his score was the second *lowest* for this tree compared to the those of the other judges. This pattern was repeated throughout the competition, which is an undesirable situation.

There is a standard solution to this sort of problem in statistical data analysis, calle "z-scoring". In technical terms, z-scoring involves subtracting off the mean, and then dividing by the standard deviation — doing so separately for each judge's list of scores. In effect, z-scoring normalizes the data so that the ranges of each of the different judges are the same. I performed this procedure, and found that it corrected the problem describe above somewhat. After normalization, the judge's influences are now a bit more equal to each other (see Fig. 6, left panel). More importantly, no one judge's scores are being dropped much more than any others' (see Fig. 6, right panel). That is, there is now more equity in how much the final total reflected the opinions of each judge. While there is still inequity in how much each judge's score correlated with the final result, the reasons for this are now "more fair" — namely, some judge's opinions (apparently Walter's, for instance) are just more predictive of the true average of than others'. In making these comparisons, I still used the trimmed average procedure, dropping the highest and lowest (normalized) score for each tree.

In my opinion, this procedure would likely have been a more consistent approach than what was applied in the official scoring. Of course, statistical consistency in and of itself hardly matters — except that it would have ended up having a significant impact on what would have been chosen as the best trees. With the normalized scores, the top three trees turn out to be:

1. First Place: Tree 32 (86.6 pts), Randy Knight's RMJ.

2. Second Place: Tree 56 (82.5 pts), Eric Schikowski's hemlock.

3. Third Place: Tree 8 (82.1 pts), Greg Brenden's southern white pine.

The first place tree doesn't change: it's scores were so robustly high that it wins regardless of what procedure is used (in fact, the separation between the first and second-place tree is even *more* evident in the normalized scores). However both the second and third place trees are different here than in the original procedure, because a broader range of judges' opinions are being taken into account.

Interestingly, with these properly normalized scores, Eric Schikowski's hemlock — which was the only one of the trees to receive more than one first-place vote (from Boon and Peter) — would receive a prize, whereas with the un-normalized rankings, it was not in the top 5. I do not feel comfortable saying that this result would have been better artistically than the official result, but I think would have been a bit more reflective of the judge's actual opinions. However, I do not believe that Artisans Cup organizers should retroactively make these normalized scores the official scores — after all, the original procedure by which the scoring was done was spelled out before the competition and was adhered to in a fair and consistent manner throughout. Transparency and consistency are the most important aspects of fairness in judging, and so changing the policy afterward would not be recommended, even to account for unintended undesirable effects (such as different judges' natural ranges). I merely mean to illustrate my reasoning behind making a clear recommendation for future competitions, namely: *use normalization whenever judge's scores are to be compared.*

Conclusion

I'd just like to end this report by saying that I think that, overall, the organizers of the Artisans Cup did a fantastic job in ensuring that the all participants received a fair judging. It is to the tremendous credit of the organizers that they released all the judge's score data immediately after the winners were announced. Their commitment to this high level of transparency is extremely heartening. Moreover, finding a 100% perfect judging system, one that will satisfy all possible artistic needs, is probably impossible. Although I have remarked on a few ways the process might be improved, I want to be clear that I feel that the process as it was implemented was already quite good. It was really inspiring to see the huge variety and stupendous quality of the entries — and the trees that won the top prizes, regardless of the particulars of the judging system, were amazing and awesome. I'm looking forward to how the energy gathered by the Artisans cup will energize American bonsai in the years to come.

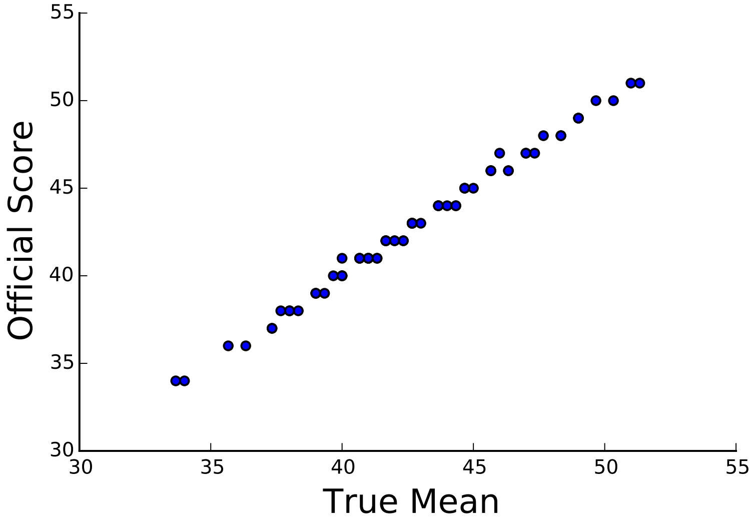


Figure 1: True trimmed average vs official score. The x-axis is the true arithmetic average of the middle three scores for each tree. The y-axis is the score as reported on the official tally. The difference between the two calculations is due to the fact that in the official scoring, the averages were rounded to the nearest whole number before the ranking was done.

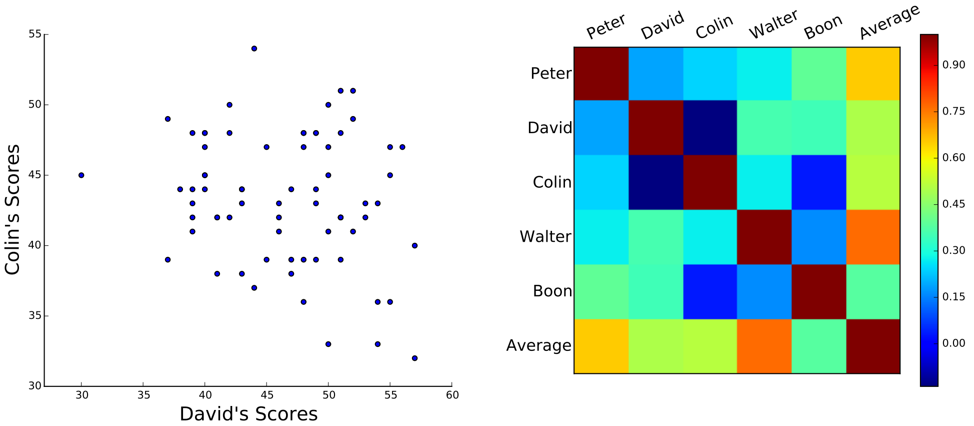


Figure 2: Correlation between judges' scores. Left: David's scores vs Colin's scores show a *negative correlation*! **Right:** heat map of correlations between judge's scores for all pairs of judges (and the score average).

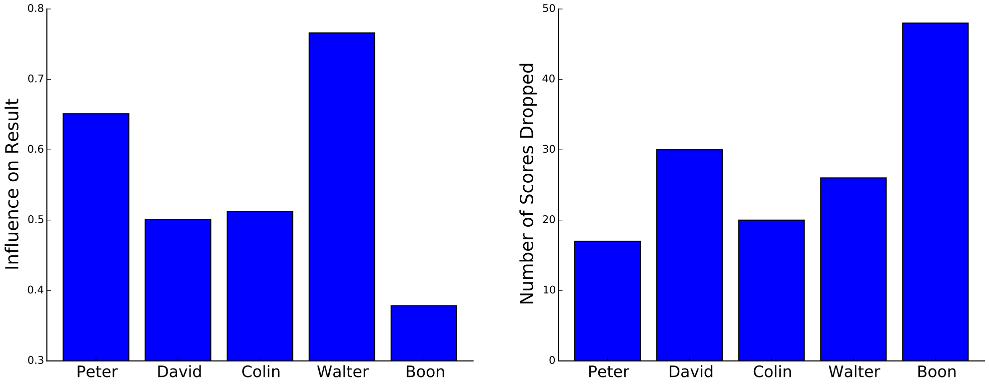


Figure 3: Amount of influence for each judge. Left: The y-axis represents the correlation of the indicated judge's scores with the final overall ranking. This shows that the different judges' influences were far from equally distributed. **Right:** Underlying this inequality is, in part, the fact that one judge (Boon) had especially many of his scores dropped, because he gave on average lower scores than all the other judges.

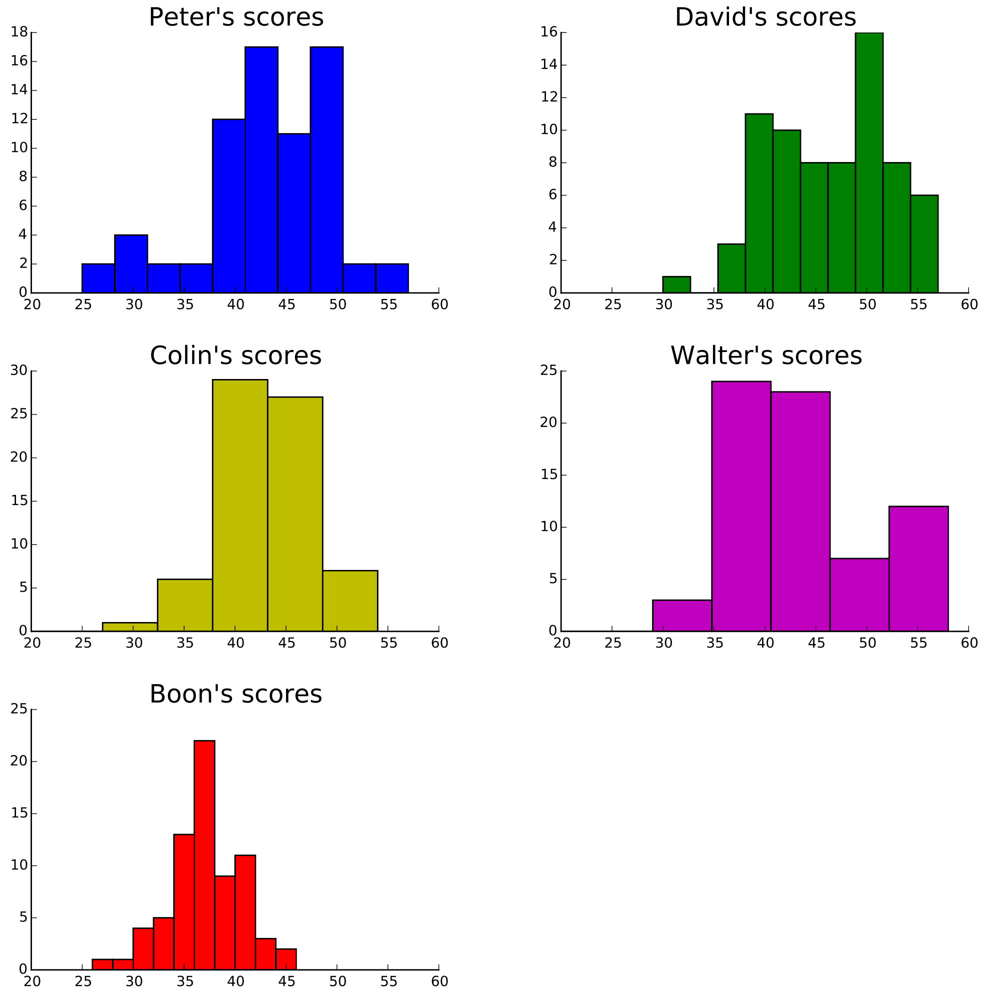
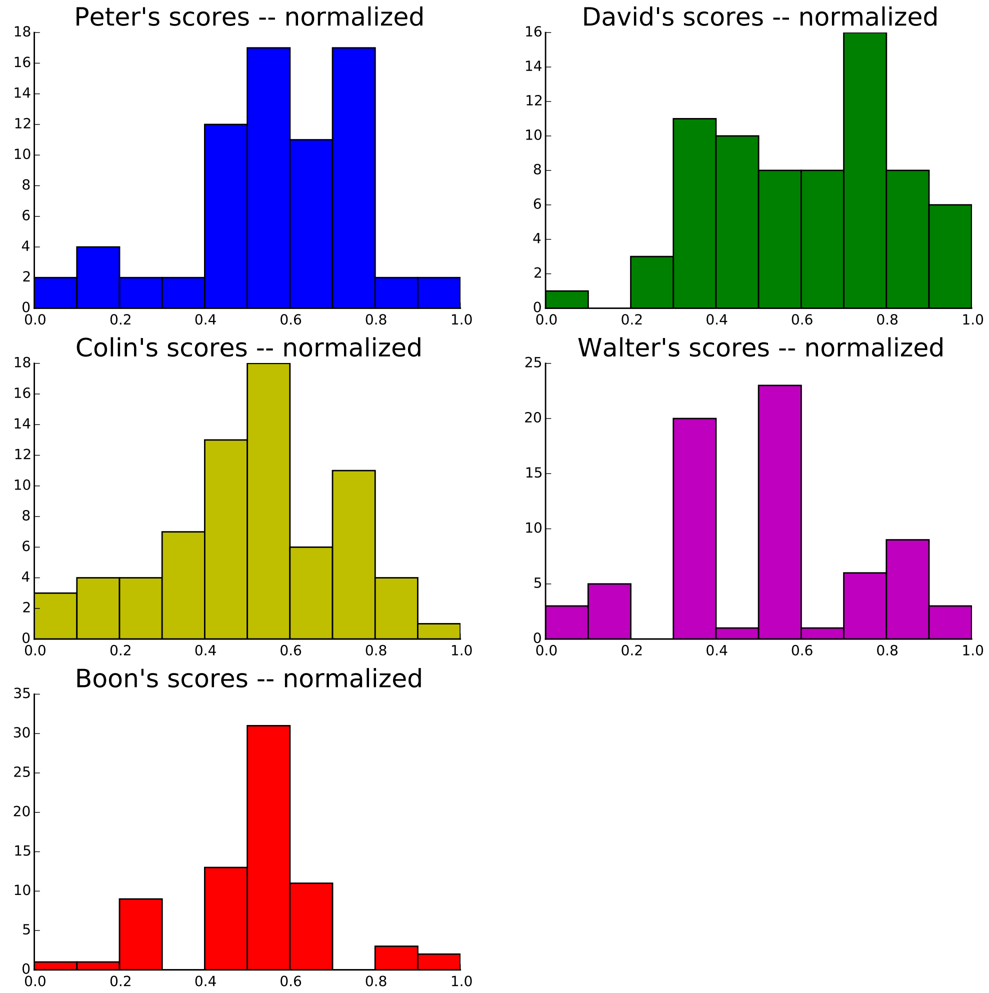


Figure 4: Distribution of judges' scores. Each panel shows the distributions of the judges scores. Notice that Boon's scores are concentrated into a narrow and lower range than the other judges' scores.


Peter's scores -- normalized

David's scores -- normalized

Colin's scores -- normalized

Walter's scores -- normalized

Boon's scores -- normalized

Figure 5: Normalized score distributions. Now that the scores have been normalized, all the distributions are on the same range. It's interesting to notice how symmetric and normal the distribution of Boon's scores are.

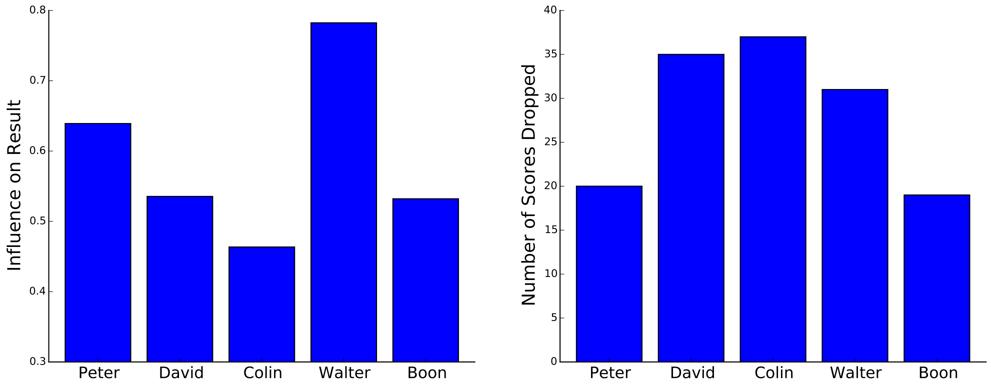


Figure 6: Effect of normalization. Left: After normalizing the scores to a standard range, the inequality in the influence of the judges has been somewhat corrected. These scores were computed using the trimmed average (dropping the top and bottom scores), just as with the original scores. **Right:** In the normalized case there is much less inequity of the number of scores dropped for each judge as compared to the original procedure (see Fig. 3, right panel for comparison).

References

[1] 2015 Artisan's Cup Scoring Rubric (2015). URL `http://static1.squarespace.com/static/54c9359de4b0f2976a2eaf2e/t/560810c8e4b03eb8da831901/1443369160697/judgingrubric.pdf`.

[2] 2015 Artisan's Cup Final Scores (2015). URL `http://www.theartisanscup.com/blog/2015/9/27/the-results-are-in`.