National Journal of Speech & Debate

Volume VI: Issue 2

January 2018

Finally... a debate camp in South Texas!

#campatTFFI

TESTING A NEW RANKING METHOD, THE LOGIT SCORE, WITH SIMULATED TOURNAMENTS

BY T. RUSSELL HANES*

* Mathematics chairperson, Northwest Academy, Portland, Oregon; Master of Arts in teaching secondary math; Lewis & Clark College, Portland Oregon; Bachelor of Arts in mathematics, Columbia University in the City of New York. The author can be reached at russell.hanes@gmail.com.

INTRODUCTION

Win-loss records are commonly used to rank debate teams. However, a win-loss record depends upon which opponents a team was matched against. Power matching is supposed to make opponent matching consistent, but my previous empirical research has shown it does not always do this as effectively as imagined (not because of errors, but because tournaments do not have much information at hand to accurately match teams).[1] Furthermore, win-loss records are affected by inconsistent judging. Because of differing opponent strength and inconsistent judging, win-loss records can be an inaccurate way to rank teams.

Speaker points provide a possible alternative way to rank debate teams. However, judges can be inconsistent in scoring speaker points, so there is skepticism in the debate community about the accuracy of speaker points-based rankings.

The ideal method to rank teams would be based on only the performance at one tournament; it would be less sensitive to variability in schedule strength if the matching is not perfect; it would be relatively simple to program into software; it would incorporate both win-loss record and speaker points into a single score; and finally, the ideal method would yield rankings that are accurate. This method is the logit score.

Previous research on the logit score, published in this journal,[2] confirmed that when the method is applied to a real data set—an entire college debate

[1] Of course, stronger teams are matched against stronger opponents. However, the data have shown comparable teams can sometimes be matched against different strength opponents. For example, Team A might have weaker opponents than team B, even though A and B are equally as strong: http://art-of-logic.blogspot.com/2015/07/study-of-speaker-points-and-power.html and http://art-of-logic.blogspot.com/2009/03/what-are-normal-opponents-wins-in-given.html.

[2] Hanes, T. Russell. (2017). "Introducing the Logit Score." *National Journal of Speech & Debate*, Volume 5, Issue 2.

season—the logit score produces rankings that are reasonable. Despite not controlling for potential lurking variables such as region, team style, and judge bias, the logit scores are highly "predictive" of actual results: the team with the higher logit score often beat an opponent with a lower score. The previous article concluded that, "The logit score ranks are both more aggressive, meaning the method sees fewer ties, and yet are slightly more accurate," which means the logit score passed a critical first test of fitting historical data.

METHOD

This article extends the research described in last year's *National Journal of Speech & Debate*. Besides historical data, another way to test a ranking method is with computer simulations. A set of teams with known, true ranks can be put through dozens of virtual tournaments. These tournaments can be simulated using reasonable estimates of speaker point variability and frequency of low-point wins. At the end of each virtual tournament, the accuracy of each ranking method can be found by comparing the observed ranks a method generates to the actual, true ranks.[3] Testing the logit score through more than a hundred simulated tournaments is a critical second test for the method, which will show whether its fit to historical data is mere chance. Simulations are experimental proof the method works under conditions of team and judge variability to rank teams accurately.

It is possible that there could be an interaction effect between the method of pairing a tournament and the accuracy of a ranking method. For example, it might be the case that the logit score is more accurate only for a randomly paired tournament but win-loss records are more accurate for power-matched tournaments. For this reason, it is necessary to compare ranking methods across various tournament conditions.

Three tournament conditions were tested: (1) a tournament using six random rounds, (2) a "pre-matched" tournament,[4] and (3) a tournament using one random preset round followed by five power-matched rounds.[5] Pre-matching and power matching are both ways to decrease the variability of opponent schedule.

[3] Code and data from virtual tournaments are available upon request.

[4] In the pre-matched tournament, teams were divided into five groups based on speaker points earned in the first, random preset round. Every team was then matched against one opponent from each of the five groups for the remaining rounds. Although not technically pre-matched, this method replicates the essential logic of a pre-matched tournament: every team debates a cross-section of the entire pool, from the weakest to the strongest teams.

[5] There are a wide variety of methods of power matching within brackets: high-low by speaker points, high-high by speaker points, by opponent wins, or by S.O.P., to name only a few. Previous research has shown that high-low by speaker points is more effective than high-high or random

For each virtual tournament, rankings were generated in three unique ways: (1) by the traditional win-loss record, (2) by the logit score, and (3) by median speaker points. Each of these was compared to the actual, true ranking to rate the method for accuracy.[6] Median speaker points were included as a baseline; in all prior research, it has outperformed every other ranking method.

To increase the accuracy of the simulations, real information from an entire college debate season was used to model speaker point variability. The speaker points on the college circuit are remarkably consistent. However, one might be able to adjust speaker points for judge inconsistency in other contexts, such as high school debate, to achieve similar consistency. See Appendix A for an effective way to adjust speaker points.

The college season's population parameters ($\mu = 56.86, \sigma = 0.54$) for the distribution of all teams' average speaker points were replicated in the experiment. An experiment size of 64 was selected as representative of a mid-to-large tournament, and also so power-matching brackets would work out perfectly. The sample of 64 thus looked like a typical spread of team ability at a tournament.

Furthermore, the college season's standard deviation in speaker points achieved in any given round for a single team is 0.67. In the virtual tournament, a team's performance in each round was calculated by adding its true speaker point average to a random Normal value consistent with the standard deviation of 0.67. Each team's virtual performance was random in each round but consistent in center and spread with a real team.[7]

Of course, the higher scoring team in a round does not always win, so the virtual tournaments must also account for low-point wins. Low-point wins occurred about 6.5% of the time in the real debate season. However, most of these happen when the two opponents were only 0.5 speaker points apart or less. Above that mark, low-point wins were rare. The formula that best fit the historical data is given in Appendix B. Based on the randomly-generated speaker points each opponent earned, a winner is assigned in each virtual round using this probability formula. This generated an appropriate distribution of low-point wins in the virtual tournaments.

within brackets, http://art-of-logic.blogspot.com/2015/01/100th-post.html, so this is the method used.

[6] Additional methods considered but which failed initial tests included weighted wins and a z-hybrid score (z-score for win-loss record plus z-score for speaker points). These produced rankings less accurate than the logit score so were dropped from further evaluation.

[7] This model does not separate out a team's variability from judging variability. Using methods to adjust judges' speaker points might improve the accuracy of the speaker points. See Appendix A for more details.

RESULTS

Three statistics were used to assess the outputs of each ranking method compared to true ranks: (1) Spearman's rho for rank order correlation; (2) mean absolute deviation; and (3) the weighted footrule.[8] See Appendix C for notes on slight modifications I made to the weighted footrule.

Table 1
Means of Spearman's rho

Tournament condition	Ranking method		
	Traditional [9]	Logit score [10]	Points-only [11]
Random	0.737 (0.055)	0.847 (0.029)	0.851 (0.028)
Pre-matched	0.751 (0.055)	0.848 (0.031)	0.854 (0.032)
Power-matched	0.823 (0.033)	0.847 (0.030)	0.858 (0.031)

Notes. n = 50 for each condition. Standard deviations in ( ).

Each statistic might pick up on different inaccuracies, so all three were included. Spearman's rho is a test of monotonicity. It is similar to the familiar r^2 statistic but more suited to rankings. Teams ranked several positions off are heavily punished by rho. The weighted footrule, however, punishes inaccuracies at the top of the rankings more heavily than inaccuracies in the middle or bottom. Finally, mean absolute deviation is neutral on the size and "location" of the errors; all errors are equally weighted.

Rho varies from 0 (no correlation) to 1 (perfect correlation). Rho demonstrates a solid pattern: changing from random to pre-matched increases the accuracy of traditional rankings but does not affect the logit score in any meaningful way; changing from random to power-matched increases the accuracy of traditional rankings even more, but not enough to close the gap with logit score rankings. Traditional rankings, in a power-matched condition, are about 82% accurate. Logit score rankings are, in any condition, about 85% accurate. Speaker points-only rankings set a baseline of 85 to 86% accuracy.

Mean absolute deviation (MAD) measures the average difference between a team's true rank and its observed rank. Higher number indicate worse correlation.

[8] Langville, Amy N. and Carl D. Meyer. (2012). *Who's #1?: The Science of Rating and Ranking*. Princeton, NJ: Princeton University Press.

[9] Ranked by wins, total speaker points excepting highest and lowest results, then median points.

[10] See Appendix D for full notes on calculating the logit score.

[11] Ranked by median speaker points.

As shown in Table 2, the same pattern as in rho appears in the MAD scores of the virtual tournaments: traditional rankings are most accurate in the power-matched condition, but not as accurate as logit score rankings in any condition. On average, traditional rankings in power-matched conditions place each team about $8\frac{1}{3}$ ranks away from its true rank, whereas logit score rankings have an average of about $7\frac{2}{3}$ ranks in error. The speaker points-only method again sets the baseline, at about 7.45 ranks in error.

Table 2
Means of mean average deviations

Tournament	Ranking method		
condition	Traditional	Logit score	Points-only
Random	10.15	7.68	7.55
	(1.21)	(0.78)	(0.75)
Pre-matched	9.86	7.64	7.46
	(1.17)	(0.77)	(0.81)
Power-matched	8.37	7.68	7.35
	(0.84)	(0.86)	(0.79)

Notes. n = 50 for each condition. Standard deviations in ( ).

Weighted footrule (WFR) scores work just as MAD scores do: measuring an average difference between true rank and observed rank. The only difference is the WFR average is weighted to punish differences in rank at the top of the tournament more heavily than differences in rank at the bottom.

As shown in Table 3, the same pattern appears in the WFR scores as appears in rho and MAD: the power-matched condition improves the accuracy of the traditional ranking method; logit score and speaker points-only rankings are more accurate in every condition and unimproved by different tournament pairing conditions in any substantial way.

Table 3
Means of weighted footrule

Tournament	Ranking method		
condition	Traditional	Logit score	Points-only
Random	9.68	7.18	7.08
	(1.24)	(0.73)	(0.71)
Pre-matched	9.38	7.11	6.94
	(1.16)	(0.71)	(0.79)
Power-matched	7.90	7.21	6.88
	(0.92)	(0.85)	(0.77)

Notes. n = 50 for each condition. Standard deviations in ( ).

The speaker points-only rankings form the best baseline for comparison between various conditions; these rankings are unaffected by opponents and therefore unaffected by pairing methods. Thus, the apparent increase in accuracy for speaker points-only rankings from random to pre-matched to power-matched is solely the result of random variability. Speaker points-only rankings represent a lowest floor of ranking error; the vagaries of team performance make it impossible to go lower.

Table 4 shows the accuracy of traditional and logit score rankings compared to speaker points-only rankings across each condition. Because traditional and logit score rankings are less accurate than speaker points-only rankings, the percent changes in accuracy are always negative.

The improvement factors show a clear story. For random pairings, logit score rankings are between 18 to 32 ***times*** more accurate than traditional rankings. Another way to understand this is that logit score rankings eliminate most of the error in ranking that is possible to eliminate.

For pre-matched pairings, traditional rankings seem to improve slightly compared to traditional rankings for random pairings (though the differences are not statistically significant). Conversely, logit score rankings seem to worsen slightly compared to logit score rankings for random pairings (though again, not statistically significantly so). In the pre-matched condition, logit score rankings are about 11 to 19 ***times*** more accurate than traditional rankings.

Power-matching, however, statistically significantly improves the accuracy of traditional rankings compared to traditional rankings for random pairings. Conversely, logit score rankings are statistically significantly worse in accuracy than logit score rankings for random pairings. Despite this convergence of traditional rankings improving and logit score rankings worsening, logit score rankings are still 1.5 to 3.3 ***times*** more accurate than traditional rankings for the power-matched condition.

Table 4
Comparisons between methods

Tournament condition	Percent change compared to Points-only Rankings		Improvement factor
Measure	Traditional	Logit score	
Random			
Rho	-13.34	-0.42	31.76
	(6.30)	(2.48)	
MAD	-35.04	-1.93	18.16
	(16.25)	(7.76)	
WFR	-37.42	-1.69	22.14
	(17.47)	(8.27)	
Pre-matched			
Rho	-12.08	-0.63	19.17
	(6.29)	(3.06)	
MAD	-33.42	-3.12	10.71
	(19.42)	(10.75)	
WFR	-36.59	-3.19	11.47
	(21.34)	(11.46)	
Power-matched			
Rho	-4.01 ***	-1.22 *	3.29
	(3.81)	(2.94)	
MAD	-14.62 ***	-5.01 *	2.92
	(13.34)	(11.04)	
WFR	-7.83 ***	-5.24 **	1.49
	(14.36)	(10.84)	

Notes. n = 50 for each condition. Standard deviations in ( ). * p < .10 compared to random pairings. ** p < .05 compared to random pairings. *** p < .01 compared to random pairings.

SUMMARY

As predicted, there is a notable interaction effect between tournament pairing condition and ranking method accuracy. The pre-matched condition did not substantially change the accuracy of traditional rankings or logit score rankings compared to random pairings. However, the power-matched condition significantly improved the accuracy of the traditional ranking method while somewhat worsening the accuracy of the logit score method. This may seem paradoxical: the ranking methods are affected oppositely by the power-matched condition. The explanation of this paradox is simple. Power-matching pushes team records to more closely reflect their true ability, though still imperfectly. Power-matching also creates more close rounds, which can allow chance wins and losses affect logit scores. This lets some error creep in to logit score rankings.

However, this interaction effect does not change the overall result: in all tournament conditions, logit score rankings were dramatically superior to traditional rankings and nearly hit the theoretical minimum of error possible. Simply put, even despite the enormous variability of speaker points,[12] logit scores are a more accurate signal of true ability than record, no matter how the tournament is paired.

DISCUSSION

The results of this experiment suggest that the debate community would better invest its time in finding ways to help judges give more consistent speaker points (such as rubrics and training) and also subtract out the variability that does crop up (devising good ways to adjust speaker points) rather than worrying about how to pair preliminary rounds.

If the community decides that power-matching preliminary rounds is good practice for the type of tournament environment it creates, then that is one consideration, but it is not necessary for accuracy in rankings if logit scores are used instead of the traditional ranking method.

If logit scores are used for rankings, it opens up several alternatives to power-matching that create different types of tournaments. Pre-matching for geographic spread is one such alternative. Tournaments might be set up so that teams are guaranteed to debate opponents from across the country, not neighboring schools. There are fancy ways to do this,[13] but a simple way is to divide teams at a tournament into seven equally populated geographic regions and have everyone debate an opponent from six, excluding only their home region.[14]

[12] It is noteworthy individual teams' variability in speaker points (standard deviation of 0.67) is so large compared with the distribution of the all teams' average ability (standard deviation of 0.54). This has profound implications because many rounds can go either way. If individual team performances were less variable, then wins would be more predictable. This would improve the accuracy of traditional rankings—but logit score rankings would also improve, since they are based on now-less variable speaker points. Although this is the empirical result from a real college season, might individual team performances be less variable than 0.67? I did not separate out team effects and judge effects. If much of that variability is due to judges—and if adjusting speaker points can adequately subtract out most of the variability—then team performances would appear more consistent.

[13] Hanes, T. Russell. (2012). "Ensuring Geographic and Skill-Level Mixing at Nationals." *National Journal of Speech & Debate*, Volume 1, Issue 1.

[14] Or perhaps one of the seven "geographic regions" could be national circuit teams.

Another alternative would be to divide teams by style: critique-heavy, policymaker, and persuasive, for example. Teams could debate two opponents of each type, once on the Affirmative and once on the Negative. Recognizing that power matching is not necessary for accurate results liberates the debate community to try different tournament arrangements. All of this depends on using a more accurate ranking method, the logit score.

Of course, there might be some resistance in the community to using non-traditional methods for ranking teams. Ranking some teams above their win-loss record is how a non-traditional method improves accuracy. A non-traditional method recognizes that a team's win-loss record may be too low because of having a tougher schedule of opponents or too high because of having weaker opponents, so the method adjusts ranks accordingly. The logit score will rank some 3-3 teams above some 4-2 teams. If a tournament is not willing to consider the possibility that the better record might belong to the weaker team, there is no point to using a non-traditional method.

However, there is a reasonable limit to pulling teams up. It would look bizarre and be unacceptable to the community to say that, for example, a 1-5 team was the best at the tournament. There is a sweet spot where win-loss records are not "overridden" too often or too egregiously, and where win-loss records are weighted into the ranking realistically. Speaker points-only rankings fail both parts of this test: win-loss records are not weighted in at all, and records are overridden to excess. Logit score rankings, however, pass both parts of the test: win-loss records are weighted in, and records are not overridden excessively.

Tournaments might be nervous about using the logit score until there is widespread community approval. But approval is unlikely before the community sees the logit score in use at a few tournaments and becomes more familiar with its characteristics. There are several ways for tournaments to use the logit score in order to ease the community in to its use:

a) **Logit score-only rankings:** The top teams based on logit score break to elimination rounds, then are ranked for brackets by logit score as well. This means including a few 3-3s and possibly a few 2-4s, and excluding several 4-2s. Excluding any 6-0s or 5-1s from elimination rounds is unlikely.

b) **Record-first rankings:** All teams with winning records break, then are ranked for elimination brackets by logit score. This means excluding some high scoring teams that just had tough opponents and went 3-3, and letting in some low scoring teams that just had weak opponents and went 4-2.

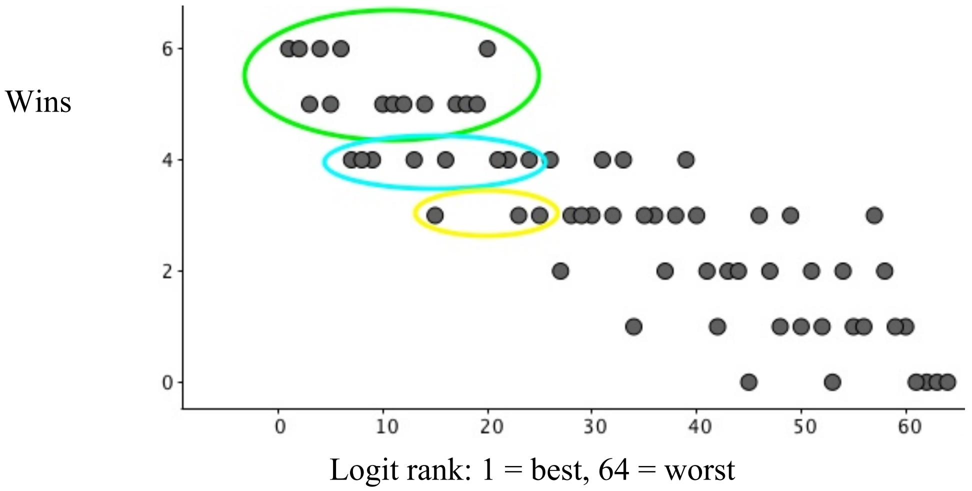
c) **Compromise rankings:** All 6-0s and 5-1s break; no 2-4s or below break; some 4-2s and 3-3s break based on logit score. Then all teams that break

> are ranked for elimination brackets by logit score. Unlike the logit score-only option, where excluding 6-0s or 5-1s is unlikely but possible, this alternative makes it impossible to exclude them. In other words, approximately the top 10% of teams by record are automatically included in elimination rounds. The middle 50% to 60% of teams—the 4-2s and 3-3s—are in the gray zone. About half of this group breaks, depending on logit scores. The bottom 30% to 40% of teams by record—the 2-4s, 1-5s, and 0-6s—are excluded automatically.

Practically speaking, the compromise has nearly the same outcomes as the logit score-only option, although it does rule out a few unpalatable outcomes. Even though it is possible that a 6-0 team is quite weak and was merely the beneficiary of good luck, the community would not stand for excluding this team. By ruling out the most unpalatable outcomes, the compromise would be much more acceptable to the debate community than the logit score-only option. Perhaps in time, as people became more familiar with the logit score, the community would move to the logit score-only option.

The record-first option is substantively the most different of the three because it includes all 4-2s and excludes all 3-3s. Yet win-loss record is not the best indicator of a team's true strength; one easier or one harder opponent is all it takes to tip the record. Because it creates no gray zone, the record-first option inevitably makes several mistakes in admitting teams to elimination rounds. While the community might have the greatest initial acceptance of the record-first option, it undermines the purpose of using a non-traditional ranking method. It would reinforce the wrong idea that win-loss record ought to be the primary way to assess a team's strength; the community would soon wonder why logit scores were used to rank teams for elimination brackets if they were not accurate enough to determine which teams break.

One randomly paired virtual tournament will suffice to show the differences that would come up using the three options:



Each dot represents one team's results. In random paired prelims, approximately equal numbers of teams end up with each win-loss record—slightly fewer for undefeateds or no-win teams—so about 9 or 10 for each record, with about 5 to 6 each in the top and bottom brackets.

The green circle represents all the teams automatically included under the compromise. The blue circle represents the 4-2s that make the cut (above about the 25th rank or so, depending on how big the partial elimination round is). Including all 4-2s would include about four weak teams. The yellow circle represents the 3-3s that make the cut. Excluding all 3-3s excludes about three moderately strong teams. No 2-4 teams rank above 25. And yes, that one 6-0 team has been disastrously lucky. It is a middling team that happened to get several easy opponents.

A philosophical objection I heard to using logit scores to rank teams is that it is so good at determining the true strength of teams, the teams' actual performance is irrelevant. Nothing could be further from the truth. The logit score is a measure of the central tendency of a team's performance; nothing more, nothing less. In the experiment, teams can and did tank their logit scores by underperforming consistently for a tournament. Though the logit score is relatively unaffected by a close win or loss, it is wrong to assume it is unaffected by a large loss or a big win. A huge upset can tank a team's score. The logit score is 85% accurate because teams are about 85% consistent in their performance. Despite the fact that the variability in speaker points is quite large, it is rare for a team to underperform or overperform for all six rounds.

CONCLUSION

The logit score has two key attributes to recommend it: it incorporates wins, schedule strength, and speaker points; and it is accurate. In this experiment, the logit score was about as accurate in ranking teams as is possible. Even with randomly paired preliminary rounds, the logit score is far more accurate than the traditional ranking method even with power-matching. No other method to combine wins and speaker points is as effective.

If the community decides to accept the logit score, then it would free tournaments to consider other methods such as pre-matching to pair rounds. Tournament directors could be confident that the pairing method will not substantially affect the accuracy of logit score rankings and would have more room to experiment. This would make traveling around the circuit more interesting, but it would also reward the best teams more accurately.

APPENDIX A: ADJUSTING JUDGE SCORES

This method for adjusting speaker points for an individual judge's bias (too high or too low, but also too spread out or too closely grouped) is straightforward. It is similar to the second-order z-scores used in current tabulation programs.

First, the raw scores from a judge are turned into z-scores. For each team $\{x_1, \ldots, x_i\}$ a judge has scored, its z-score is then

$$z_1 = \frac{x_1 - \bar{x}_j}{s_j}$$

where $\bar{x}_j$ is the judge's average score and s_j is the judge's standard deviation.

Second, average speaker points for each team in the set are calculated: $\{\bar{x}_1, \ldots, \bar{x}_i\}$. These average scores reflect several judges' opinion about each team; they reflect the community's general sensibility. The community average could be calculated either by including or excluding the judge in question.

Third, for the judge in question, an adjusted average score needs to be calculated:

$$\bar{x}_{j^*} = \frac{1}{i}\sum_{n=1}^{i} \bar{x}_n$$

This adjusted average, $\bar{x}_{j^*}$, shows what the community would have given the set of teams the judge saw. If $\bar{x}_j > \bar{x}_{j^*}$, it means a judge gives higher scores than the community. If the latter average is larger, it means the judge gives lower scores than the community.

A similar calculation needs to be done for an adjusted standard deviation, s_{j^*}, of the set $\{\bar{x}_1, \ldots, \bar{x}_i\}$.

Finally, the adjusted score is $x_{1^*} = z_1 s_{j^*} + \bar{x}_{j^*}$.

APPENDIX B: LOW-POINT WIN FORMULA

Probability of low-point win = $0.215 - 0.18748\, x^{0.131} + 0.285\, (x+1)^{-4.75}$

where $x = \left|points_{aff} - points_{neg}\right|$.

This had an $r = 0.985$ on the historical data, but it only makes sense up to a difference of 2.89 points. Beyond that difference, the probability of a low-point win must be set to zero.

APPENDIX C: WEIGHTED FOOTRULE

I used a modified version of the weighted footrule:

$$\frac{1}{19.71} \cdot \sum \frac{|deviation|}{(true\ rank)^{0.39}}$$

where deviation is the difference, for each team, between its true rank and its observed rank in a tournament. Raising the denominator to the 0.39 power has a special result: half of the weight of the weighted footrule is put on the rankings given to the top 22 teams—those that have winning records.[15] The remaining two-thirds of the teams in the virtual tournament count for only half of the weight.

To scale the weighted footrule, I divided by the sum of weighted deviations by 19.71 because

$$\sum_{n=1}^{64} \frac{1}{n^{0.39}} = 19.71$$

which is the assumption of uniform deviations. Thus, because of dividing the sum by 19.71, if all teams are off by 1 rank position, then the weighted footrule score will be 1. If all teams are off by 2 rank positions, then the weighted footrule score will be 2. The weighted footrule is therefore on the same "scale" as the mean absolute deviation.

[15] At a tournament of 64 teams run with power-matching, one team will go 6-0, six will go 5-1, and fifteen will go 4-2. The exact number with each record would differ for the random and pre-matched tournament conditions.

APPENDIX D: CALCULATING THE LOGIT SCORE

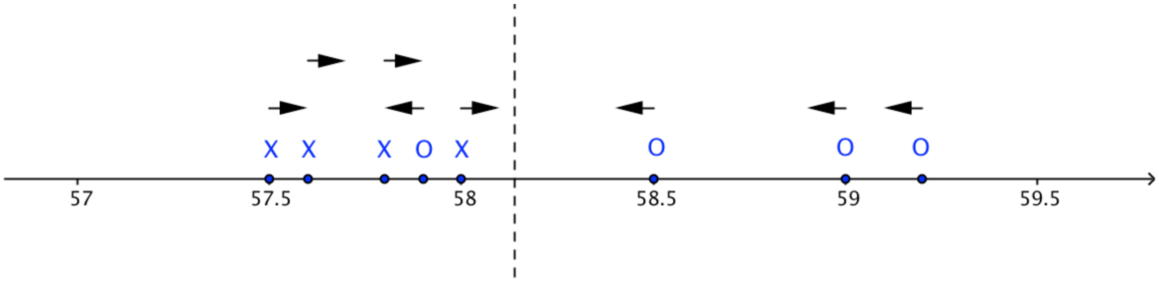
For each team, all of its opponents, each opponent's median (or average) speaker points, and the win or loss are listed. Here is a hypothetical list for team E:

Opponent	Opponent median speaker points	Win?
A	57.6	1
B	57.8	1
C	57.9	0
D	59.2	0

Furthermore, for each team, its speaker points in every round are listed. These results are coded 0 or 1, above or below its median. If two speaker point results are both exactly at the median, one is coded 1 and the other 0. If three results are exactly at the median, the third is one coded 0.5. Here is this list for team E based on its median speaker points of 58.25:

Round	Team speaker points	Below median?
1	57.5	1
2	58.0	1
3	58.5	0
4	59.0	0

The two lists are put together. The combined list serves as the basis for calculating the team's logit score.



In the graphic above, the 1s are marked by Xs. The logit score, marked at the dashed line, is lower than an average of all the data points. With the logit score, all wins exert upward pressure on the score—marked by the right-pointing arrows—and all losses exert downward pressure—marked by the left-pointing arrows. But the pressure diminishes the farther away the results are from the central grouping.

In other words, the loss to the 57.9 team is more consequential to this team's final score than the losses to the high-scoring teams or unrealistically high speaker points. The logit score estimates each team's strength based on **the score that is most consistent with:**

a) the opponents it beat,
b) the opponents it lost to, and
c) the speaker points it received.

The logit score is extremely resistant to inconsistent speaker points, to wins against weak opponents, and to losses to strong opponents. Results in close rounds and typical speaker points are more important in calculating the logit score.

A team's median speaker points can serve as the initial estimate of its logit score, L_0. A logistic function is used to "retrodict" the wins and losses for team using the initial estimate of the logit score and each opponent's speaker points, O:[16]

$$\text{Probability of win} = \left[1 + e^{-2.436(L-O)}\right]^{-1}$$

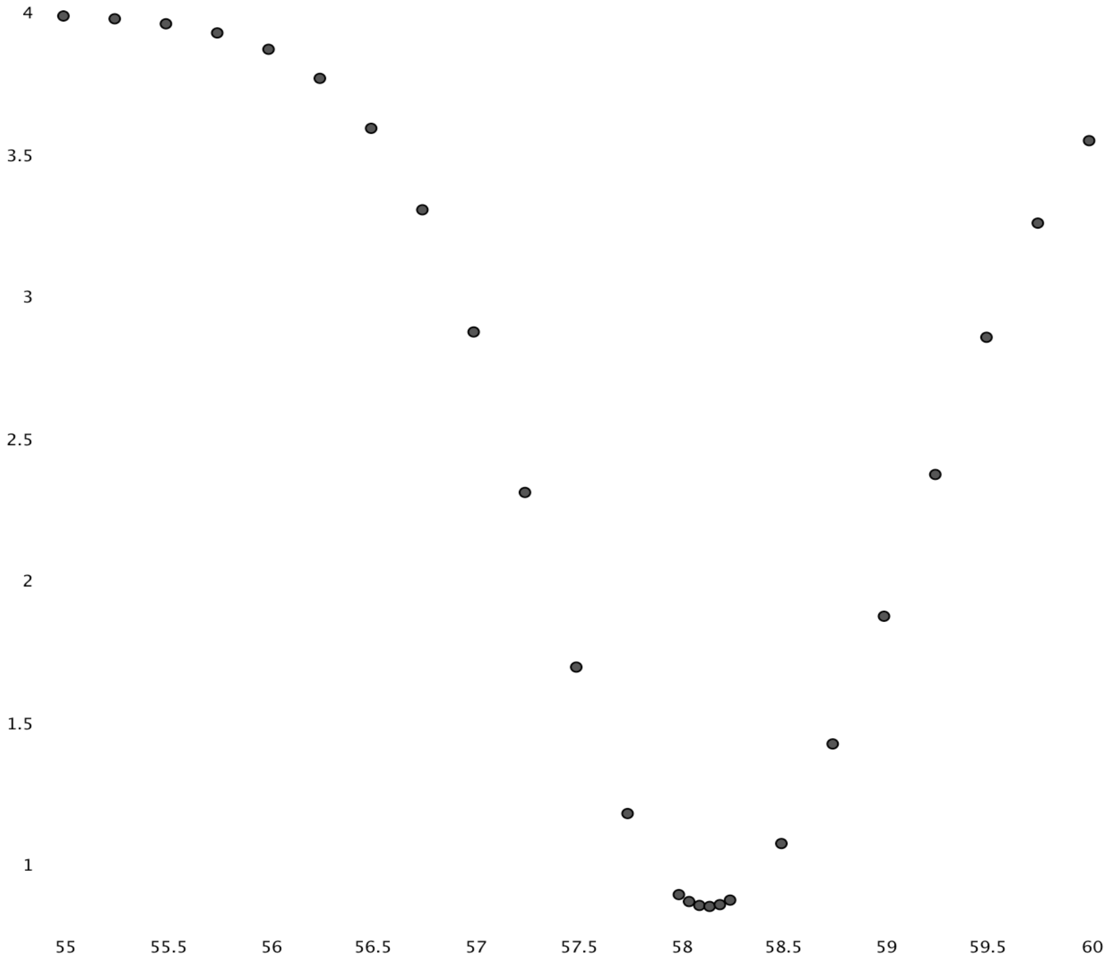
The same is done for each team's speaker points per round, P, replacing O. Thus, the list for team E contains the following retrodictions based on the team's $L_0 = 58.25$:

Opponent/Round	O/P	Win/Below median?	Retrodiction	Error
1	57.5	1	0.86	0.02
A	57.6	1	0.83	0.03
B	57.8	1	0.75	0.06
C	57.9	0	0.7	0.49
2	58.0	1	0.65	0.12
3	58.5	0	0.35	0.12
4	59.0	0	0.14	0.02
D	59.2	0	0.09	0.01

Retrodictions near 0.5 mean a team's average performance could have resulted in a win or a loss in that round; retrodictions near 1 indicate a near certain win; retridictions near 0, a near certain loss. The error for each retrodiction is calculated,

[16] This is the logistic function that best fit the college debate season studied. This can be updated with further data but works well: http://art-of-logic.blogspot.com/2012/07/probability-of-upsets.html.

squared, and summed to produce a sum of squared errors (S.S.E.).[17] The logit score is then raised or lowered to produce the minimum S.S.E.



The horizontal axis notes the logit score tested; the vertical axis records the S.S.E. This technique is known as a logit regression and is a well-established mathematical tool. In our example, $L_{final} = 58.14$, as this is the best-fitted logit score to the observed data.[18]

[17] In a previous iteration, I advocated using weighting as well, but it is not necessary for six-round tournaments. It may be beneficial for season-long analyses, however.

[18] An important note for programmers seeking to use the logit score: the S.S.E. vs. possible logit score does not *always* follow this simple shape. In some cases, the shape is best described as a

The revised table looks like this:

Opponent/Round	O/P	Win/Below median?	Retrodiction	Error
1	57.5	1	0.83	0.03
A	57.6	1	0.79	0.04
B	57.8	1	0.7	0.09
C	57.9	0	0.64	0.41
2	58.0	1	0.58	0.17
3	58.5	0	0.29	0.09
4	59.0	0	0.11	0.01
D	59.2	0	0.07	0.004

By pushing down the team's logit score from the initial estimate, the error on team C has been noticeably lowered (from 0.49 to 0.41) while the error on round 2 has increased slightly (from 0.12 to 0.17). The errors on B and 3 changed by only 0.03, and the other four errors changed by 0.01. The results in the middle matter the most; outliers do not affect results significantly.

To save time, computer programs might calculate each team's logit score to one or two decimal places, then return to calculate further decimal places only in the case of ties.

Although it is not possible for debaters to check the calculation of the logit score, it should be stable enough from tournament to tournament for teams to be able to track their progress. In some ways, the logit score is quite conceptually similar to adjusted speaker points, except instead of only being adjusted for judges, it is also adjusted for wins, losses, and opponent strength.

To distinguish the logit score from speaker points, it may be helpful to multiply the final result by $\frac{5}{3}$ to make the logit score on a 100-point scale. This team would then have a logit score of 96.9.

plateau with a sharp divot at the best logit score. For these teams, the most extreme scores may have S.S.E.s slightly *lower* than the plateau. When searching for the best logit score, a sufficiently robust search method needs to be used, given such an odd-shaped distribution.