# National Journal of Speech & Debate

Volume V: Issue 2

January 2017

# TABLE OF CONTENTS

# INTRODUCING THE LOGIT SCORE: A NEW METHOD TO RANK DEBATE TEAMS

BY T. RUSSELL HANES*

* The author has a bachelor's in mathematics from Columbia University and a master's in teaching mathematics from Lewis & Clark College. He has spent several years investigating and blogging about tournament mathematics at http://art-of-logic.blogspot.com. He is finishing a textbook entitled *Arguing with Data*, an introduction to statistical concepts for debaters. The author can be reached at russell.hanes@gmail.com.

## INTRODUCTION

The purpose of collecting team measures (win–loss record and speaker points being the main two) is to rank teams. The ranks determine which teams move on to elimination rounds and which teams are invited to round robins and championship tournaments. The importance of getting ranks right is clear. Wringing out a bit more accuracy has led the debate tabulation community to develop new methods and measures, including different power-matching procedures and second-order z-scores. For several years, I have been looking into ranking methods from non-debate fields, such as sports and technology.[1]

An ideal ranking method would be based on only the performance at one tournament; it would be less sensitive to outliers resulting from inconsistent judging and varying schedule strength; it would be simple to program into software; and finally and most importantly, it would yield accurate rankings. Based on my research, I believe that the ideal method is the logit score.

For this analysis, I looked at the varsity or open rounds in cross-examination college debate in 2014-15. This included only invitational tournaments, not round robins, districts, or nationals. Only preliminary rounds were included because elimination rounds lack speaker points. The data set included 510 teams and 5,313 rounds. The measures included in the data set were wins, points, opponents, and judges, among others.

The primary method used in this research is retrodiction analysis. The teams were ranked using various methods. The purpose is to see whether the resulting rankings make sense. Each set of rankings was used to "retrodict" the actual results. For example, suppose that ranking method 1 ranks team A better than B. If team A actually beat team B during the season, then method 1 has successfully retrodicted the result. On the other hand, suppose that ranking

[1] I would highly recommend AMY N. LANGVILLE AND CARL D. MEYER, WHO'S # 1? THE SCIENCE OF RATING AND RANKING (2013) a source of possible methods.

method 2 ranks B the superior team, which means method 2 has made an incorrect retrodiction of the actual result. Retrodiction analysis allows multiple ranking methods to be compared on a real data set for empirical validity. Of course, actual results are not necessarily "correct"—there may be judging errors and inconsistent performances from either or both opponents, and there is a degree of randomness in any decision in close rounds—but a ranking method that repeatedly disagrees with the real outcomes can be dismissed as hopelessly invalid. Because the goal is to look at possible alternatives to ranking teams at tournaments by win–loss record, the traditional method serves as the baseline for comparison.

## THE TRADITIONAL METHOD

The traditional ranking method starts with the win–loss record, then incorporates some form of speaker points as the first tiebreaker. The general sense of the debate community is that wins and losses are objective, while speaker points are subjective, though in truth neither is entirely objective nor subjective. They both exist in the gray zone of rigorous but human-derived judgments, like medical diagnosis or expert poker play. If we do not trust judges to give accurate speaker points, why do we trust them to give accurate wins? Both wins and points can be valuable information to rank teams if used properly. A social scientist would say both wins and points have high levels of intersubjective agreement; judges for the most part concur with each other.

Mathematically, a win–loss record could be considered less informative than average speaker points because the data are less precise. For an example, one team in the data set with 32 rounds has a confidence interval around its win percent of (56%, 88%),[2] which could move it from the 72nd to the 99th percentile. If this seems over-broad, think about this: flipping one win to a loss would make an enormous difference on the team's ranking—there are so many teams that are so close together—which is why the confidence interval is so large. On the other hand, the same team's confidence interval around its average speaker points is (57.3, 57.9),[3] which could move it from the 88th to the 96th percentile—a much narrower range. Average speaker points for a season are so stable that there is little one judge could do to affect them, aside from giving a zero. This example team shows how imprecise even *season-long* win–loss data are; take several steps back to figure out how imprecise *tournament* data are.

[2] The 95% confidence interval for a proportion: in this case, 72 win percent ± 16. This assumes that rounds are assigned randomly. Because of power-matching, this confidence interval is an underestimate. More close rounds should mean wider confidence intervals and therefore less precision.

[3] The 95% confidence interval for a mean: in this case, 57.6 points ± 0.3.

Yet despite the narrower confidence intervals, ranks based on average points retrodict fewer rounds correctly than ranks based on win percentages. The issue is that speaker points, while being more precise information, also exhibit some biases. This could include teams that are good at speaking but not doing sufficient research to win rounds or weaker teams that win rounds primarily on the strength of their squad's research. Wins are the more meaningful currency in spite of the fact that win–loss record is a less precise, blunter measure. Wins are both more important and yet also more vague information than speaker points.

## WEIGHTED WINS

Of course, this mathematical imprecision should accord with practical experience, too. A win does not indicate the margin of the victory; blowout rounds and near-tied rounds all receive one, equal win. Crucially, the win–loss record depends on the strength of opponents debated.[4] Several attempts have been made, by me and others, to correct win–loss records for the schedule strength. My preferred method used "weighted wins" and "weighted losses" that added extra weight into wins against strong teams and less weight into wins against weak teams. In theory, this could work well and would have been simple to add to tabulation programs.[5] When I tested it on real data, weighted wins–losses failed. The main problem seems to be that wins can be blowouts or near-ties, and weighting for opponent strength does not address this difficulty.[6]

While it is conceivable that a better method of weighting rounds is possible, the clearer path forward is to find a way to rank teams that utilizes speaker points to incorporate the margin of victory into wins. This kind of ranking method would use all the available information and should be able to deliver more accurate ranks with narrower confidence intervals and less bias than either wins alone or points alone. The traditional method (wins then points as separate categories) made sense before more sophisticated tools existed to fuse wins and points into one combined measure, but sufficient computing power today allows advanced calculations to be made in the blink of an eye.

[4] Power-matching does not entirely solve this problem. In some cases, two teams that were adjacent in rank had average opponents 40 percentiles apart over a season.

[5] I also extensively tested matrix-based methods where linear algebra is used to find the best fitting value for each team considering judges, opponents, and wins. While some of these methods would be practical for round robins, they are prohibitively computer-intensive to be used for large tournaments or entire seasons.

[6] The other problem is that differing schedules artificially inflate or deflate opponent records as well, which makes it difficult to weight the wins and losses using this as a measure of opponent strength.

## ELO/GLICKO METHOD

An Elo or Glicko method seems a good candidate for calculating a fused measure. The basic method for both is simple: each team has a running, season-long rating. When they debate, the winning team takes points off the losing team's rating; both teams' ratings are updated after the round based on the result. If the teams are close in ratings, few ratings points are traded whichever team wins. If the result is an upset, the lower-rated winner wins a lot of points off the higher-rated loser. On the other hand, a higher-rated winner gets very few additional ratings points off the lower-rated loser. It is fair to think of the points being traded relative to a "surprise factor": the more surprising the result, the more points are traded. The Elo/Glicko methods can also use the margin of victory as well as result to move teams up or down the ratings. A win by a larger margin of victory can move a team up by more than a win by a smaller margin. In debate, the margin would be the difference in speaker points between the two opponents in a round. However, Elo and Glicko methods may elect not to use margin of victory at all. Elo and Glicko methods have been used successfully for sports rankings from football to baseball.[7]

Teams can start the season at parity (or carry over last year's ratings), but as results come in, the teams are re-rated. As such, earlier results are less important than later results in determining final ratings. These methods work well for season-long scales, but it is tricky to use them for ranking during a single tournament. If the teams start each tournament with their season-long scores in place, little reshuffling might happen during the tournament. In that case, the highest-rated teams before the tournament would be the teams that break to elimination rounds. On the other hand, if the teams start a tournament with a fresh slate, a few chance results during the tournament could cause odd teams to move into elimination rounds.

Overall, this is a known difficulty for Elo and Glicko ratings: setting the sensitivity to change, known as the K factor, just right is hard. If it is set with too much sensitivity to unusual results, then the rankings will lack stability; if the scale is set with too little sensitivity, then the rankings will not allow for movement to happen.[8] In sports, statisticians try to set the right sensitivity by analyzing past seasons, but there is no way to empirically determine that a particular sensitivity is "correct" for a short timeframe. Over an entire season, this

[7] For practical descriptions, I would recommend reading:
http://fivethirtyeight.com/datalab/introducing-nfl-elo-ratings/ and
http://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/
[8] https://en.wikipedia.org/wiki/Elo_rating_system#Mathematical_issues

matters little.[9] More or less sensitive methods will likely converge on the same ranks. However, it matters immensely for a single tournament. Elo or Glicko ratings are terrifically well suited to season-long ratings but are incompatible with deciding which teams break at a tournament. There are too few rounds at one tournament to give an accurate rating.

## LOGIT SCORE

The logit regression is a well-known method in statistics, and the logit score is an adaptation of this method to the debate context. A detailed method is described in the appendix, so this section only overviews its basic details. A logit score takes a team's various points, wins, and opponents and finds the likeliest team strength resulting in all those data. For example, say that team A has won blowout wins against weak opponents but lost blowout rounds to strong opponents. To calculate a logit score, all the results from a tournament are put into one probability model. Team A is likelier to win against weaker opponents by large margins if the team is highly rated—but that makes it less likely to lose against stronger opponents by large margins. Conversely, team A is likelier to lose against stronger opponents by large margins if the team is poorly rated—but that makes it less likely to win against weaker opponents by large margins. The logit score is the balance point that makes its wins and its losses, together, maximally probable. In the case of team A, it will be rated as mediocre. The logit score is much less sensitive to single outlier results than wins alone or points alone because all the information—including opponent strength—is fit into a single probability model.

For the 2014-15 season, I calculated the logit scores for each team and ranked them from 1 (the best) to 510 (the worst). I also ranked the team by win percentage.[10] How do the two methods of ranking compare? The Pearson correlation coefficient was 0.86, a high degree of correlation. (Pearson's measures how well the score in one variable matches to the score in another on a scale of 0, no matching, to 1, perfect matching.) The mean absolute value of difference in rank between the two methods was 10.5 percentiles, although the median was only 7.4 percentiles. This means that, for example, a team ranked 52nd percentile in one measure might be 59th or 62nd percentile in the other measure (or better). However, about 96% of the logit score ranks are within the confidence interval for

[9] However, one season-long problem is that teams might avoid tournaments to protect a high rating; ratings need to slowly "decay" to encourage activity.

[10] Then by total rounds, then by average speaker points. I also tested out several other possibilities for ranking by win–loss record, including the binomial probability of a team's wins by chance, the number of wins over 50%, total number of wins, and others. Nothing worked as well as ranking by win percentage.

each team's win rank. Along the same line of evidence, 87% of the logit score ranks are within $\pm 2$ rounds of the team's actual wins. It is easy to imagine that, over an entire season, any team could win or lose two close rounds it should not have. The difference of $\pm 2$ rounds is slight, but it does move a team substantially down or up in the win ranks because of the imprecision of the win–loss record and the clumping together of teams. The 10.5 percentiles of difference between a team's win rank and a logit score rank mostly represents shuffling among neighboring teams. These facts suggest the logit scores are not wild and off base, but it is also consistent with the possibility that logit scores do bring in new, useful information that is ignored in win ranking.

Given that the goal was to look for a measure for ranking at single tournaments, examining an entire season may seem counterintuitive. However, it is the best way to assess empirical validity. In the 2014-15 season, the logit scores retrodicted slightly more rounds correctly than win percentage, 79% to 78%. This is a tie. The logit score is the only ranking method I tested that did as well as win ranking. Most other ranking methods did substantially worse.[11] While a tie is not evidence to reject win ranking in favor of logit score ranking, it is evidence to *not reject* logit score ranking. It passes a crucial test of reasonability as a ranking method. On the basis of this result, I looked at logit score ranking in more detail.

## LOGIT REGRESSIONS

Ranks can be used, not only to retrodict a result, but also to assign a probability to the outcome. Two closely ranked opponents should be assigned probabilities in the 45-55% range: a coin toss. On the other hand, a strong and weak opponent might be assigned probabilities of 90% and 10% of winning respectively. Analyzing these probabilities allows us to investigate whether a ranking method is identifying teams' strength in finer detail than we can determine from the overall retrodictions-correct percentage alone. To do this requires using a best fitting logistic regression model. For win ranks, the best fitting regression model is:

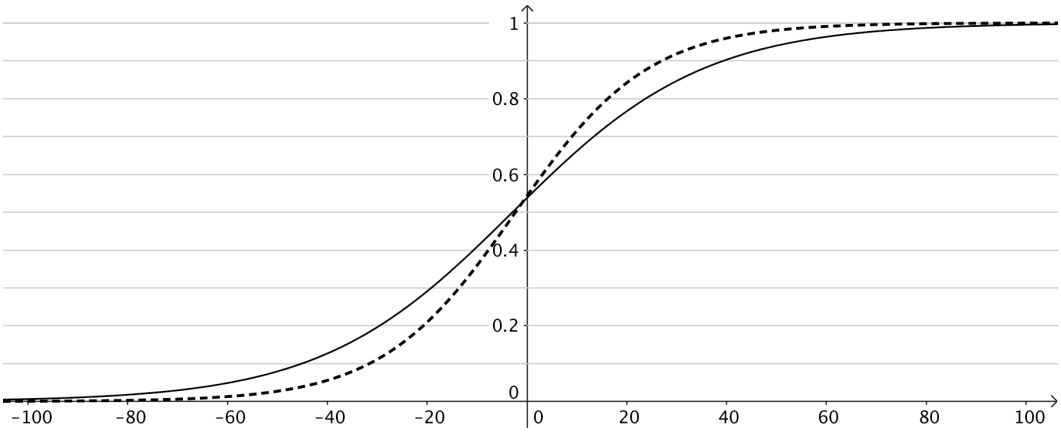
$$probability\ of\ AFF\ win = \left[1 + e^{-0.052 \cdot (AFF\ percentile - NEG\ percentile + 2.9)}\right]^{-1}$$

For logit score ranks, the best fitting regression model is:

[11] I did not evaluate Elo or Glicko ratings in this analysis because of their inappropriateness for single tournaments.

$$probability\ of\ AFF\ win = \left[1 + e^{-0.075 \cdot (AFF\ percentile - NEG\ percentile + 2.3)}\right]^{-1}$$

Both models worked best with some affirmative bias built in, represented by the 2.9 and 2.3 in the respective equations. These show that the Negative needed to be about 2 to 3 percentiles better than the Affirmative to make the round evenly matched. (In this season, Affirmatives won 52.5% of rounds.) Here is what the models look like:




The horizontal axis shows the affirmative percentile minus the negative percentile. The vertical axis shows the probability of an affirmative win. The solid curve is the win model; the dashed curve is the logit score model. Neither model passes through (0, 0.5) because of the affirmative bias on this topic. Rounds where the two opponents are similarly matched, around zero on the horizontal axis, are near toss-ups. Rounds at the extreme right, where the Affirmative has a huge advantage, near a 100% chance for an affirmative win; rounds at the extreme left, where the Negative has a huge advantage, near a 0% chance for an affirmative win.

Probabilities are not certainties, however. A 95% chance that the Affirmative wins means just that: it is expected the Negative wins about 1/20 of these rounds. These upsets are not necessarily low-point wins. Upsets are rounds when the lower-ranked team wins, but the winning team could receive lower, higher, or tied points to its opponent. The lower rank is about the team's season-long performance, not the result in that individual round. Both models call about 20% of all the rounds in the season upsets. Low-point wins, however, account for only about 6% of rounds in the 2014-15 season. Most low-point wins occurred when the losing team only slightly edged out the winning team in points,

indicating a closely matched round. Upsets occur up and down the scale; low-point wins are mostly concentrated near zero on the horizontal axis.

### RETRODICTIONS BY CERTAINTY

The logit score model is more aggressive in assigning the probabilities. The two models differ the most around +23 percentiles, with the win model assigning 79% probability for an affirmative win and the logit score model assigning 87% probability; and at -26 percentiles, with the win model assigning 23% probability for an affirmative win and the logit score model assigning 14% probability. Although these are the largest gaps, the logit score model assigns probabilities more aggressively than the win model at every given percentile difference in team strength. Yet despite the increased aggressiveness across the board, logit score ranks edge out win ranks for overall accuracy and also accuracy in most categories:

| Certainty by win model | % correct by win rank retrodiction [12] | % correct by logit rank retrodiction | Difference in % correct | Agreement between models | Count of Rounds | Estimated distribution of rounds[13] |
|---|---|---|---|---|---|---|
| 0.50 – 0.55 | 0.571 | 0.589 | 0.018 | 0.591 | 992 | 422 |
| 0.55 – 0.65 | 0.644 | 0.681 | 0.037 | 0.756 | 891 | 577 |
| 0.65 – 0.75 | 0.738 | 0.740 | 0.002 | 0.833 | 884 | 579 |
| 0.75 – 0.85 | 0.779 | 0.780 | 0.001 | 0.913 | 929 | 790 |
| 0.85 – 0.95 | 0.884 | 0.862 | -0.022 | 0.952 | 1202 | 1183 |
| 0.95 – 1.00 | 0.957 | 0.958 | 0.001 | 0.992 | 911 | 1973 |
| **Overall average** | 0.784 | 0.788 | 0.004 | 0.866 | | |

In this table, the rounds are organized into six categories by the confidence the win model gives the retrodicted results. The first category includes all the rounds in which the teams are so evenly matched, the model says either team winning would be a nearly equally likely outcome (45-55%). The second category includes all the rounds in which one team has only a slight advantage over the other and the model gives the favored team a 55% to 65% chance of winning. The final category includes all the rounds that are so unevenly matched that the favored team is given above a 95% chance to win. I picked these categories because they divide all the rounds into approximately equal-sized groups.

[12] An astute observer might wonder why the retrodictions do not seem well calibrated. Calibration refers to whether model-assigned probabilities obtain over observed data. For example, a weather forecasting model is well calibrated if, for the set of all days it forecasts a 10% chance of rain, it does in fact rain about 10% of the time. It seems like the win model is not calibrated because columns 1 and 2 are discrepant. However, this is entirely a result of using percentile-based models, which is necessary to make a fair comparison between the two models. Teams follow a Normal distribution of strength, so the gap between the first and second place team is larger than the gap between two adjacent, middle ranked teams. Using percentiles has the effect of exaggerating some differences, between two mediocre teams, and also minimizing others, such as between two excellent teams. I checked the raw-score versions of the models; they are perfectly well calibrated.

[13] The counts of rounds for the different categories confused me at first. It seems as if power-matching should have produced more close matches—more rounds in the first and second rows—than it did. I ran a simulation of 5,313 rounds paired at random, the result of which is listed in this column. Random matching would produce many more lopsided matches than had occurred in the real debate season, so power-matching is in fact working to pair more close matches than expected by chance.

In two categories, logit score ranks produce more accurate retrodictions: the first and second. In three categories, the win ranks and logit score ranks are tied. Only in one category, the fifth, do win ranks produce more accurate retrodictions.

## CYCLES

How could the ranks differ so much—on average by 10.5 percentiles between the win rank and logit score rank for each team—and yet both ranks be approximately equally accurate in retrodictions? The agreement column sheds some light on this, showing in what percentage of rounds the logit score rank and win rank pick the same winner. In the sixth category, for example, the win rank and the logit score rank nearly always agree in retrodicting the winner: about 99% of rounds. The two opponents are so far apart that both ranking methods pick up on the difference. In the first category, on the other hand, the two ranks agree on the winner in only 59% of rounds—yet both the win rank and logit score rank retrodict about the same number of rounds correctly. The two methods arrive at nearly equal accuracy by picking *different sets of winners*.

The answer to this "paradox" is explicable with an example. Imagine three teams competing in a round robin. Team A beats B, team B beats C, and team C beats A. This is known as a cycle, the shortest one there can be. One can come up with six possible rankings using any number of tiebreakers, but all six rankings retrodict exactly one result incorrectly. For example, the ranking A > B > C misses C beating A, but the ranking C > A > B misses B beating C. Any ranking will retrodict 67% of rounds correctly, but two different rankings could agree on 0%, 33%, or 67% of the ranks.

This problem occurs on a much larger scale in the analysis I did. Teams ranked near to one another will, when they debate, split rounds about 50-50, but the exact outcomes will be random and thus create cycles of varying length. Changing the ranks may move around some ineluctable contradictions. These results show that, for the most part, the two ranking methods agree on the big sort of teams from top to bottom and mostly differ on the specific tiebreaking of neighboring teams.

## CAVEAT

The logit score ranks are both more aggressive, meaning the method sees fewer ties, and yet are also slightly more accurate in retrodicting results. Given the same information, the logit score works better as a tiebreaker than the traditional win-loss method. It is important to note, however, that there is a low variability of speaker points in the college debate world, which allows their use in

determining margins of victory for the logit score. For the teams in this population, the median standard deviation in speaker points is only 0.65. Except for a handful of obvious coding errors, there is a clear upper limit of 1.3 for the standard deviation of speaker points for any team. I did not even need to bother adjusting speaker points for judge inconsistency. Despite the impression that speaker points are inaccurate and subjective, the college debate community is more or less one circuit with a fairly homogenous standard of excellence and low variability in speaker points given. High school debate, on the other hand, is another matter. Given that various high school circuits are so distinct in character, it seems reasonable to assume speaker points are highly variable. A season-long use of speaker points as the basis for determining a regression model might doom logit scores to fail. However, it could be possible that creating a regression model *per tournament*, rather than for an entire season, would account for circuit variability and would work for high school debate as well. Whether or not this is true for high school debate, it is clear that the regression model for college debate works well and should be studied for use in ranking teams at single tournaments.

## FURTHER RESEARCH

There is limited utility in doing more retrodiction analyses. Logit scores pass this first test. The better test now is tournament simulation. Researchers can take a representative sample of teams with known strengths; set up a model for realistic speaker point variability, low-point wins, and judging effects; run the computer simulations; and then see which ranking method yields the most accurate ranks.[14] The simulation tournaments could be run under various pairing conditions—power-matching, random, and presets by strength groupings (Group A vs. Group A, B, C, etc.)—to test whether the accuracy of the ranking methods are affected by tournament type. Various questions could be tested, such as: Would adjusting speaker points for judging inconsistencies, perhaps using z-scores or second-order z-scores, improve the accuracy of logit scores? Would replacing each team's opponents' average speaker points with their logit scores, then re-running the logit score analysis, improve accuracy?

## CONCLUSION

Why bother to do the logit score calculation, even though it only has similar accuracy to win percentage? My answer is two-fold. First, logit scores

[14] Although point rankings and z-hybrid rankings (z-score of win percentage plus z-score of points) were notably less accurate in my retrodiction analysis, it might be worth including them as baselines in tournament simulations. One might also chose to include weighted wins–losses for comparison as well.

have similar accuracy to win percentage *over a season*, but they might have greater accuracy *for a single tournament*. A season-long record is great information, not available during individual tournaments, so it is possible logit scores may have greater accuracy at a short time scale. Second, if logit scores are merely equally accurate, then why not decide which one is superior on the basis of the richness of information it is based on? Logit scores are based on more data: wins, opponents, and speaker points; the beauty of the method is allowing these to be fused together into one score. It matters immensely for who breaks. A good team that does not break because of tough opponents might get an extra nudge up and break because the logit score factors in its opponent strength in a fair and reasonable way. The spirit of fair competition compels us to think about what we reward.

## APPENDIX: CALCULATING THE LOGIT SCORE

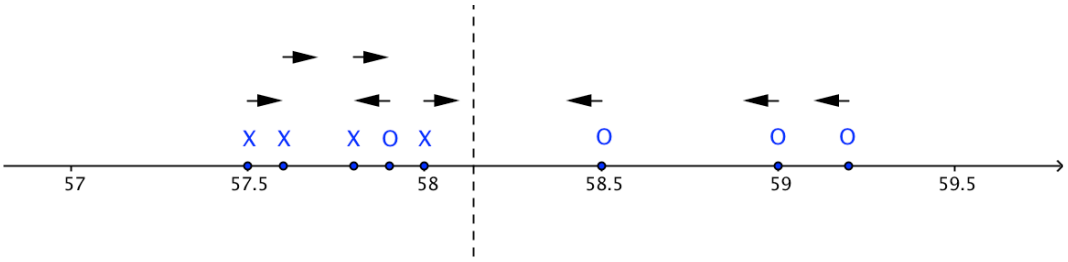
For each team, all of its opponents, each opponent's average speaker points, and the win or loss are listed. Here is a hypothetical list for team A:

| Opponent | Opp. spkr. pts. | Win? |
|---|---|---|
| B | 57.6 | 1 |
| C | 57.8 | 1 |
| D | 57.9 | 0 |
| E | 59.2 | 0 |

Furthermore, for each team, its speaker points in every round are listed. These are coded 0, 0.5, or 1 for above, at, or below its median speaker points, respectively. Here is the list for team A based on its median speaker points of 58.25:

| Round | Team spkr. pts. | Below median? |
|---|---|---|
| 1 | 57.5 | 1 |
| 2 | 58.0 | 1 |
| 3 | 58.5 | 0 |
| 4 | 59.0 | 0 |

The two lists are put together to make one list for the team. This list serves as the basis for calculating its logit score.




In the above graphic, the 1s are marked by Xs and the zeroes by Os. The logit score, marked at the dashed line, is in this case lower than the mean of all the data points. This is because losses to high-scoring opponents inflate the mean but not the logit score. With a logit score, all wins exert upward pressure on the score—marked by the right-pointing arrows—while all losses exert downward pressure—marked by the left-pointing arrows. The logit score settles at the point where these forces balance. In other words, the *single loss* to the 57.9 team and *single round* of 58 speaker points are far more consequential to this team's final logit score than the *three extreme results* at 58.5, 59, and 59.2. Those are outliers and not so influential on the logit score.

There is a zone or a range of opponents against whom we expect a team to go 50-50, and that zone for team A is clearly below 58.5 and clearly above 57.5. Based on the evidence of the given results, the evenly matched zone is around 58. A team's median speaker points can serve as the initial estimate of its logit score. A logistic function[15] is used to assign the probability of a 1 using the initial estimate of the logit score and each speaker point result (opponent or self) listed in column 2 of the tables above:

$$Probability\ of\ 1 = \left[1 + e^{-2.436 \cdot (logit\ score\ -\ points)}\right]^{-1}$$

Thus, the list for team A contains the following retrodictions listed below based on the team's median of 58.25:

[15] This is the logistic function that best fit the speaker points and actual results of the 2014-15 season.

| Round | Points | Actual Result | Probability of 1 | Error |
|---|---|---|---|---|
| 1 | 57.5 | 1 | 0.861 | -0.139 |
| B | 57.6 | 1 | 0.83 | -0.17 |
| C | 57.8 | 1 | 0.75 | -0.25 |
| D | 57.9 | 0 | 0.701 | 0.701 |
| 2 | 58 | 1 | 0.648 | -0.352 |
| 3 | 58.5 | 0 | 0.352 | 0.352 |
| 4 | 59 | 0 | 0.139 | 0.139 |
| E | 59.2 | 0 | 0.09 | 0.09 |

The error for each retrodiction is squared, weighted, and summed to produce a sum of weighted squared errors (S.W.S.E.). The weighting is based on the number of 1s, 0.5s, and 0s. If there are twenty 1s and ten 0s, then the squared errors for the 1s are multiplied by 1/3 and the squared errors for the 0s are multiplied by 2/3. Failing to weight means a team with a losing record could have its logit score pushed to zero, or a team with a winning record an infinite score. At the heaviest weighting, a team could have nothing but losses, and thus the ratio of 1s to 0s would be 3:1. The losses (n rounds) plus the speaker points above the median (n/2) would be three units compared to the one unit of the speaker points below the median (n/2).

The logit score is raised or lowered to produce the minimum S.W.S.E.[16] This is the "where the forces balance" analysis mentioned above. In this example, the final logit score is 58.14. 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 order to avoid confusion with speaker points, perhaps the logit score could be reported on a 0 to 100 scale.

[16] An important note for programmers seeking to use the logit score: the S.W.S.E. vs. possible logit score does not follow a simple parabola. The shape is best described as a plateau with a sharp divot at the optimal logit score. In fact, the most extreme logit scores can have S.W.S.E.s slightly *lower* than the plateau. When searching for the optimal logit score, a sufficiently robust search method needs to be used for this oddly shaped distribution.