Monash Debating Review

An annual publication of the Monash Association of Debaters

Volume 12. 2014

2014 EDITORIAL TEAM

Editor in Chief: Rebecca Meredith
Associate Editors: Gemma Buckley, Brett Frazer, Vihasini Gopakumar, Freddy Powell, Aditya Shetty, Milan Vignjevic
Publication Manager: Sam Whitney

Changing Landscapes: Proposals for Debating Reform

Adjudication and Motions

Commentary and Critique: the Functioning of Tournaments

How (not) to Run Worlds: Advice from two people who needed it – Harish Natarajan and Michael Baer

Transgender exclusion in debating: A case for pronoun introductions – Crash Wigley

An Evaluation of Four-Team-Per-Contest Swiss (Power Paired) Tournament Structures Using Computer Models in Python – Neil du Toit

It Actually Has a Real-Life Function: Debating as a Pedagogical Tool in Singaporean Education Introduction – Huiyi Lu

Proudly powered by WordPress

Monash Debating Review

An annual publication of the Monash Association of Debaters

Introducing Elo Ratings in British Parliamentary Debating

Ashish Xiangyi Kumar, Michael Dunn Goekjian, Richard Coate[1]

Introduction: The Elo Ranking System

Debating is a competitive hobby. Part of the pleasure of debating comes from being able to know how good one is compared to other debaters. This explains, inter alia, speaker and team tabs. However, speaker and team tabs, as well as results from individual debates, do not often provide us with information we might want to have.

We propose implementing the Elo rating system in British Parliamentary debating ("BP debating") to solve this problem. The Elo rating system calculates the relative skill levels of players in competitor-versus-competitor games. A detailed explanation of the mathematics of the Elo mechanism is to be found in the next section, but our proposal can be summarised thus:

1. To begin with every speaker is given a certain number of Elo points – we propose 1500. This is a player's *Elo rating*. (1500 points will also be given to any individual who is beginning British parliamentary debating.)
2. When speakers form teams, their team will be given a *team rating* – this is the average of the two speakers' Elo ratings.
3. When a team wins, it will *steal* points from the losing team. These points will be added to the speakers' Elo ratings. A team loses to any team ranked above it in a room and wins against any team ranked below it. So a team that is 3rd in a debate wins against 1 team and loses against 2 teams.
4. The number of Elo points stolen is determined by the *gap in the team ratings*

and *not* the gap in individual speaker Elo ratings[2]. Winning against a relatively weak team results in a small number of Elo points stolen; winning against a relatively strong team results in a large number of Elo points stolen.

5. Over time, speakers' Elo ratings will change to reflect their debating ability.
6. Speakers are globally ranked according to their Elo rating. There should also be ESL and EFL rankings. We hope that there will be regional rankings too.
7. For the Elo rating system, *both* in-round and out-round performance can be taken into account. This is because even in out-rounds where full team rankings are not produced, we know *for certain* that the teams progressing to the next out-round have beaten the 2 teams that have not progressed to the next out-round. We do not see distortions arising from including out-rounds in the Elo calculation.
8. Speakers will fall off[3] the *public* Elo rating list if they
    1. Finish university education
    2. Are inactive for 1 year[4]
    3. Indicate that they intend to cease competitive debating
    4. Otherwise do not wish to be included on the rating list
9. In principle the Elo rating system can be extended to include all debating tournaments. Practical concerns *might* dictate that only relatively major tournaments are included in the system, although the system should not be excessively difficult to put into place.

The Elo rating system has been implemented in chess, basketball, and Major League Baseball. An instant-update Elo ranking of all professional chess players with Elo ratings of 2700 and above can be found here, and might illustrate what an Elo ranking system if implemented for debating might look like: http://www.2700chess.com/[5]

In the Section 1 the Elo mechanism is explained in detail and illustrated with a hypothetical example. In Section 2 we point out some of the benefits that implementing the Elo rating system might have. In Section 3 we illustrate Elo implementation by running the Elo mechanism for Zagreb EUDC 2014, and make some brief comments on the results. In Section 4 we briefly list some possible further avenues of exploration with regards to Elo implementation.

Section 1: mathematical outline and hypothetical example

The Elo rating system adjusts a debater's score after every debate based on how their team's performance compares to that implied by the difference between their score and those of the other debaters in the room. If a debater exceeds that expectation, their score moves up. If they underperform, their score is reduced. Given a large enough sample of debates, the implied probabilities of victory will approach the actual probabilities, given that they are frequently adjusted to reflect speakers' performances.

The Elo rating system treats each four-team debate as a series of six[6] pairwise matchups between the four teams. If team A ranks above team B, team A is treated as winning against team B and vice versa. There is no additional adjustment for beating another team by more than one place in the final ranking: if team A also beat team C they would receive credit for that *independently*.

To understand how the adjustment process works, consider the following scenario:

– Two teams: team 1, debaters A and B; team 2, debaters C and D.

– ELO ratings of A, B, C and D are R_A, R_B, R_c and R_D, respectively.

– Team 1 beats team 2.

First, we calculate the team rankings of teams 1 and 2, T_1 and T_2, respectively, namely the linear average of the individual ranking of the two players:

$$T_1 = \frac{R_A + R_B}{2}$$

$$T_2 = \frac{R_C + R_D}{2}$$

This is fairly intuitive – both team members clearly contribute to the overall strength of a pairing. The type of average used is arbitrary. We picked the arithmetic mean because it is simple, but some other, larger average (e.g. a quadratic mean) may be more appropriate, given the propensity of the stronger team member to dominate their combined performance.

The difference between team rankings of 1 and 2 yields the expected probabilities of victory, P_1 and P_2, respectively:

$$P_1 = \frac{10^{T_1 - T_2/400}}{1 + 10^{T_1 - T_2/400}}$$

$$P_2 = \frac{10^{T_2 - T_1/400}}{1 + 10^{T_2 - T_1/400}}$$

This is an intuitive way of deriving probabilities of victory:

– The probabilities sum to 1. This is reassuring; one team should indeed beat the other.

– If the two teams are equally ranked, the probabilities will both be 0.5.

– As $T_1 - T_2$ increases, P_1 tends to (i.e. gets arbitrarily close to) 1 and P_2 tends to 0.

Note that we divide the difference in scores by 400 (the *divisor*) in the probability calculation. The choice of 400 here is arbitrary. Roughly, it determines how much the implied probabilities change given a shift in the score difference – a larger divisor gives rise to smaller change in the probabilities. 400 is the divisor used in chess.

We now adjust the scores based on difference between the expected and actual outcomes of the match. If a team scores 1 point for a victory and 0 points for a loss, we would expect team 1 to win P_1 and team 2 P_2 points in any given match. Given this, we calculate the changes in scores for members of team 1 and team 2, Δ_1 and Δ_2, respectively:

$$\Delta_1 = 32(1 - P_1)$$

$$\Delta_2 = 32(0 - P_2)$$

We are simply multiplying the difference between the expected and actual

outcomes for both teams by 32 (the *K-factor*). The K-factor determines the magnitude of the Elo adjustment. To calculate the new Elo ranking of the various debaters, we simply add the Δ-values for the relevant team to the ratings of each of its constituent debaters. It is worth noting the following:

– The multiplier 32 is arbitrary; it is the maximum number of points a given matchup can move a player's score. So given that each faces three others in any given debate, a single debate can move a player's score by as much as 96 points (though this is practically impossible).

– Since $P_2 = 1 – P_1$,a little algebraic rearrangement will show that team 1's gain is team 2's loss and vice-versa. So, unless new debaters join the system, *points are merely redistributed, not created.*

– We calculate the score adjustment for each pairwise matchup in a given debate before adjusting the scores – so if team 1's scores change by Δ_1 as a result of their beating team 2, we do not add this change on until we have calculated how much they gain or lose from their results against teams 3 and 4.

This following is an example debate that illustrates what an Elo adjustment might look like. The exact Elo ratings of the speakers have been chosen arbitrarily.

– Team 1, pro-am, debaters A and B, rankings 2500 and 1500 respectively.

– Team 2, strong team, debaters C and D, rankings 2200 and 2300 respectively.

– Team 3, intermediate team, debaters E and F, both ranked 1900.

– Team 4, novice team, debaters G and H, rankings 1700 and 1500 respectively.

Suppose team 1 wins, team 2 comes 2^{nd} , team 3 3^{rd} and team 4 comes 4^{th}. We will consider the various pairwise matchups and the adjustments to each of the debater's rankings.

1. 1 vs 2
2. 1 vs 3
3. 1 vs 4

4. 2 vs 3
5. 2 vs 4
6. 3 vs 4

We then add up the relevant Δ-values to obtain the following rankings:

1. 2540.3 (2500 + 40.3)
2. 1540.3 (1500 + 40.3)
3. 2179.7 (2200 – 20.3)
4. 2279.7 (2300 – 20.3)
5. 1891.3 (1900 – 8.7)
6. 1891.3 (1900 – 8.7)
7. 1688.7 (1700 – 11.3)
8. 1488.7 (1500 – 11.3)

Several things should be noted. First: the large increase in the number of Elo points A and B have is due to the fact that it was a pro-am team: the relatively low Elo ranking of B meant that Team 1's ranking was pulled down, and it was hence rewarded more for victory. The second is that the change in the Elo score of Team 4 is small despite its loss: this is because it is a novice team. The third is the fact that despite Team 2 coming second in the debate, it *lost points* overall because it lost more points to Team 1 than it gained from defeating the relatively weak Teams 3 and 4. A "guaranteed second" does not always gain a strong team points.

Section 2: Why implement ELO?

Before we discuss the positive reasons to implement the Elo rating system we would like to point out that the Elo system does not require much more information than is currently captured and publicly shown in tournament tabs. All that is required for the Elo system to work are:

1. Records of the composition of each team. This is currently captured on all tabs.
2. Records of the wins and losses of teams in in-rounds and out-rounds. This is currently captured on all interactive tabs[7], but not non-interactive tabs. Richard Coates, one of the tab engineers of the Oxford and Cambridge IVs 2014 and EUDC 2014, is currently developing an online central database that would capture all the information needed for the Elo rating system to work

across multiple tournaments. The Elo system would effectively require the use of interactive tabs across most tournaments.

Performance across time

The first benefit of an Elo rating system is that it allows for the accurate tracking of performance across time. This is currently very difficult to do. Looking at speaker and team tabs across different tournaments is a helpful guide, but team tabs not take into account the varying strengths of the field at tournaments, as do rankings on speaker tabs. Speaker score averages are problematic as judges in different regions, circuits and tournaments might have different scoring standards. A novice speaker might fail to break at three tournaments at a row even though that speaker might be consistently improving; the Elo system would allow for this speaker to observe real improvements in performance and encourage the novice to continue speaking. Often the illusion of stagnation is discouraging to novices. Conversely, a speaker might break top at a tournament and fail to break at another despite performing equally strongly. It would be helpful to have a metric which can detect improvement or consistency in these two cases.

Another time-based issue arises when a "snapshot" of a speaker's strength is used as a proxy for strength *over* a certain period of time, even though that "snapshot" is not representative. For instance, the person who is tops the speaker tab at the WUDC is often called the "world's best speaker" or "World No.1" for a period of a year, even though that speaker's strength will fluctuate over the course of a year. The rankings generated by the Elo system will probably be a lot more generous to a larger number of speakers; we might see, for instance, several speakers continuously jostle for the no.1 ranking. We might speak of "so-and so being the best speaker from June to October", for instance, which would be a more accurate way of capturing global rank and performance.

Note that speaker tabs and the Elo rating system measure different things. First, performance on speaker tabs is based on the *numerical score* a judge rewards in the round (an *absolute* measure), while the Elo rating is based purely on *relative* performance. Speaker tabs account for the fact that one might have won against a strong team in a terrible debate (which means speakers get low speaker scores despite a "good" *relative* performance), while the Elo rating system cannot. Second, speaker tabs are in some sense more fine-grained than the Elo rating system, since they account for variation *within* teams. Third, the Elo rating system does not take into account margins of

victory, and so a 1-point and a 20-point win are treated the same, while speaker tabs capture (albeit indirectly) such margins. We highlight these factors to point out that the Elo rating system cannot claim to replace speaker tabs, which will continue to remain important.

Comparisons with speakers against whom one has not competed

The second benefit of the Elo rating system is that it allows for comparison with speakers against whom one has not competed; more precisely, it allows for strength comparisons of speakers across circuits. Currently there is no reliable way of telling if a speaker in one regional circuit is stronger than a speaker in another regional circuit. Educated guesses are always possible but are imprecise. It is plausible that a speaker who dominates a particular circuit is not in fact performing particularly well; or that a speaker who is not doing particularly well in a circuit is in fact performing very well relative to the rest of the world. The Elo rating system helps to clear away some of this uncertainty.

Of course, if different circuits had no contact with each other at all the Elo rating system would not be able to provide these comparisons, since the "Elo pools" of each circuit would be closed and Elo points could not be stolen by or from other circuits. This would mean that the weakening or strengthening of a circuit relative to the rest of the world would not be detectable. This concern can be addressed. Regional competitions such as the EUDC, Sydney Mini, ABP, and the US BP Nationals provide one valuable place for pools to mix. The most important competition from the perspective of getting accurate comparison across regions is WUDC, since representatives from all debating circuits will be present, and will determine (together with the number of individuals beginning to debate) the size of their circuit's Elo pool for the rest of the year.

Consider two speakers A and B in two circuits X and Y respectively, both of whom have never participated in the same tournament. Currently it is very difficult for A and B to compare their debating strength. However, circuit X and Y both send (their strongest) teams to WUDC. If circuit X happens to be strong relative to circuit Y, then its teams will increase the size of circuit X's Elo pool relative to circuit Y (by winning more debates than circuit Y's teams at WUDC). If A and B perform roughly equally against teams in their own circuits, it is then likely that A will have a greater number of Elo points than B, since more Elo points collected from WUDC will diffuse into circuit X than circuit Y. A might then be able to say with a reasonable degree of confidence that he/she is

a stronger debater than B.

No problem arises even when a circuit's WUDC teams are highly unrepresentative of the quality of the circuit in general. If the WUDC teams are particularly strong, then they are also unlikely to have the Elo points they gained at WUDC stolen from then by other teams in their circuit. The circuit's Elo pool increases, but the Elo points are also more tightly locked up in a few teams. The converse logic applies where the WUDC teams are particularly weak.

Large-scale comparisons

The third benefit of the Elo rating system is that it allows for certain large-scale comparisons to be easily made. One has been mentioned to above – the relative strength of different circuits. However, Elo ratings could also be helpful in detecting bias in circuits towards or against certain genders or races. If circuit X has a large number of female speakers in its regional top 20 ranking and circuit Y has a small number of female speakers in its regional top 20 ranking (controlling for factors like the participation rates of people with different sexual orientations), this suggests that circuit Y might have a bias against female speakers. More prosaically if, say, half of the debaters in a circuit are female but none of them are ranked in that circuit's top 20 speakers, something is probably wrong. Thus, the Elo rating system is of interest not just to individual speakers who want to become better debaters, but to tournament organisers and bodies like the WUDC Council that have a general interest in making debating fair and inclusive.

Determining tournament/room strength

The fourth benefit of the Elo rating system is that it allows for accurate categorization of tournament strength. For example, for the purposes of novice competitions or pro-arms, we currently determine who an "am" or "novice" is in debating by reference to how many university-level tournaments they have broken in. It might be worth considering *broadening* the definition of "am" or "novice" to include individuals whose Elo ratings fall below a certain number. A person could have debated for a long time and still benefit hugely from being partnered with a strong debater. The Elo rating system would also let us determine what the overall strength of a tournament (or room) is by simply obtaining the average Elo rating of the relevant speakers. Universities deciding which tournaments to send their teams to might find objective measurements

of tournaments' strength useful. Furthermore, knowing the strength of a particular room in a competition might aid CA teams in judge allocation; they might want to put the best judges in rooms that fall within a certain Elo bracket, for instance.

Including out-rounds

The fifth benefit of the Elo rating system is that it allows us to integrate performance over in- and out-rounds in a single measurement. This means that Elo ratings capture more information about team performance than team tabs do. A team that progresses from the quarter-finals of a tournament to the semi-finals must beat the two teams that do not progress from the quarter-finals. Thus, it steals points from two teams but, assuming that the judges did not come to a comprehensive team ranking, should neither steal points from or lose points to the team that progresses through the out-round with it. Loosely speaking, we might say that a team that progresses through an out-round takes a "1.5" ranking, while teams that do not progress take a "3.5" ranking. This makes sense; half of the teams that progress come 1^{st}, and half 2^{nd}, and teams that do not progress come 3^{rd} and 4^{th} half of the time respectively. Of course, these assumptions do not hold true for particular teams; note, however, that *including* this data is certainly less distortionary than *excluding* it altogether, since we are *certain* that each team has won/lost against two other teams, and these wins are just as valid as wins against any other team in an in-round.

Estimating individual tournament performance

The sixth benefit of the Elo rating system is that it allows for a speaker to estimate to a reasonable degree their *performance rating*. The performance rating measures the strength of performance at only one tournament; knowing his/her own performance ratings for each tournament would allow a speaker to know which tournament represented their strongest or weakest performance in terms of debating strength, without distortions relating to the strength of the tournament field. It might also allow us to determine the *strongest tournament performance by any person recorded in a certain period* – a person might not win a tournament, but still be responsible for a stunning performance overall. One way of estimating a speaker's performance rating for a tournament[8] is to:

1. Take the rating of each team beaten and adding 400;
2. Take the rating of each team lost to and subtracting 400;

3. Sum the figures obtained; and
4. Divide by the number of debates multiplied by three (the number of teams debated against)

A possible advantage of being able to calculate performance rating relates to tie-breaks. Ceteris paribus, we want the team with the higher performance rating to break to out-rounds. Current tiebreak measures tend to arbitrarily favour either consistency or variance in performance (e.g., counting wins) or provide only a limited snapshot of the team's performance that overemphasises team-specific interactions (e.g., head-to-head records and tiebreak debates). Performance rating might provide a better measure of overall debating strength, although this requires the Elo rating system to be relatively well-developed (i.e., implemented for a significant period of time) so that team ratings accurately capture team strength.

Encouraging pro-ams

The seventh possible benefit of the Elo rating system relates to pro-ams. It is plausible that strong speakers will see pro-aming as a way to gain rating points, since pro-aming lowers the team rating. Provided that strong speakers believe that they will continue to perform relatively well even when pro-aming, this lowered team rating makes it appear easier for them to gain Elo points from wins. Of course, this effect is not at all a mathematical certainty – the fact that we employ *team ratings* when determining the size of the Elo point transfer ought to mean that a strong speaker is neither punished or rewarded when speaking with a novice – but our experience indicates that it is at least plausible that strong speakers perform very well (i.e., not significantly worse than if they were not speaking with a novice) when speaking with novices.

We do not believe that the Elo rating system will be particularly humiliating or off-putting for individuals with low Elo ratings. We should first note that there is no reason to believe that Elo ratings are more embarrassing than speaker tabs, which already list all individuals from best to worst regardless of language category. Being part of the debating community appears to already involve being willing to publicly share one's successes and failures, as in any other competitive activity. The Elo rating system finesses information that is already available in the form of interactive tabs. We also note that, in relation to speaker tabs at large tournaments, individuals tend to be interested only in (1) their own ranking; (2) the rankings of individuals they know personally, and (3) the top 20 speakers. We no reason to believe things will be different in

relation to Elo rankings. This means that a speaker who is world no.255, for instance, has absolutely nothing to feel ashamed or worry about. If this is in fact a problem, however, the solution would be to only publicly display the Elo ratings of the world's top 100 speakers. Furthermore, we have reason to believe that Elo ratings might be especially encouraging for novice speakers, who might not see clear indicators of improvement at their first few tournaments if they do not break. And Elo ratings will also tell individuals when they have stagnated so that if they want to they can do something about it.

Does the Elo rating system make debating *too competitive*? This is hard to tell. Some speakers will want to debate more to improve their ranking; others might want to debate less for fear of damaging it. And (we hope) people will continue be motivated to debate or not to debate by factors unrelated to Elo rankings: the general need to live a full life, the desire to see friends (debaters or otherwise), the enjoyment of debates, and the desire to do well. It is hard to imagine the Elo rating system making a huge difference to people's decision-making. What we will know for certain is that people will have more information upon which to base their decisions. This is good.

Section 3: Zagreb EUDC 2014: what would Elo look like?

We assigned speaker who participated in Zagreb EUDC 1500 points. Therefore each team started the tournament with a rating of 1500. We calculated Elo ratings both after the in-rounds, and after the entire tournament. The top 50 teams were ranked according to their post-EUDC Elo ratings. Since team and speaker ratings are identical (given that everyone began with the same Elo rating) we do not explicitly consider individual ratings.

Several things should be noted:

1. Hebrew A broke into both the Open and ESL out-rounds. Since it debated in the Open out-rounds first, its out-round-inclusive Elo was calculated by making the relevant Elo adjustment from the Open quarter-final before making the adjustments from the ESL quarter-final and semi-final.
2. Since everyone started the tournament with 1500 Elo points, the post-tournament Elo rankings also function as a measure of tournament performance strength.
3. The relevant calculations were not particularly difficult to carry out. Once

the Elo formula was provided, the relevant coding for Tabbie took less than 1 hour to complete, although several corrections had to be made later. We estimate that, if told in advance, individuals familiar with Tabbie will be able to perform the relevant Elo calculations for a tournament in less than 30 minutes, assuming that the relevant coding has been completed. The relevant data input and calculations were made easier for this EUDC illustration by the fact that all teams and individuals started off with the same rating, but we do not believe that obtaining speakers' Elo ratings pre-tournament will be difficult. Obtaining Elo ratings can be integrated into current tournament registration procedures. If there is a central database that immediately updates and stores Elo ratings, this can be consulted. For individuals who wish to write programs that calculate Elo ratings, note that:

1. Elo point transfers *in each debate* must be calculated *independently*. Thus, the team that takes a 1st *does not* have its Elo rating adjusted after the size of the point transfer from *one other team* has been calculated: all the Δ-values for all teams must be added up before the point transfer is made. (See the hypothetical example provided in Section 2.)
2. If a team's rating changes by X over the course of a tournament, then each speaker will also have his/her Elo rating change by X.
3. In an in-round, a team's rating can change by a maximum of 96 Elo; in out-rounds, 64 Elo.

EUDC 2014 Elo ratings (top 50)

Elo (Final)	**Elo (after in-rounds)**	**Team Tab**
1. SHEFFIELD A 1791	1. CAMBRIDGE C 1779	1. CAMBRIDGE C
2. OXFORD A 1767	2. OXFORD B 1769	2. OXFORD B
3. OXFORD B 1754	3. GUU A 1737	3. CAMBRIDGE B
4. CAMBRIDGE A 1748	4. CAMBRIDGE B 1734	4. OXFORD A
5. BELGRADE B 1738	5. OXFORD A 1732	5. GUU A
6. CAMBRIDGE C 1736	6. CAMBRIDGE A 1701	6. CAMBRIDGE A
7. EDINBURGH A 1723	7. OXFORD C 1701	7. OXFORD C
8. GUU A 1701	8. DURHAM B 1681	8. EDINBURGH A
9. CAMBRIDGE B 1697	9. LSE A 1673	9. LSE A
10. BERLIN A 1683	10. EDINBURGH A 1673	10. SHEFFIELD A
11. LUND A 1677	11. NOTTINGHAM A 1672	11. DURHAM B
12. NOTTINGHAM A 1676	12. SHEFFIELD A 1671	12. KCL A
13. KCL A 1675	13. KCL A 1671	13. HEBREW A
		14. NOTTINGHAM A

14. OXFORD C 1670
15. BPP A 1653
16. DURHAM B 1649
17. DURHAM A 1646
18. DURHAM C 1641
19. UCD L&H A 1640
20. TCD PHIL A 1640
21. LSE A 1639
22. BIRMINGHAM A 1638
23. WARWICK B 1638
24. WARWICK A 1638
25. HEBREW A 1623
26. TARTU A 1623
27. MANCHESTER A 1610
28. SOAS A 1608
29. ABERYSTWYTH A 1608
30. BGU A 1608
31. GUU B 1607
32. UCD L&H C 1607
33. UCC PHIL A 1606
34. BBU A 1606
35. LEIDEN A 1604
36. TEL AVIV B 1603
37. BELGRADE A 1581
38. HULL A 1581
39. ELTE A 1579
40. TCD HIST B 1579
41. IMPERIAL B 1578
42. TILBURY H A 1578
43. LSE B 1577
44. WARSAW A 1577
45. BRISTOL B 1577
46. UCC LAW B 1577
47. TCD HIST A 1576
48. STRATHCLYDE A 1576
49. ULU C 1576

14. DURHAM C 1669
15. HEBREW A 1668
16. DURHAM A 1646
17. BELGRADE B 1645
18. BERLIN A 1644
19. BPP A 1642
20. TCD PHIL A 1640
21. UCD L&H A 1640
22. TEL AVIV B 1639
23. WARWICK B 1638
24. WARWICK A 1638
25. BIRMINGHAM A 1638
26. TARTU A 1638
27. LUND A 1636
28. BUCHAREST A 1610
29. MANCHESTER A 1610
30. TILBURY HOUSE A 1609
31. ELTE A 1608
32. SOAS A 1608
33. BGU A 1608
34. ABERYSTWYTH A 1608
35. UCD L&H C 1607
36. MANNHEIM A 1607
37. GUU B 1607
38. BBU A 1606
39. UCC PHIL A 1606
40. LEIDEN A 1602
41. HULL A 1581
42. TCD HIST B 1579
43. IMPERIAL B 1578
44. LSE B 1577
45. UCC LAW B 1577
46. BRISTOL B 1577
47. WARSAW A 1577
48. TCD HIST A 1576
49. STRATHCLYDE A

15. DURHAM C
16. BPP A
17. BELGRADE B
18. LUND A
19. UCD L&H A
20. TCD PHIL A
21. DURHAM A
22. TARTU A
23. WARWICK A
24. BERLIN A
25. WARWICK B
26. BIRMINGHAM A
27. TEL AVIV B
28. LEIDEN A
29. BUCHAREST A
30. SOAS A
31. UCC PHIL A
32. UCD L&H C
33. GUU B
34. ABERYSTWYTH A
35. BBU A
36. TILBURY HOUSE A
37. MANNHEIM A
38. MANCHESTER A
39. ELTE A
40. BGU A
41. TCD HIST B
42. TCD HIST A
43. STRATHCLYDE A
44. LSE B
45. GUU C
46. HULL A
47. LANCASTER A
48. BRISTOL B
49. ULU C
50. STRATHCLYDE B

50. LANCASTER A 1576	1576 50. LANCASTER A 1576	

Several things should be noted:

1. The changes in Elo rating are relatively large, often approaching 300 points. This is because many speakers began with a score (1500) that was highly unlikely to represent their debating strength, and because EUDC is a large tournament where each team must debate against at least 27 others. There are hence at least 9 ratings adjustments, each with a hypothetical maximum size of 96 Elo points, to be made.
2. The ranking according to in-round team ratings corresponds fairly well the team tab, with some minor divergences (see Durham B, Nottingham A, Elte A, and Lund A, for example.) This is unsurprising, given that the EUDC has (1) a relatively large number of in-rounds and (2) employs power-pairing. Less correspondence will tend to be seen in smaller tournaments.
3. The out-rounds have a significant impact on Elo rating. Sheffield A, ranked equal 12th based on Elo after the in-rounds, gains 120 Elo points by defeating 7 strong teams in the out-rounds to come 1st in the final Elo rankings and very close to crossing the 1800 mark. Belgrade B also moves from 17th to 5th position in this manner.
4. Even though EUDC 2014 is a large tournament, it is unclear if the Elo rankings above are representative of the speakers' relative strength; more time might be needed for estimated and actual performance to match and for Elo ratings to stabilise. We did not calculate Elo ratings on a round-by-round basis, and so do not know if Elo rankings stabilised before Round 9. For teams at the upper and lower ends of the Elo ranking, we suspect that this is unlikely to be the case.

Section 4: Further issues for consideration

Issues that we have not had time or space to discuss but which are relevant and might merit exploration include:

1. Modifying any one of the arbitrary parameters used in our Elo calculation, such as the initial number of Elo points (1500), or the size of the divisor in the probability calculation (400).
2. Using the geometric rather than arithmetic mean to determine team ratings.
3. Specific K-factor issues:

    1. Having higher K-factors for tournaments deemed to be important.
    2. Having lower (or higher) K-factors for out-rounds.
    3. It might be especially useful to have a K-factor that starts out large but shrinks down to a minimum value over time, to ensure that people can rapidly move towards their representative Elo rating from the initial 1500. A simple formula for achieving this might be to have a K-factor of: 500/(number of debates), with a minimum K-factor of 32. This drastically reduces the time it takes to move away from the 1500 rating, since the first few (rated) debates will have a very large impact.
    4. Having a rating-staggered K-factor. E.g.: a K-factor of 32 for ratings between 1200 and 1600, 24 for ratings between 1600 and 2000, and 16 for ratings above 2000.

4. Excluding certain tournaments from Elo calculations.
5. Implementing (separate) Elo ratings for non-BP debating formats, with which we are not intimately familiar. We note that the relevant calculations ought to be simpler where debates only include 2 teams.
6. Integrating the Elo ratings for BP and non-BP formats. This is worth serious consideration, since debaters in the Australian and Asian circuits debate mostly in the Australs and Asians formats. Implementing Elo ratings only for BP debating means that (1) these debaters have few chances to have their Elo rating adjusted, sometimes as few as 3 a year, and that (2) both UADC and Australs are excluded from Elo calculations. Separate Elo ratings might be necessary if the Australs and Asians formats are considered too different from the BP format for a single Elo rating to make sense. Since we are not intimately familiar with the Australs/Asians formats, however, we do not take a stand on this issue.
7. Using Elo as an aid in team allocations for WUDC. Given that the demand for WUDC spots appears to be growing faster than WUDC can accommodate it, Elo ratings might be useful in determining which one among two institutions gets, say, a 3rd team for WUDC. We might wish to give the spot to the team with the higher Elo rating. Of course, this assumes a certain set of aims of the WUDC, and we do not take a stand in this article on this issue.

1. We would like to express our gratitude to the many individuals who discussed our proposal with us and provided us with important insights and suggestions. ↩
2. This has important implications for teams where the Elo ratings of the speakers differ significantly. See the discussion of pro-ams in Section 2. ↩
3. Note that this might appear to pose a problem for tournaments which

include speakers who are not on the Elo rating list. However, the solution is so *calculate and store* the Elo ratings of individuals *not* on the list, without publicly revealing their Elo ratings. So speakers can still gain or lose points fairly from open tournaments by debating against "retired" debaters whose Elo ratings will be resurrected for the purposes of making the relevant calculations. ↩

4. This period can be modified to reflect different levels of BP debating activity in different circuits. 6 months might be more suitable for the IONA circuit, for instance. ↩
5. The site is maintained by chess enthusiasts and shows: (1) Elo ratings; (2) world rankings; (3) recent changes in Elo rating and ranking; (4) recent games played; and (5) progress charts over time for each player. ↩
6. 4!/2!(4-2)! ↩
7. This is the interactive tab for the Cambridge IV 2014: http://www.tabbieballots.com/tabs/cambiv2014/teamtab.html. Clicking on a team's name shows its win-loss record. ↩
8. Used by some chess clubs. ↩

December 23, 2014 Ashish Kumar, Michael Goekjian and Richard Coates Volume 12. 2014 Changing Landscapes: Proposals for Debating Reform

Proudly powered by WordPress

Monash Debating Review

An annual publication of the Monash Association of Debaters

Text Tab

Debate tournaments, especially big ones, can be messy events, when it comes to disbursement of information, be it the Tab or Motion for the round, or other major announcements. Especially with all the intense pre-debate prep demanding people's attention.

In many instances debaters find it difficult to properly note down their rooms, names of other teams and correct wording of the motion. Having information slides exacerbates the problem, especially if any word requires further clarification. Moreover, multiple running of the draw contributes to significant delays as well. Projector displays and oral announcements require the audience to be attentive all the time, which can be both inconvenient and stressful, and we always run the risk of someone missing out on an important announcement for any number of reasons, especially towards the end of the day when there is a rush to get back to the buses on time.

A recent development in the Bangladesh debate circuit has tackled these issues with ingenuity & pragmatism.

A project consisting of five developers, headed by Nazmus Sakib from Islamic University of Technology, Textab has been quite a success in its short stint in the local circuit, playing a substantial role in terms of distributing essential information to the relevant parties & ensuring the overall efficiency of the tournament.

TexTab ensures that each individual receives the information they need to know at the right time, and in case of announcements, allows them to stretch their memory to events that happened before that late night party started. Most importantly, it allows adjudication core and organizational committees greater flexibility at making adjustments to the schedule as the tournament progresses, since there remain no worries about people missing out on an announcement.

So how does it work?

TexTab is a personalized SMS-integrated debate tabulation system. When deployed, it sends text messages to each participant's mobile phone via bundle SMS's on their carrier. Currently the system sends three types of information – the draw for each round (Image 1), the motion for each round and any accompanying information slides (Image 2), and alerts (reporting notice and as such, Image 3). The system also offers custom masks, which means the name of the tournament can be the SMS transmitter's ID.

In terms of compatibility the system operates seamlessly with Tournaman and Tabbie. The team is adept at running BP tournaments. However, the dearth of Asians or Australs format tournaments in this part of the world meant the development for 3v3 formats are still in their initial stages.

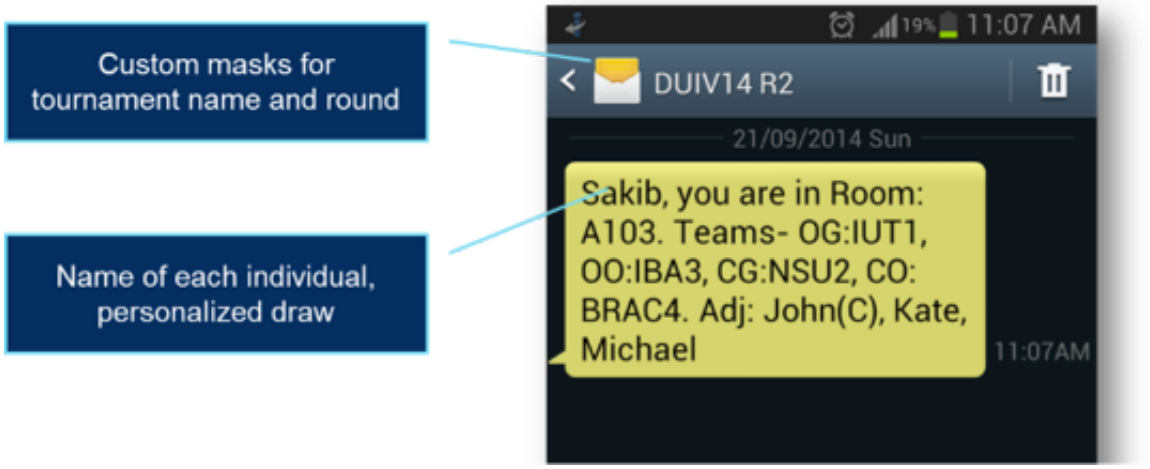


Image 1: Sample SMS – Personalized draw for debater

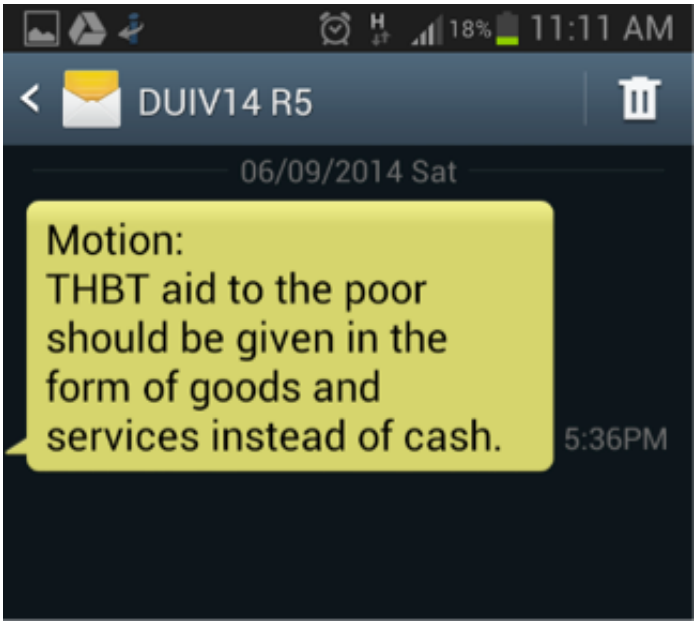


Image 2: Sample SMS – Motion of specific round

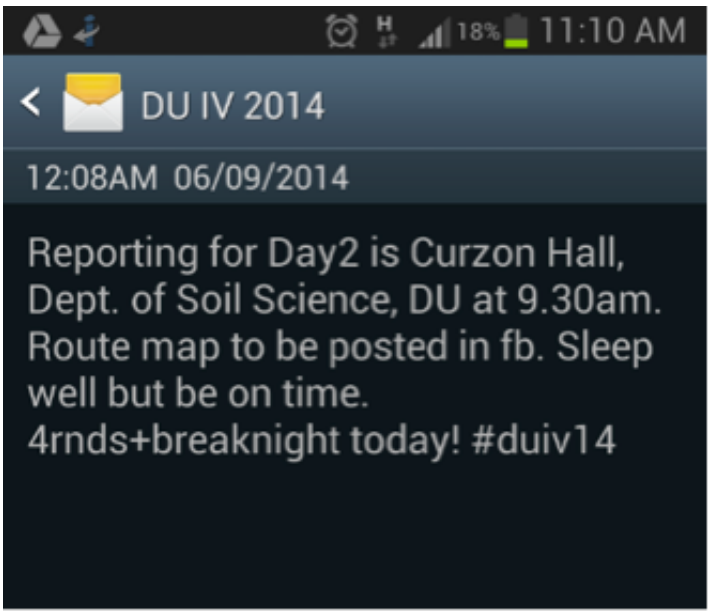
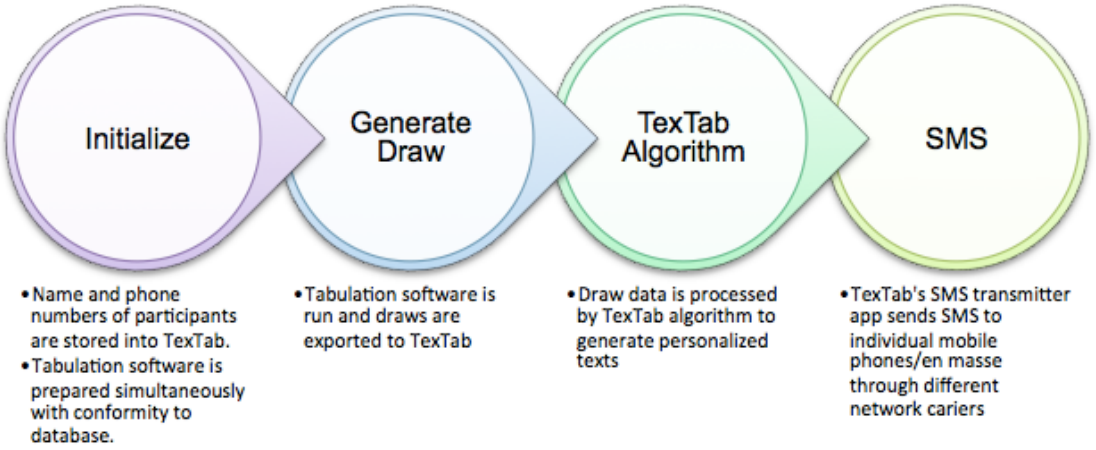


Image 3: Sample SMS – Alert for reporting

About the operations

The system consolidates three programs into one platform – the tabulation software, the TexTab data processor and the TexTab transmitter application. For the draw SMS, all three work together.

For motion and alert SMS, only the transmitter app is required. The steps are outlined in the illustration below:



For an example of time involved: In a 40-team tournament, time required between end of draw generation and SMS received in phones is 12 minutes on average.

Once recruited for a typical tournament, the TexTab team sets up the database and tab room beforehand. One assignment team consists of a Tab Director and two TexTab executives. The team takes care of all tabulation-related tasks of a

tournament – from briefing and directing runners, coordinating with organizers and adjudication core and providing tab feedback, break summary and tabulation reports in the end. The approach is to allow the org.com to be able to completely outsource anything regarding tabulation to the team, and focus on more important things, like managing food & drinks, ensuring the debates run on time. It also reduces the burden on the organizational comities in terms of training the volunteers specifically for ballot and feedback form related tasks as well.

Other platforms:

There are certain considerations in juxtaposing SMS with WiFi, such as – reliable internet connection and availability, costs, SIM cards and need for developing iOS and Android apps. There are plans to expand the system and cover alternative communication modes and platforms in future, with particular focus on WiFi based app in the works.

Challenges

Still in its infancy of both development and operation, the software does have its hitches, however minor. Most of the issues are purely logistical. In some instances, the SMS went to the wrong person due to incorrect entry of the person's mobile number, or that individual having swapped the team combinations from the ones originally submitted.

In a few other instances, SMS to some recipients were delayed by up to 15-20 minutes; this author received a long information slide on a gas deal during the Prime Minister Speech (Much obliged).

However, these are purely down to the connectivity and network quality of the local carriers; and were not faced by people who subscribe to better service packages within the same events.

So far, the system has been used successfully in two major local tournaments, the Dhaka University IV 2014 and the IUT IV 2014, both with more than 11 debate rooms with over 8 rounds of debate each, with no additional issues reported.

In terms of scaling up, the only concern remains for international tournaments due to some participants having limited or no access to a working SIM card.

However, that can be partially circumvented by them submitting one number per team/institute, allowing them access to all the announcements regarding the tournament.

The system has been tested to be reliable with carriers in a number of countries, so this remains a valid option for regional or national level tournaments. Realistically, a WiFi based app will need to be the default platform for the system to fully function on the international platform.
Message from the developers:

"The field of tabulation should evolve with the pace of technology and scale of tournaments. Investment is required to encourage young developers to create new methods in tab system, analysis and service. We designed TexTab as a model that changes the way tabulation is engaged intellectually and incentivizes creativity. As such, the TexTab system and service is a commercial one.

Our tab directors and executives are always excited to volunteer at tournaments where just the tabulation software is required, at no-cost. Subscribing to the full package of SMS-integrated TexTab, however, has its costs. We bill a fee which covers service- and SMS-charges. We are looking to try our system on a broader scope, to both cater to the needs of the tournaments, and to learn the specific requirements of participants and tailor our system to suit them better in the future.

We eagerly await feedback from the international community"- Nazmus Sakib, Developer, Textab.

Contact Us

Email: nsakib002@gmail.com

Call: +880-175-553-0753

December 23, 2014 Saad Amjad Volume 12. 2014 Changing Landscapes: Proposals for Debating Reform

Proudly powered by WordPress

Monash Debating Review

An annual publication of the Monash Association of Debaters

Making Judge Feedback More Representative

Good judging is a crucial part of any tournament. There are many skills that a good adjudicator should have. In general we say a good judge is able to accurately understand and describe the debate as it happened, objectively evaluate and comparatively weight contributions of each of the teams and is capable of participating constructively in a panel discussion while also allowing other judges to voice their views. It is difficult for CA-teams to know how good someone is at all of these different skills. Feedback on judges (teams on chairs, chairs on wings and wings on chairs) is one of the only ways to assess these attributes and help determine the quality of a judge. That makes feedback an essential tool in the debating community to further the overall quality of judging at our competitions.

Last summer, the European Universities Debating Championships took place in Zagreb from August 18 to August 23. During this yearly event (with 9 preliminary rounds and a break to Quarterfinals for both ESL as well as Open teams), the CA-team and the Tab-team put in place a feedback system to be able to evaluate judges. Every open round, teams could give feedback on their chair judges (through a virtual or physical form). In all of the rounds, chairs gave feedback on their wings and wings on their chairs. This led to 1777 pieces of feedback that were submitted to the tabroom. In this article we (Maja Cimerman & Tomas Beerthuis; DCAs and Calum Worsley; Tab Co-ordinator) would like to share with you the things we found and what we've learnt from that. This way we hope to make feedback in the debating community more effective and through that, help improve the quality of judging.

What did we do with this feedback?

Let's start by saying that every piece of feedback was looked at by a member of the CA-team. We can assure you that we were very much sleep deprived, but

also that this helped us tremendously in determining how judges were performing at our competition. Feedback at Euros worked in the following way:

– Every piece of feedback was submitted to our system. In this system we could look at the scores on a set of determinants for every individual judge for each round. This allowed us to establish whether the ranking we had allocated to a judge was consistent with the score or if the former needed to be raised/lowered. By that we mean if a judge received very poor feedback when chairing, then this would be a reason to make this person a wing and look at their judging with more scrutiny.

– Next to that, we closely inspected very high and very low ratings every round, to understand the extreme cases (and take appropriate action where necessary).

– We also inspected comments closely, to ensure we learned more about our judges (particularly those that none of us knew from previous competitions).

– Every round, 2 members of the CA-team would 'sit-out' (not judge) in order to look at feedback and determine if the rankings of judges needed to be changed.

Looking at so much data and especially putting it all together and analysing it after the tournament gave us some insights into how people give feedback and how useful feedback is at (large) competitions. We found a number of things that are valuable to share and may help to improve the quality of feedback for future competitions.

Finding #1: People do not use the scale

For every question and irrespective of the specific content asked, respondents could choose from a 1 to 5 scale (with 1 being the lowest score and 5 being the highest score). Looking at the results of our feedback forms, we realised 5 was a disproportionately popular answer across all questions asked, indicating that people start their evaluation at 5 and work down from there (see Graph 1). At best this kind of scale can tell us something about judges that people are really dissatisfied with, but fails to differentiate among good judges, meaning it has little value at determining the judges who should break. Thus any judge of average quality would receive a 5, but an absolute top judge would also receive

a 5. On the other side of the spectrum we can interpret 1's as judges people are really dissatisfied with, but it is not clear what 2's, 3's and 4's are. While it might be that some respondents use the full scale, the fact that it is not used equally across all the respondent skews the results. This makes it very hard to determine the relative difference between judges, apart from the extremes. And even with the extremes, people tend to go to the '1' very quickly (perhaps also out of resentment sometimes), while that may not be an accurate reflection on the person judging.

To address this, we propose rethinking how we define the answer scale, making 3 the response that would be expected most frequently and also closest to the average response. This seems more logical, because it allows CA-teams to better understand differences between judges. 3 would be a rank you would give to most judges that perform as expected, indicating the judge was solid. 5 would be the rank for an exceptional judge and 1 would be the rank for a judge you would be *really* dissatisfied with. While this might require a bit of redefining how we think about judges (mental shift from awarding a good judge a 3 and not a 5), it is actually something we already, very successfully, do with speaker points where the distribution is very close to a normal distribution.

To implement such a change 2 things need to be done:

1. The feedback scale should be revised and explicitly included and explained in both speakers and judges briefings. Raising more awareness with participants on how to use the system will help contribute to making this mental shift.
2. The scale on the feedback forms should be adjusted to reflect the discussion. This is an on-going process, and different scales might be used, [1], but the authors of this article are most fond of keeping the 1-5 scale, while adding a description of each of the values rather than focusing on the number. Obviously this would depend on the question, but we see it as something like:

How well did the judge explain the reasoning of the decision?

> **[] Poor performance** *(Poor explanation of the debate. Did not agree with their reasoning of the ranking at all.)*
>
> **[] Acceptable** *(Somewhat acceptable reasoning explaining their decision.*

Was not fully convinced by their explanation of the ranking.)

[] Meets expectations *(Good reasoning explaining their point of view. I could see and understand why they decided as they did.)*

[] Exceeds expectations *(Great reasoning explaining their point of view. I was convinced that was the correct reading of the debate.)*

[] Top performance *(Excellent explanation of the debate. Not only did I fully agree with their explanation, it gave me new insight in the debate.)*

Although the system would still capture 'Poor performance' as a 1, this way of framing feedback would trigger people to think in a more nuanced way about the actual performance of a judge rather than thinking about a number. Sometimes there is a tendency for people to give a 5 when they are satisfied, but that doesn't always adequately capture the performance of the judge. This is a way to make feedback more consistent across the board and give the CA teams more useful information on the quality of judges.

The same descriptive scale can be applied to the majority of other questions as well by simply reformulating their grammatical structure, while keeping the same content of the questions. For example the current question "Was this person active in the discussion?" could be changed to: "How helpful was this person in the discussion (for reaching the final decision)?". Along with the structure of the questions, obviously the answers would be changed as well, where the answer on number 3 would be the one we expect to be the most common or average. For the specific example above:

How helpful was this person in the discussion (for reaching the final decision)?

[] Poor performance *(Mostly disruptive or not involved at all.)*

[] Acceptable *(Only somewhat helpful and/or barely involved.)*

[] Meets expectations *(Helpful and active in the discussion.)*

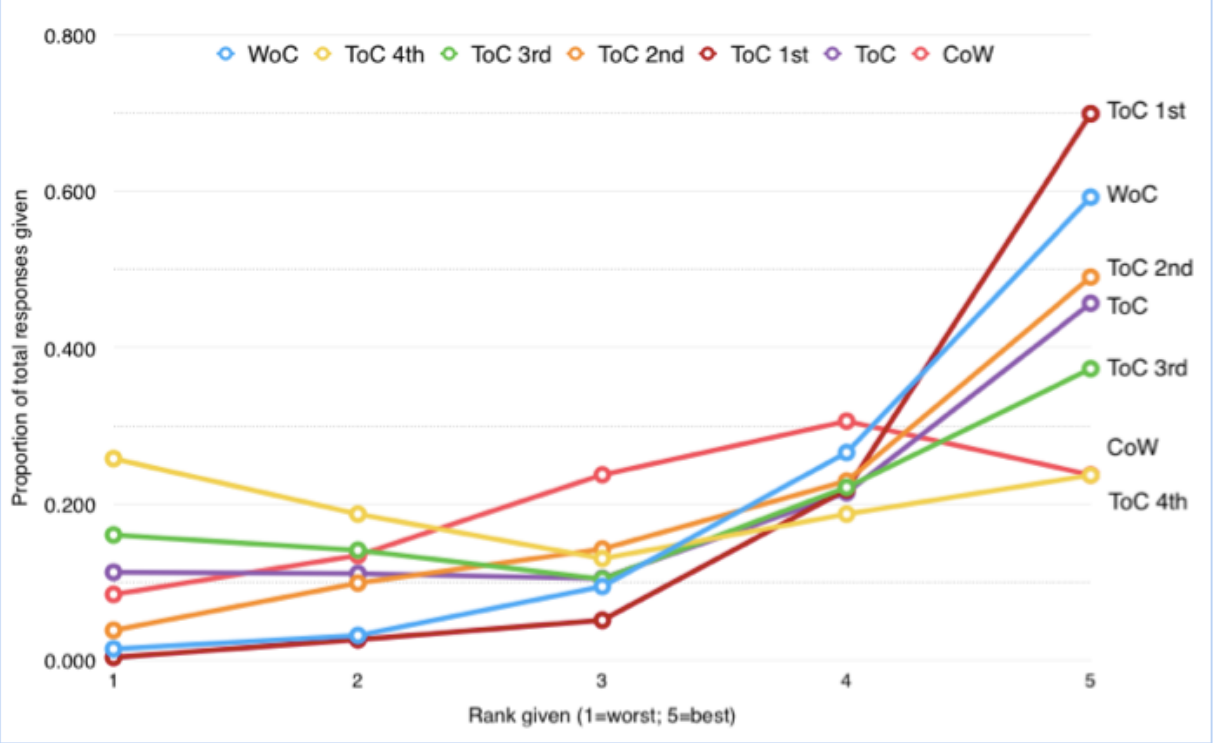
[] Exceeds expectations *(Very good contribution to the discussion, all relevant and excellent.)*

[]Top performance *(Great contribution, changed some of my views of the*

debate.)

Finding #2: Your ranking in a debate determines what kind of feedback you are going to give

For a community that prides itself for reasoning and critical thinking, it is interesting to see the role emotions play when giving feedback. More specifically, data shows (see Graph 1) 1st placed teams give feedback which almost exclusively evaluates judges positively, 2nd placed teams are a bit more critical of their judges, 3rd placed even more and 4th placed teams are most likely to give judges bad feedback (the only group where "1" was the most common answer). This might be unsurprising, given that worst placed team were probably least happy with the outcome of the adjudication and best ranked teams were the happiest, however it also means this kind of feedback tells us little about the actual quality of the judge.



Graph 1: Frequency of responses on a scale 1-5 for judging evaluating questionnaires, based on different answering groups. [CoW = Chair on Wing, WoC= Wing on Chair, ToC = Team on Chair, ToC 1st = 1st ranked Team on Chair, ToC 2nd = 2nd ranked Team of Chair, ToC 3rd = 3rd ranked Team on Chair, ToC 4th = 4th ranked Team on Chair]

We already control for team's position when weighing their feedback, and in the feedback module the team feedback always comes with the position the team took in that round next to the scores for CA team's information. This data

possibly calls for even greater consideration of a team's position in determining the value of the feedback they give us. For instance, a first ranked team delivering horrible feedback on a judge necessitates greater CA's consideration than a first ranked team praising the judge.

However, adjusting the weight of feedback based on ranking will not contribute gravely to tackling the real problem – on average, when teams win, they applaud their judge and when they lose they punish the judge with bad feedback. This is something that needs to be seriously discussed and considered within the community (and possibly even having a debaters' briefing to flag out the role their emotions play so they might be more vigilant about them), otherwise there is little value in reading, triaging and entering the feedback we get from teams. Although emotions in debating competitions are normal, we should realize that this (currently) is seriously affecting what kind of feedback people give their judges. We should also realize that complex debates (with sometimes unsatisfying outcomes) may further trigger this effect. All of this distorts the credibility of feedback and makes it more difficult to evaluate the performance of judges. In turn, this makes it more difficult for CA-teams to adjust the rank of a judge appropriately, which again has an effect on the quality of judging at the competition.

Some other comments

We would also like to add some pragmatic issues of incorporating feedback in judges evaluation, which do not stem from empirical analysis of feedback rather they reflect issues we stumbled upon when looking at feedback.

a. In retrospect, we found the questions to chairs regarding their wings about the participation in the discussion (*Was this person active in the discussion?*) less useful, as a wing judge might get 1 on all other questions and 5 on this question. We believe a better phrasing might be: *How helpful was this person in the discussion?* (Something we have already discussed in Finding #1.) This way we could possibly also scrap the question about how willing they are to compromise (*If you disagreed, did they show willingness to take your view on board?*) and overall reduce the number of questions.

b. In terms of Wings on Chair feedback we realised some wings got confused by the initial call question (*On reflection, do you think this person's initial call was reasonable?*), as some chairs do not disclose

their ranking during the discussion. We propose either scrapping the question or reducing its relative importance.

c. Some things to look out for when interpreting feedback:

Feedback should not be determined only by the aggregate score, we should look at scores for individual questions/rounds and see what these tell us. For example:

1. A fresher that received phenomenal feedback as a wing but terrible feedback as a chair might be a really good judge, but inexperienced or unconfident as a chair. If this person would break as a talent, this could very much contribute to their development, making them a potential chair at a future competition.
2. A chair who consistently scores very low on taking other judges seriously, should probably not be chairing (outrounds), because they will be too dominant in the discussion and thus might stifle it.

Conclusion

Reading and evaluating feedback is time consuming, especially when the aggregate score is insufficient for a holistic evaluations and relevant information needs to be extracted from minor scores and specific answers. This, most times, results in lengthy discussions regarding the merit of a specific feedback, which constitutes too big of a time toll on the CA team in such a fast paced tournament. Thus a different way of doing and interpreting feedback is necessary. Some of the changes we discussed touch on how we ask questions and others touch on a mental shift that is necessary in the debating community to make feedback a little bit more reasonable. This article provides some suggestions on how to do that, however we see it as an ongoing process where the discussions we have within the community will play a crucial role.

1. As for example a 1-9 scale or a Likert scale. ↩

December 23, 2014 Maja Cimerman, Calum Worsley and Tomas Beerthuis Volume 12. 2014 Changing Landscapes: Proposals for Debating Reform

Monash Debating Review

An annual publication of the Monash Association of Debaters

The ‘Fairness Principle’ in Debating

Joshua Taylor and Gemma Buckley

Debating is an inherently arbitrary, and therefore unfair, activity. However, much of the arbitrariness and unfairness is a result of how we currently choose to implement the rules and procedures around the activity. This article will propose a radical shift in the way we go about adjudicating debates, in hopes of achieving fairer outcomes.

Part One: Fairness matters

It might seem redundant to begin this article by establishing that fairness is of the utmost importance in awarding the result of a debate; hopefully all readers agree intuitively that the goal of debating should be to find a fair result. We can probably all acknowledge that individuals spend countless hours, huge sums of money and much emotional turmoil to hone their debating skills in pursuit of success. This reason alone is enough to suggest that we ought to ensure the right, or the most deserving, teams are rewarded.

However, it is also in the interests of the debate community as a whole to ensure fairness in the adjudication of debates. Debating as a sport gains its uniqueness and credibility by rising above many of the worst parts of political discourse – a platform which is most definitely unfair. Half the reason some of the brightest university students from around the globe dedicate their lives, often at the cost of other pursuits, to competitive debating is that it rewards merit and logic and creates constructive discussion. As such, there is no doubt that a frustration debaters face – something which no doubt drives people away – is the feeling that the results of debates are unfair. The idea that considerations beyond the skill and talent of a team will ultimately play a part in the result of a debate is enough to discourage effort and reduce the overall

quality of debating. Why would any reasonable person expend resources on trying to make themselves the best debater they can be, when there is a good chance that their individual merit will not be a decisive factor in the result of any given round? The answer is of course that they would not, and we believe this is something regrettable for all who share a passion for debate.

As such, we would argue that establishing a level of fairness ought to be the number one priority for the debating community. For the purposes of this article we will advocate a view that fairness is achieved when the merit of each team is the sole consideration of the judge in coming to a decision, meaning that factors outside of the control of the team ought not be determinative.

Part Two: The way we assess debates right now is unfair

There are obviously countless ways in which the assessment of debates right now is unfair. Many former articles in this publication and others have dealt extensively with some of these issues, including gender, nationality and language biases. As such, this article will take issue with a very specific manifestation of unfairness, which can be described as unfairness in the substratum, or parameters, of debates. A judge ought to be believe that all four teams have an equal opportunity to get every result. However, often this basis is changing. This article focuses on situations where teams find themselves in substantially unequal positions due to a range of factors, which in application gives teams unequal opportunities to succeed.

Firstly, unfairness occurs when the very foundation nature of the debate is unequal. For example the topic itself sets up a much more difficult task for one team or side than the other, teams have unequal prospects of success.. Secondly, even if the motion has potential of fairness, actions of teams required to set up the parameters of the debate may create unfairness within the debate. For example, when a team squirrels or challenges a definition, teams within the debate may find themselves, through no fault of their own, in untenable positions. Thirdly, when the intuitions and biases of the judge establish and unequal playing ground, for example a judge is simply more receptive to arguments from a particular philosophical framework, teams may find themselves fighting from unequal ground. In this section we will deal with each

of these situations in turn and make an argument that the status quo often prevents the most meritorious team from achieving success.

When debates are unevenly weighted.

It is (hopefully) a rare situation, but there are times when the very ground on which the debate is built is uneven by nature of the motion. Some ideas are simply good or bad. No matter what you believe about the capacity of debate to create logical arguments for anything, you must acknowledge that there are some cases where that is not true, or at the very least it is differentially difficult on one side or team. It is certainly very uncommon for something genuinely indefensible to be set as a topic (although not unheard of), but as with anything, there is a spectrum of unfairness. Topics that are counterintuitive or incredibly complicated may force a certain team or bench to do more work than another in order to convince an average judge. We would argue that even that is unfair.

The WUDC in Chennai provides a clear example of this. Six of the top nine ranked teams on the tab (teams ranked 2nd, 3rd, 6th, 7th, 8th and 9th to be specific) were all knocked out in the Octo-final round by lower seeded teams from the government side of the controversial motion “THBT Japan should shame its soldiers who participated in WWII, including those who did not commit war crimes themselves”. To be clear, it is not our argument that there is nothing to be said on the government side of the debate, nor is it our belief that higher seeded teams are always the most meritorious in a round. However, we would suggest that the overwhelming preference for opposition teams getting through, despite their seed, seems to suggest that the topic was far more difficult for those on the government. Some teams did manage to get through from the government side. However, that does not invalidate the claim that the task of the opposition was easier in this round. We believe that an unbalanced topic is one of the external factors which prevents merit from being rewarded.

When teams muddy the parameters of the debate.

Even when the topic is balanced and has sufficient depth to allow all teams an equal shot at winning, the actions of teams within the round can create a more difficult task for some. We are all familiar with the concept of squirreling, and it is true that we discourage it under the status quo. However, we also ask

opposition teams to default to accepting the set up of the Opening Government team, so long as it is “debateable”. As it is now, Closing Government can only use a new definition if Opening Opposition challenges. In the instance where no challenge is offered by the Opening Opposition, the Closing Government is left with two choices – either defend the silly thing that their opening set up (something which the adjudication core did not intend) or be penalised for knifing. Through no fault of their own, this team is now in a more difficult position than other teams in the round, and has limited capacity to equalise the score. Similarly in this case, presuming the Closing Government choose to knife, the job of the Closing Opposition arguably becomes harder, with them having to oppose two different entire cases.

Again to offer an (albeit extreme) illustrative example from the authors own experience – a few years back in an out round at the Red Sea Open, the Opening Government misinterpreted or squirreled a motion and ended up running the opposition line rather than that prescribed for the government. The Opening Opposition, recognising that there was still ‘a debate’ to be had chose not to formally challenge the definition and instead offer some casual mockery. The Closing Government chose to run the topic as it was, therefore knifing the OG completely and also running the same case as the Opening Opposition. Finally the Closing Opposition were left to argue against two entirely contradictory cases, whilst also being incapable of defending their own for fear of agreeing with half of the Government bench. Obviously this kind of debate does not take place every day, but it does illustrate that sometimes the actions of a certain team create unfathomably difficult burdens for others. Much less egregious versions of this happen every day, when teams choose to block out other teams or simply make strategic errors that affect the capacity of the other team on their bench to win. We contend that the actions of team can create unfairness within the debate.

When the judge has prejudice.

Finally, we would suggest that the intuition and bias of the judge can artificially make one team’s job more difficult. Some judges have a preference for certain philosophical frameworks or simply believe particular arguments to be true. For example if a judge vehemently believes in a utilitarian framework, there is no doubt that any side of the debate that is mounting a rights based argument will have a more difficult time convincing that judge. This can often occur in debates that others may consider fair. An Adj. Core may value arguments on a

side of a motion, but by virtue of landing a judge who rejects the same arguments, teams find themselves in substantially unfair positions. Regardless of which of these circumstances unfolds, it is clear that there is a pervasive unfairness with competitive debate.

Part Three: There are unacceptable deficiencies in our current approaches to solving unfairness

Obviously we are not the first people to notice or identify these problems, nor will we be the last. Many potential solutions have been discussed and implemented, however it is our belief that they have been hitherto ineffective at meaningfully resolving the issues of unfairness explained above

With regards to unbalanced topics, there have been moves to create more rigorous assessments of the success of topics, with tournaments offering motion balance analysis as common practice. However, of course there are limitations to this approach. In the first place, it does nothing to resolve the unfairness that teams at that tournament faced, although it may prevent unbalanced topics from being set again in the future. Secondly, it is unlikely to even do that, given there is no systematic approach to ensuring balance employed by most adjudication cores. The current approach seems to be to presume that adjudication cores know best and assert that all topics that are set can be won by any team. The fact is that this is wrong. This is in no way intended to be accusatory – the authors of this article are guilty of setting unbalanced motions too. However, it is important to recognise that setting unbalanced motions is always a possibility, given the way in which topics are set. Adjudication cores are predominantly made up of excellent and accomplished debaters, meaning their interpretation of what can be reasonably expected of teams is skewed. They also spend a long time talking about and thinking about their topics, leading to an echo chamber effect and also probably the belief that good cases are possible on either side – it is easy to forget that what is possible in days and weeks is near to impossible in 15 minutes. Forcing teams to suck it up and just try their best to overcome the inbuilt bias in a topic is unacceptable and frustrating.

To the credit of adjudication cores and the debating community, there have been greater attempts to resolve the issues of unfairness within debates, for

example with moves towards mandating engagement through the POI rule. However, there is still a significant problem in the way we deal with squirrels and knifing. We ask the opposition team to simply run with whatever they are dealt, and penalise a closing team who choose not to defend something preposterous set up by their opening. Why should we choose to punish these teams and make their jobs more difficult through no fault of their own? We believe unfairness pervades these arenas of adjudication.

Finally, right now we attempt to deal with judge bias by asserting that judges ought not have bias. Of course we offering training, judge tests and feedback systems to try to weed out those with bias and demote them, however that does not really deal with the core problem. Bias will always exist, even amongst the best and most well enlightened judges, and no methods that we have right now are sufficient at overcoming that fact. Some of the current strategies can obviously be expanded upon: stricter selection of adjudication cores, more rigorous assessment of topic balance, mandatory POI's and clearer rules about squirreling etc. may all help to limit the degree of unfairness within debate, but they will not eliminate it.

Part Four: The application of the 'degree of difficulty' metric will help to correct these problems

When all else fails, and by virtue of either the topic, actions of the teams or the inbuilt biases of the judge, the capacity of one team or bench to win on merit is reduced, the question remains: what do we do? This article advocates a controversial solution which advocates for judges to explicitly account for the degree of difficulty faced by each team in coming to a result. While this intuitively seems quite extreme, this is actually an extension of an already accepted principle, both in the adjudication of debates and other sports. In the debating status-quo, we tell judges to 'be lenient' to Opening Oppositions who are faced with squirrels, given their relative disadvantage of having to make up a case on the spot. We also ask judges to weigh the contributions of the closing teams, accounting for their relative advantages of additional preparation time and seeing the issues of the debate play out in front of them. However, we are unwilling to codify the right of a judge to explicitly take into account, and reference in their judgements, the degree of difficulty faced by each team. To

the extent we do, we limit it to very specific circumstances. We believe the same principle that underpins these already accepted norms ought to be formalised and extended to all situations – essentially that we reimagine one of the core roles of the judge.

What is the 'degree of difficulty' principle, and how would it be used?
At this point you may be asking: what does any of this actually mean for a judge in a round like one we have described? No longer is the winner just the team that most convinced you; it is the team who convinced you most given their likelihood of convincing you at all. This doctrine is a fundamental rethinking of the scope of the judges role, and would require judges to ask themselves a new set of questions when coming to their decision. We would require judges, as a last step in coming to a decision, to adjust results based on the 'degree of difficulty'. Judges would need to ask themselves: do I believe, given the way the debate unfolded, that this team had an equal opportunity for success? If the answer is no, we would require judges to adjust their result to account for this. This assessment will need to account for the topic itself and any perceived balance or depth issues; things that took place within a round that artificially changed the difficulty of one team's task; and the judges own preconceived bias.

With regards to accounting for the topic itself and the way the debate plays out, judges would be required to think about any issues regarding the topic and reward teams who made the best arguments they could, even if those things were ultimately defeated within the debate. This will serve to address the situations outlined above.

Assessing judge bias may seem impossible to do, and to some extent that is true. However, we believe that encouraging a judge to think about their own opinions on the issue and be cognizant of those when coming to a decision is the best approach. For example, if a judge is encouraged to acknowledge (rather than hide) their bias, then they can attempt to correct for it. For example, a judge who is very predisposed to believe that liberty is good and government intervention is bad, may penalise a team that argued things they agreed with in a basic superficial way as compared to a team that argued things which they disagree with in a more logical way – this would be independent to some extent of who ultimately convinced them at the end of the day. For example, perhaps it is true that they still believed a world with small government was good, but might reward a team on government for making the best of (what they perceive to be) a bad situation. This is more likely to ensure

that the most meritorious teams succeed.

Principally justifying the ‘degree of difficulty” standard.

The underlying principle of this policy is to account for substantial inequality that occurs in a debate – even though all four teams, on paper, have the same chances of success. In defending this principle, we would draw an analogy to the underlying principle of Affirmative Action policies or redistributive taxation – the idea that merit can only truly be understood and rewarded when all begin from the same point. Those who support AA policies would make the argument that right now ‘merit’ based entrance requirements, for example to universities, do not really reflect the true merit of the individuals, given that one group begun from a disadvantaged position. It is our opinion that the way the role of the adjudicator is currently constructed limits their capacity to account for the relative difficulty in the starting positions of each team within a debate, and therefore prevents fair results from being achieved.

Difficulty is fundamental consideration in fairness. Consider a second analogy – a diving contest. Each dive that a contestant undertakes is given a difficulty rating and then the ultimate score that is awarded accounts for the degree of success at that dive, meaning that someone who poorly executes a very difficult dive may get a similar score to someone who executes a very easy dive quite well. The only real difference between this situation and the aforementioned ones is that these divers get to make the choice to attempt a more difficult manoeuvre, whereas the relative difficulty a team faces is out of their hands in debates. If anything, that suggests the introduction of a degree of difficulty criteria is even more important in the realm of debate than anywhere else in order to maintain the integrity of the competition.

Part Five: There are legitimate criticisms of this approach, but on balance, it will ensure greater fairness in results

The most fundamental criticism of this way of thinking is the idea that it jeopardises the key aim of debate, which we all know is the elusive ‘persuasion’. If the point of the debate is to convince the judge that they should

agree with your side and want to do or not do the thing you say, surely failure to do that means you have lost. To some extent that is true. However, we would suggest that winning a race when you had a 50 metre head start is not truly winning at all. More than that, we would of course suggest caution with the use of this principle – the degree of difficulty should only be one of many considerations a panel takes into account, and of course they should weigh it against whom they ultimately thought was the most persuasive in the round. However, the alternative is that they are precluded from accounting for it at all. For us this is a far greater concern.

Another legitimate grievance with this approach is that it is highly subjective and essentially means that the beliefs of the individual judge are determinative in the result of the debate. We acknowledge that, but would suggest a couple of important caveats. Firstly, British Parliamentary has the benefit of consensus adjudication. What that means is that issues of fairness and assessments of the degree of difficulty would be discussed and fought over by the panel, in the same way as all other aspects of the debate. If one person on the panel thought a topic was very Government weighted, one thought it lent to the Opposition and the other thought it was balanced, it would be unlikely that any degree of difficulty consideration would come into play. However, in the circumstance when the entire panel agrees that a topic is near impossible from one side, should they not have the right to account for that in their decision?

Secondly, all adjudication is subjective and that is a reality we ought to embrace rather than ignore. The particular opinions of the judge about the arguments, style and rules are decisive in results right now. What this approach does is to use that subjectivity in a more constructive way, by asking that judges embrace their preconceived bias and make an attempt to correct for it. We would argue that the consistency is found, counterintuitively, in the realisation that there is no consistency.

For those who fear overcorrection (for example that it will reverse the bias and make it basically impossible to win from the intuitive side of a motion), we would suggest again that it ought only be one factor judges consider. We would also question whether it is likely that a judge is going to go so far as to say that they cannot be convinced by the side they are naturally more convinced by. But again, we must compare with the alternative, which is a world where we make no effort to correct for the difficulty some teams face in winning from a particular side or position in front of a certain judge.

Finally, people may sensibly argue that all of these considerations (regarding motion balance, and the application of the knifing rule etc) should remain in the hands of the adjudication core. After all, these people are appointed due to their achievements, experience and respect, and we ought to trust them, more than any random judge, to fairly resolve these issues. Of course we will not question that adjudication cores do a generally fantastic job, and are certainly made up of incredibly qualified people. However, we would point out that there are some things that are structurally problematic about the way adjudication cores function, as discussed earlier. Beyond that, we would say that an adjudication core can never really account for the individual circumstances of any debate, nor the particular biases of any given judge. As such, we believe that the greatest amount of flexibility ought to be given to judges to deliver a fair decision in the context of the round they were entrusted to adjudicate.

Part Six/Conclusion: Setting an agenda for adjudication

We accept that many of the suggestions in this radical are both radical and controversial. Rather than discounting them out hand, we hope that readers will consider some of the problems we have outlined – problems we believe are widely recognised – and whether the 'degree of difficulty' principle is an acceptable and an effective solution to them. We contend that current strategies, at best, limit the regularity of unfair outcomes, rather than establishing a solution to them in their entirety. By no means do we argue that this paradigm shift is without fault or risk, but we believe it to be a necessary step in ensuring the integrity of the activity of competitive debate. Considering the starting position of teams within a debate, and their capacities to achieve a successful result is something we ought to allow judges to do in pursuit of creating a more enjoyable and rewarding experience for everyone.

December 22, 2014 Gemma Buckley and Josh Taylor Volume 12. 2014 Adjudication and Motions

Proudly powered by WordPress

Monash Debating Review

An annual publication of the Monash Association of Debaters

Comparing Experienced Judges and Lay Judges

R. Eric Barnes & Matthew McPartlon

Abstract

Though the wording of this platitude varies slightly when repeated at various judge briefings, it is commonly accepted that the goal of judges in British Parliamentary debate is to emulate the typical, educated, intelligent person. The primary question we are looking at in this study is whether actual BP judges are really doing this. We examine this by comparing the decisions made by normal judging panels at a tournament with decisions made by a panel of educated and intelligent people who have no familiarity with competitive debating. In investigating this question, we come across some other insights about judging as well.

Data Gathering

The HWS Round Robin ("HWS RR" hereafter) is an elite debating competition that invites 16 of the best debate teams and about 16 highly regarded debate judges from around the world each year. To be more precise, of the 16 judges in 2014, when this research was conducted: 13 had broken as a judge at Worlds (the other 3 had never judged at Worlds, but had accomplishments that would no doubt warrant them being invited as subsidized independent adjudicators); 4 had been Worlds grand finalists or had won the ESL championship; 5 had judged in Worlds semis or finals; 1 was top speaker at Worlds; 1 was a Worlds DCA; and 2 were Worlds CAs. Of course, this leaves out countless judging

credentials outside of the WUDC. Suffice it to say that this is an exceptionally strong set of judges. The judging pool was 25% female. A total of 6 nationalities were represented.

Over the course of 5 rounds, each team debates every other team exactly once. Judges are allocated such that no judge ever sees the same team more than twice and two judges are never on the same panel more than once.

In 2014, we ran a research study on judging by adding a panel of "lay judges" to each of the preliminary debate rounds. We recruited 40 people who had had no prior experience with competitive public speaking. These lay judges were recruited from faculty, staff and academically high-performing students at HWS. All lay judges were given a very brief (about 30 minute) orientation to judging BP debate, which was as neutral as possible regarding what constituted good debating. (See Appendix A for a summary of what was said at this orientation.) The primary purpose of the orientation was telling them what we were asking them to do and to encourage them to set aside any preconceptions about competitive debating.

The lay judges were assigned to rooms in panels of 3, with 1 person randomly designated as the chair. These people watched their assigned debates silently, as typical audience members would. After the debate was over, they were moved to another room and given 15 minutes to come to a decision about the debate, consulting with no one else. But, before discussing the debate among the panel, they were instructed to write down their initial call on a slip of paper, which we then collected. After the lay judges came to a decision (by consensus or vote), they filled out a ballot indicating team ranks and individual speaker points. In a few cases, there was more than one set of lay judges in the room, and in these cases, they deliberated entirely independently.

The pro judges stayed in the room after the debate and came to a decision, just as a panel ordinarily would. The only difference was that pro judges were also instructed to write down their initial call on paper that was collected.

Almost all of the 20 preliminary debates were video recorded and almost all of the judge deliberations were audio recorded.[1] This paper will not discuss any of the information from these recordings, though we hope to engage in some careful qualitative analysis of those recordings in a future publication.

All the quantitative data was entered into a spreadsheet and analyzed using the

methods described below. This included:

- Pro judge panel ballots (including speaker points)
- Lay judge panel ballots (including speaker points)
- Individual pro judge initial calls
- Individual lay judge initial calls

Methods

A central element of our analysis concerns comparing team rankings provided by individual judges and panels of judges. To do this, we developed a method of measuring the degree of difference between two *complete rankings* (i.e., ordinal rankings of all four teams). The difference between two complete rankings can be measured on a scale from 0 (representing an identical ranking) to 6 (representing a maximally divergent ranking). A complete ranking can be translated into a set of 6 bilateral rankings, comparing each possible pairing of teams out of the four teams in the room. Each bilateral ranking was scored as a 0 if the two complete rankings agreed on which of those two teams should be ranked higher, and was scored as a 1 if they disagreed. These six scores were then summed to provide the final divergence between the two complete rankings on the 0-6 scale. [2] So, the least divergent rankings (other than full agreement) would be a situation where the rankings are the same, except for two adjacently ranked teams being switched. See the examples below:

Divergence = 1				
	OG	OO	CG	CO
Judge Call	1	2	3	4
Final Decision	1	3	2	4

Divergence = 3				
	OG	OO	CG	CO
Judge Call	1	2	3	4
Final Decision	3	2	1	4

Divergence = 5				
	OG	OO	CG	CO
Judge Call	1	2	3	4
Final Decision	3	4	2	1

We also wanted to measure how similar the initial calls from an entire panel were. To do this, we simply created three pairs of complete rankings from the three judges, calculated the divergence for each of these pairs, and then summed these. This gives a scale from 0 (no disagreement) to 12 (maximum disagreement).[3] To make this easier to grasp, consider the table below, where the “call difference” is the degree to which the three judges calls differed.

Typical Examples of the Various Degrees of Call Differences

Call Difference = 2	OG	OO	CG	CO
Judge A	1	2	3	4
Judge B	1	2	3	4
Judge C	1	3	2	4

Call Difference = 4	OG	OO	CG	CO
Judge A	1	2	3	4
Judge B	2	1	3	4
Judge C	2	1	4	3

Call Difference = 6	OG	OO	CG	CO
Judge A	1	2	3	4
Judge B	1	4	2	3
Judge C	2	1	3	4

Call Difference = 8	OG	OO	CG	CO
Judge A	1	2	3	4
Judge B	2	4	1	3
Judge C	4	2	1	3

Call Difference = 10	OG	OO	CG	CO
Judge A	1	2	3	4
Judge B	4	1	2	3
Judge C	4	3	1	2

Call Difference = 12	OG	OO	CG	CO
Judge A	1	2	3	4
Judge B	2	1	4	3
Judge C	4	3	2	1

We used averages of these measures to answer the following questions:

- Did pro or lay panels show greater differences in their initial calls?
- Did pro or lay judges tend to alter their rankings more to arrive at a final call?
- How different were pro and lay panel rankings from each other?

To test for statistical significance of these differences (between A & B), we used a t-test for sample means, controlling for unequal variances:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{N} + \frac{s_2^2}{N}}}$$

We tested the following hypothesis to determine the likelihood that the differences were random:

$$H_o: A = B$$
$$H_o: A \neq B$$

As a point of comparison, we sometimes include what a set of random rankings would look like. To generate this random data for initial call differences between a panel, we numbered all 24 possible rankings for a BP debate, then we used a random number generator in Excel to create three independent random numbers from 1 – 24. We then calculated the call difference of those three rankings and recorded it in a spreadsheet. We did this 100 times and used that sample as our random data for call differences. To get a "random" data distribution for the divergence between just two rankings, we calculated the divergence of the ranking (1,2,3,4) against each of the 24 possible rankings and used that as our "random" distribution. Although not generated randomly, any arbitrarily large set of paired rankings (each randomly selected) would

converge on this distribution, so it should more than suffice as a stand in.

Findings & Discussion

Based on our analysis, there are five areas that we want to discuss: 1) the correlation between lay and pro judges regarding team point decisions; 2) the relative similarity between the initial calls of the two kinds of judges; 3) the movement between initial calls and final decisions for the two kinds of judges; 4) situations in which we placed two lay panels in the same room; 5) judge bias toward particular positions in the debate.

Similarity of Final Decisions

The data clearly shows that there is a correlation between the winners chosen by the lay judges to those chosen by the pro judges. It would have been both horribly depressing and a damning indictment of our activity if this had not been the case.

At the same time, we want to note that the break would have looked very different if the lay judges had been deciding the winners. The top breaking team would not have changed and the second team would have squeaked in as 4^{th} seed (on a tie-breaker), but the other two teams who broke to finals would have been 8^{th} and 9^{th} on the tab. What stands out to us in the comparison of the results from lay vs. pro judges is that there were 3 teams whose total team points from the two groups differed by 5 or 6 over just five rounds. An additional 4 teams had results differing by 3 or 4 team points. Putting this another way, there were 2 teams that the pro judges liked much more than the lay judges (5 points), and there were 3 teams that the lay judges liked much more than the pro judges (4-6 points). The average difference for a team at the end of five rounds was 2.625, which is substantial, since the average point total is 7.5. The chart below shows the different results from the two sets of judges. Team names were alphabetized to show the order in which they would have finished by the lay judges' rankings (i.e., "Team A" would have broken first, "Team B" second, "Team C" third, etc.).

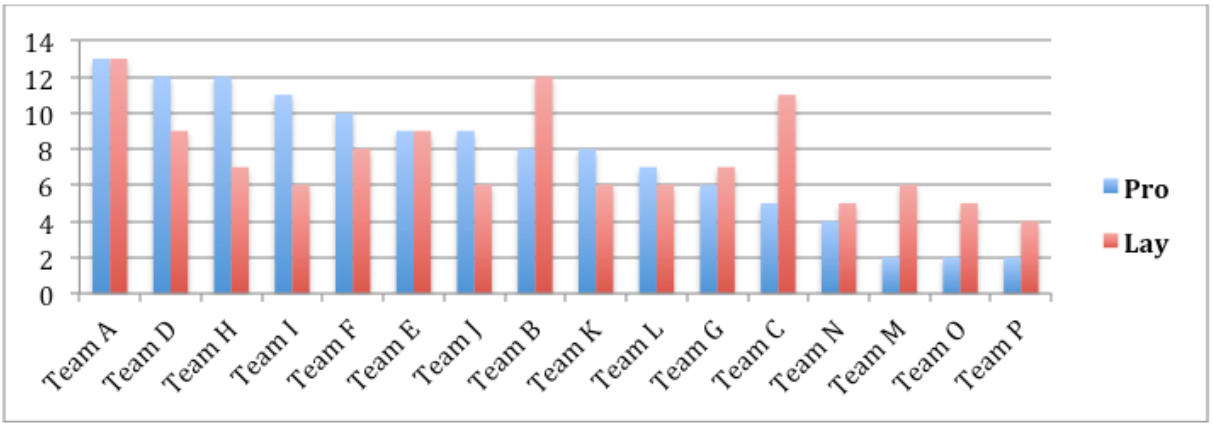


Figure 1: Total team points given by pro and lay judges

Although there is a correlation between the two sets of results, there are some significant aberrations. In fact, the differences between the two sets of results are greater than the chart above suggests, since the data presented in the chart above only considers the teams' final score, not the *accumulated variance* in the decisions from each rounds.[4] So, the chart below may better represent the amount of disagreement between lay and pro judges. What is striking here is that even in cases where there appeared to be strong agreement on the results (e.g., teams A, E, G and L), the reasons for that result were very different, varying by 4-6 points in these four cases. Though we did not represent it in the chart below, the expected accumulated differences between any set of team rankings and a random set of rankings is 6.25 for each team over five rounds. So, they two sets of judges are coordinating better than random, but that's not a high bar.

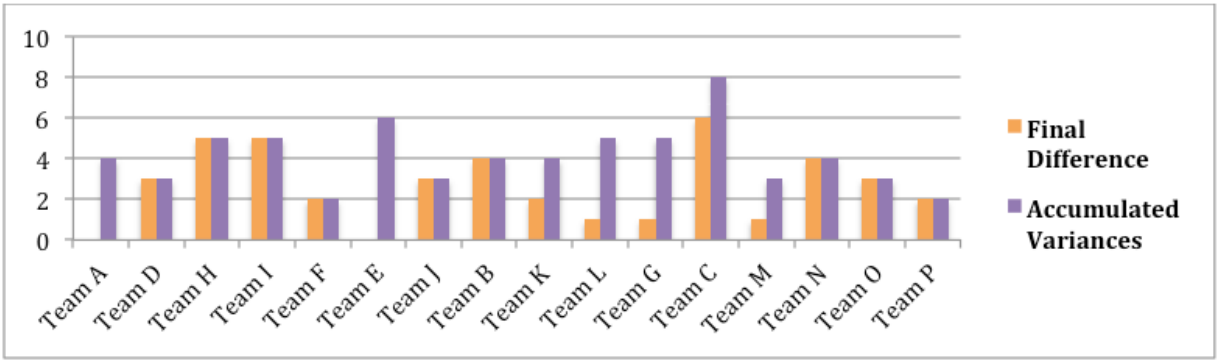


Figure 2: Accumulated differences in team points (between pro and lay rankings)

Clearly, the pro and lay panels saw some debates very differently, and the quantitative data that we have will not answer the question of *why* this is the case. Our intention is to move forward with this research by engaging in a qualitative analysis of the audiotapes that we have of the deliberations of pro and lay panels, particularly in the rounds where they disagreed markedly.

While the accumulated differences shown above make it seem as though the decisions by the lay and pro panels were quite substantial, things look

somewhat different when we view the data in a different way. We calculated the divergence between the rankings of the pro and lay panels in each round and found the distribution of these. For comparison, we added what a distribution of divergences from random rankings would look like and we also added the distribution of divergences from individual pro judges on the same panel at this tournament.

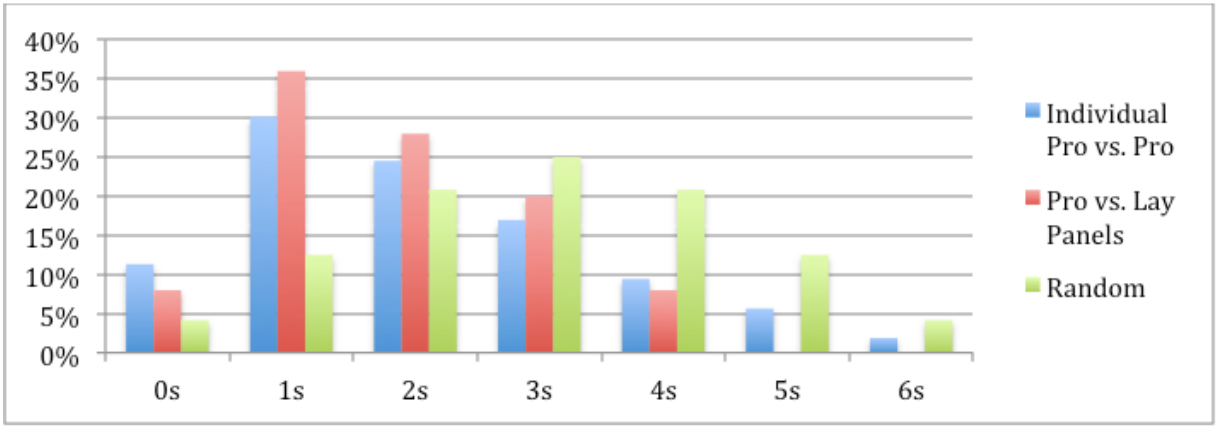


Figure 3: Divergence of pro and lay final panel rankings

There is only the smallest possible divergence or no divergence at all between the lay and pro panels in 44% of all cases. In another 28% of cases, there was a divergence of 2, which we still consider a fairly similar ranking. Although a divergence of 3 or 4 is definitely substantial, it is important to note that there were no cases where the calls of the two panels diverged as much as 5 or 6. The lay panels diverged from the pro panels slightly less than the individual pro judges on the same panel diverged from each other.[5] This suggests that there is not such a big difference between how pro and lay judges see the debates.

Similarity of Initial Calls

Regarding the differences in the initial calls of the judging panels, the data reflected what we expected to see, but not to the degree that we expected. Pro judge panels tended to be more consistent (i.e., less divergent) than lay judge panels. However, there was a greater average difference than we had expected between the initial calls of the pro judge panels. In other words, we expected the pro judges to agree even more before the panel discussion began.

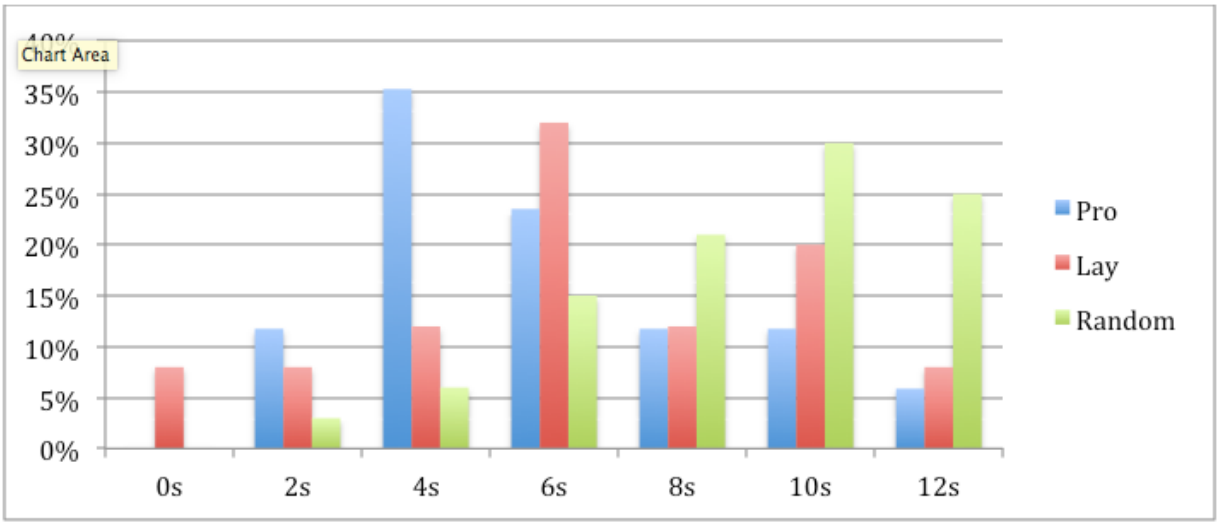


Figure 4: Degree of call difference (divergence in rankings among a 3-judge panel)

In the 20 prelim rooms, there was never a case where the pros completely agreed, though perhaps that isn't quite as remarkable once you consider that the odds of this agreement randomly happening are 1 in 576. In 47% of the rounds the pro panel's call difference was minimal, meaning that it was either 2 (the smallest possible difference, outside of complete agreement) or 4 (the next smallest). We see these differences as relatively minor, and indicative of a panel being largely on the same page at the end of the debate. These are situations that would likely set the stage for a fairly easy deliberation. We consider call differences of 6 to be moderately divergent. Although panels with call difference of 6 will find it somewhat more difficult to reach consensus, there will be some clear commonalities in the three judges' rankings that can help to find a path to consensus. About 24% of pro panels fell into this range. We consider panels with a call difference of 8 (12% of pro panels) to be significantly divergent. These panels will likely struggle to find commonalities in their rankings, though some will likely exist. We consider call differences of 10 or 12 to be extreme, since these calls indicate virtually no agreement. In 18% of pro panels, there were such extreme differences. We feel sorry for the people engaged in these deliberations. Of course, we acknowledge that in some cases, these extreme call differences can dissipate quickly once the panel resolves one or two central questions about the debate. But, many times this is not what happens.

In contrast, the lay judges had minimal or no call difference (0-4) in 28% of their panels, many fewer than the pro panels. About 32% of lay panels had a moderately divergent call difference of 6. About 12% of lay panels had a significantly divergent call difference of 8. The remaining 28% of panels had extremely divergent call differences.

We note that even with panels of uniformly excellent judges, about 30% of panels will disagree to a significant or extreme degree in their initial impression about who won a debate. This fact strongly suggests that even the most confident judge among us should cultivate a sense of humility regarding their call in a debate. This is even clearer for those people considering criticizing a decision without participating in the deliberation process.

The average call difference for lay panels was 6.5, with a standard deviation of 3.38. This compares to an average of 5.9 for pro panels, with a standard deviation of 2.87. An average random set of three rankings had a call difference of 8.9, with a standard deviation of 2.67. We had expected that pro judges would have a certain uniformity of expectations and criteria and that this would result in more uniformity in their initial call. While our findings were not strictly inconsistent with this, as mentioned above, we had expected to find a larger gap between the pro and lay panels in this respect. The gap we found was not even statistically significant.[6]

The size of these call differences suggests that all judges should remember that the panel deliberation is a essential element in coming to a good decision and that judges (chairs in particular) should not see their job in the deliberation as ensuring that the other judges are willing to go along with their initial call.

Movement from Initial Calls to Final Decisions

We use the term “movement” to refer to how much a judge’s initial call diverges from their panel’s final call. When there is an initial call difference among the judges on a panel, there will necessarily be some movement by some of the judges. But panels will not always come to a final decision that minimizes how much the judges move. For better or worse, in practice, panels sometimes engage in deliberations that cause everyone on the panel to change their mind about a ranking that they had all agreed on. So, the existence of an initial call difference sets a minimum amount of movement that needs to happen to reach consensus, but judge movement can significantly exceed this. In theory, a panel could start with complete agreement (i.e., no call difference) and end with a call that is completely different from what everyone initially thought. So, movement measures something new.

Given that pro judge panels had a lower call difference on average, one might reasonably expect that the pro judges would tend to move less than the lay judges. However, the data showed that the lay judges moved an average of 1.3

between their initial call and their final judgment, with a standard deviation of 1.41. The pro judges moved an average of 1.8 between their initial call and their final judgment, with a standard deviation of 1.43. This difference of .5 is both statistically significant and potentially revealing.[7] One might try to explain the fact that lay judges moved less than pro judges, by focusing on the 2 lay panels that agreed immediately, but this only accounts for 6 of the 27 instances of 0 movement. Moreover, these 2 panels with 0 call difference equally affected the call difference average, and so cannot really explain the fact that lay judges moved less even though they started by disagreeing more.

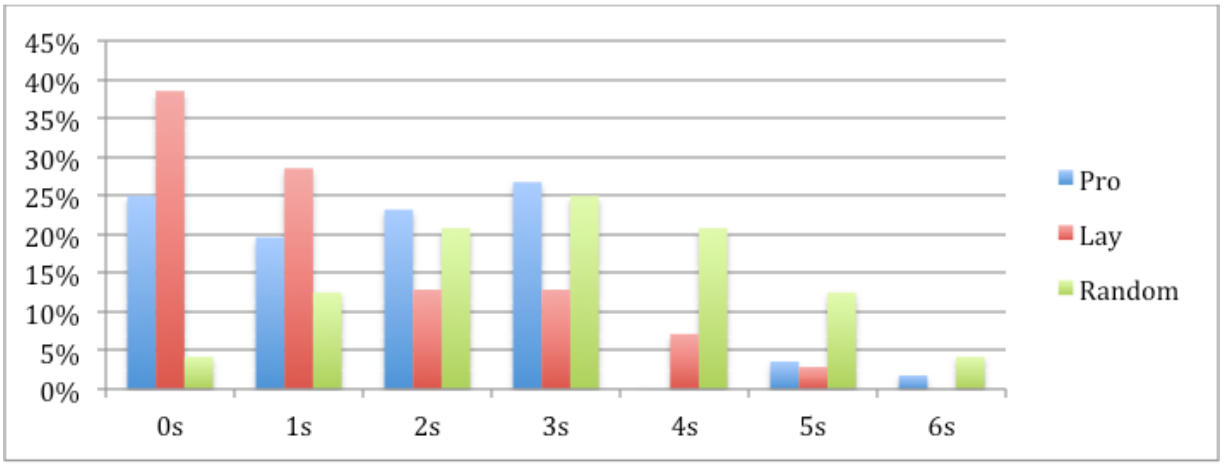


Figure 5: Movement between initial judge rankings and final panel rankings

The chart shows that lay judges most frequently do not move at all from their initial call, and their tendency to move tappers fairly steadily as the divergence increases. In contrast, the pro judges were about equally likely to move 0, 1, 2 or 3 degrees to the final panel decision, but the likelihood that they would move more than 3 drops precipitously. It is unclear if this precipitous drop is just a statistical aberration based on our small sample size or if there is a real cause to why pro judges are dramatically less likely to move beyond 3 degrees of divergence.

One possible explanation of why the lay judges had a smaller average movement despite starting further apart is that lay judges were more conciliatory and attempted to minimize the degree to which the panel members needed to move by being willing to compromise (i.e., split the difference). This is only one possible hypothesis and we make no judgment about whether a conciliatory attitude is beneficial to judging or not. It is possible that discussions among the pro judges revealed deeper insights into the debate that caused many people on the panel to reevaluate their initial calls. It is also possible that pro judges attempted to do this, but actually just ended up distracting themselves from their more accurate first impressions.

Perhaps our future qualitative analysis of the deliberation recordings will shed some light on this.

Do distinct lay panels come to similar conclusions

As we said above, at the start of this research we anticipated finding that pro judges were more internally consistent (less divergent) in their rankings, both as individuals and as panels. The analysis of call differences suggests that as individuals, pro judges are more consistent with each other than lay judges are. However, we have no direct evidence about the extent to which different pro *panels* would be consistent. Our data did provide us with some modest evidence about consistency between lay panels because we had enough volunteer lay judges during some rounds to put two lay panels in the same room. We were able to do this five times and the results seem worth reporting.

Rooms with 2 Lay Panels	Degree of Divergence (0-6)		
	Lay1 v. Lay2	Lay1 v. Pro	Lay2 v. Pro
A	2	1	1
B	1	1	0
C	0	1	1
D	1	3	2
E	0	2	2

There was a high degree of consistency between the two lay panels in these five rooms. In two rooms, they were in perfect agreement. In two others, they had the smallest degree of divergence and in the final room, they diverged by 2 still were largely in agreement. So, the average divergence between 2 lay panels was 0.8. In contrast, the average divergence between these panels and the pro panels that were in their respective rooms was 1.4. The sample size is too small to determine statistical significance, but it seemed to us that it was worth remarking on.

Judging bias towards debating positions

We looked at the data on how well the various team and speaker positions did according to the points that the two sets of judges awarded them. The clear trends in the data are:

- Closing opposition teams were likely to do better with both pro and lay judges
- Opening government teams were likely to do worse with both pro and lay judges

- These biases were more pronounced with the lay judges, especially the preference for closing opposition teams.[8]

To provide a frame of reference, we compare our data from the 2014 HWS RR with data from the past seven years of the HWS RR, and also with the data from the 2014 WUDC in Chennai. We compared these by adding up all the points won by teams in each position during the preliminary rounds of these tournaments and then calculating the percentage of the total points that this represented.

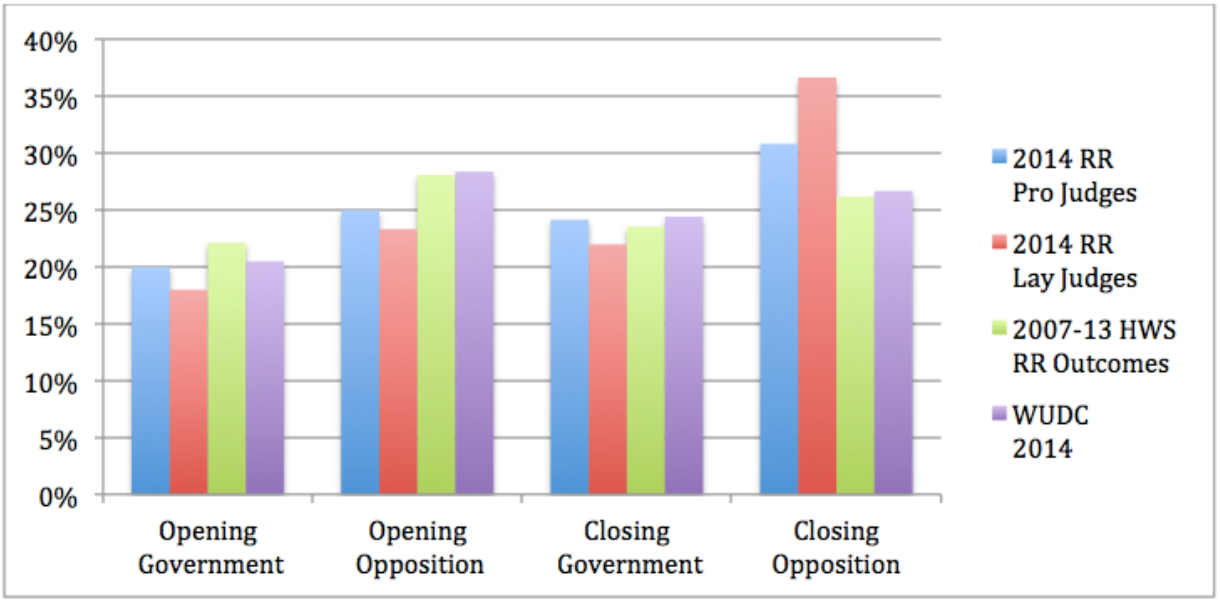


Figure 6: Percentages of points awarded by judges to teams in the four positions

The result was both interesting and remarkably boring. The results are boring because all of these sets of judges award points in basically the same zigzag pattern. But, the results are interesting partly because there is this consistency, and particularly because the lay judges not only replicated this pattern, but did so in an exaggerated manner. This strongly suggests that the bias in favor of opposition teams (and against the opening government team) is not a function of some set of habits or expectations developed within our debating community, but rather is an outgrowth of something about how the nature of those positions relates to an audience.

As a final note, we hope that this short publication will spark discussion about these issues and will also prompt people to suggest new ways for us to analyze the data that we have at our disposal.

Limitations & Directions for Future Research

There were several limitations on our research.

- Obviously, with only 20 preliminary rounds, the data set we are working with is a fairly small sample size.
- Because the HWS RR has such an unusually high caliber of debaters and judges, one might question the extent to which we can generalize to more typical debates.
- Because the HWS RR uses team codes, the debaters schools were anonymous with lay judges (who were also unaware of particular debater reputations), but most of the judges were likely aware of who all (or almost all) of the debaters were.
- A small amount of our data needed to be discarded because forms were incomplete or not filled out correctly (e.g., a judge would fill out the initial call sheet without giving each of the four teams a unique rank from 1-4).
- There were no rooms with more than one pro judging panel, so we are unable to determine the consistency between pro panel decisions after deliberation.

As mentioned above, we plan to pursue further qualitative research based on the audio recordings made at the 2014 HWS RR. This will hopefully provide a significantly more textured and nuanced view of what was happening within the deliberations of panels with the two kinds of judges.

Conducting this research again at the HWS RR or at other tournaments could increase the sample size. Additionally, it would be fascinating to gather more data on the consistency of pro panels. On possibility would be to hold a (presumably small) tournament where each room had two pro panels. Teams would simply accumulate points from both panels. This would be very simple to do in a round robin format, but would also be possible in more traditional formats, though it would need to be hand tabulated (or software would need to be developed). Such a tournament could provide a wealth of useful data about how consistent judging panels are.

Appendix A: Instructions to Lay Judges

The handout below was given to all volunteer lay judges along with an explanation of each point on the handout. Volunteers had an opportunity to ask questions as well. All volunteers were screened to ensure that they had had no previous exposure to any form of competitive debating.

HWS Debate Research

Before you start:

– Please set aside everything you think you know about what competitive debate should be like.

– We are interested in your perspective as an intelligent and thoughtful listener.

– It is not easy, but please do all you can to *set aside your own personal biases and beliefs*.

– Try to forget whether you actually agree with one side or the other.

– Try to forget any particular pet theories that you tend to favor.

– Try to adopt what you take to be the *bland* beliefs of a typical, intelligent, educated person.

– If an ordinary, intelligent, educated person would accept or reject a claim, you should too, regardless of whether other debaters refute it.

– Ask yourself: Who would have persuaded me most if I really were an unbiased person?

This is a contest of who is best at rational persuasion, not a contest of who presents the most eloquent speech. Obviously, good speaking style helps one persuade an audience, but we are asking you to judge what would actually persuade a rational, intelligent and educated audience. This is **a holistic judgment** that is not exclusively about style or content. The question is "Who was most persuasive?" and we offer no formula for coming to that decision.

Things you must know:

– There are 4 teams competing in each debate.

– The 2 teams on the left are supporting the plan or proposition stated by the first speaker.

– The 2 teams on the right are opposing this plan or proposition.

– But, judges do not declare either “side” of the debate (i.e., either “bench”) the winning side.

– Rank them “Best”, “Second”, “Third” and “Fourth” based on how persuasive they were.

– Which team, considered as a whole, was most likely to ACTUALLY persuade an unbiased, intelligent and well-educated audience.

– *Before* your panel begins its discussion, please take just one or two minutes to write down the ranking of the teams that you (on your own) think is most appropriate. But, after this, please be willing to revise this ranking if the discussion actually makes you see things differently.

– Judging a debate is a COOPERATIVE exercise. DO NOT VIEW THIS AS A COMPETITION to convince the others that your initial impression is correct. The goal is to work together to find the best answer to the question of which team was more persuasive of an intelligent, educated and unbiased audience.

– After coming to a decision on the team rankings, we ask that your panel assign points to each individual debater on a scale of 50 (poor) – 100 (excellent). These points should reflect the speaker’s overall contribution to persuading an intelligent, educated and unbiased audience that their side is correct.

– So, this includes quality of argumentation and quality of style.

– The average points at this tournament are typically about 79.

Things you should know:

– One person in each panel of 3 judges has been assigned to be the “chair”, which means only that they keep an eye on the time and try to ensure that the deliberation moves along so that your panel is ready to render a decision at the end of 15 minutes about how all 4 teams ranked.

– The 2 teams on the same side need to (largely) agree with each other.

– Disagreeing with a team on the same side is called “knifing”.

– This is to be considered a negative exactly to the degree that it undermines the overall persuasiveness of their side's position. (So, a small disagreement about an unimportant element can be mostly ignored.)

– The debate is about the main *proposition* articulated by the first speaker, which may be somewhat more specific that the general 'motion' (i.e., topic) announced before the debate. Focus on the proposition, not the motion.

– You are permitted to take notes, but you are not required to do so.

Things you might want to know:

This is a guide to some unfamiliar terminology that might be used in the debate. Below are the names of the various teams (in the outside columns) and names of the individual speaking positions (in the inside columns):

<table>
<tr><td rowspan="2">Team: Opening Government</td><td>Prime Minister (PM)</td><td>Leader of the Opposition (LO)</td><td rowspan="2">Team: Opening Opposition</td></tr>
<tr><td>Deputy Prime Minister (DPM)</td><td>Deputy Leader of the Opposition (DLO)</td></tr>
<tr><td rowspan="2">Team: Closing Government</td><td>Member of Government (MG)</td><td>Member of Opposition (MO)</td><td rowspan="2">Team: Closing Opposition</td></tr>
<tr><td>Government Whip (GW)</td><td>Opposition Whip (OW)</td></tr>
</table>

– During the middle 5 minutes of a speaker's 7 minute speech, debaters on the other side can stand up for a point of information (POI). The speaker can either accept or turn down these POIs, but typically they are expected to accept 2 during each speech. The perception is that failing to do this demonstrates a lack of confidence.

1. There were some technical difficulties that prevented recording in some of the debates. ↩
2. In other words, for any two ordinal rankings of four teams in a room (e.g., CG/OG/CO/OO and OG/CG/OO/CO), we asked the following six questions: Did they agree on whether OG placed above OO?; Did they agree on whether OG placed above CG?; Did they agree on whether OG placed above CO?; Did they agree on whether OO placed above CG?; Did they agree on whether OO placed above CO?; Did they agree on whether CG placed above CO? Using the example just given, the answers would be: yes, no, yes, yes, no, yes. Answers of "yes" were represented with a 0, while answers of "no" were represented by a 1. So, in the same example, the answers were represented

as (0,1,0,0,1,0). The sum of these represents the divergence between two rankings. So, in this example, these rankings diverge by 2 degrees out of a possible 6. ↩

3. Only even numbers are possible on this scale, but we chose not to simplify it to a 6 points scale in order to make it more obvious when we were talking about comparing panel rankings, as opposed to bilateral comparisons between rankings. ↩
4. In each round, the difference between the team points given to particular team by the pro judges and the lay judges will be somewhere between +3 and -3. The "accumulated variances" for a team is the sum of the absolute values of all individual divergences in the five prelim rounds. The "final difference" is the sum of these values (not the absolute values). So, for example, imagine that a team got the same points from the pro and lay judges in the first three rounds, then in round four got ranked 1 point higher by the lay judges than by the pro judges, and then in round five got ranked 2 points lower by the lay judges than by the pros. That team would have a final difference of 1, but an accumulated difference of 3. ↩
5. Comparing the lay and pro panel divergence to the divergence between individual pro judge rankings and their pro panel rankings would not be useful, because those are not causally independent rankings. Below, we do discuss the distinct issue of how much judge rankings move from their initial call to the final panel ranking. ↩
6. The t-statistic for the test on this data turned out to be 0.6124. Given this statistic, we fail to reject the null hypothesis that the average difference in initial rankings for lay judges and debate judges are equal. ↩
7. The t-value = 3.33, significant at a 99% confidence level (i.e., a significance level of 0.01), found by t-test for sample means controlling for unequal variances. Given this statistic, we do reject the null hypothesis that the average movement for lay judges and debate judges are equal. ↩
8. We are not using the word "bias" in a pejorative sense. We mean it merely in the statistical sense. ↩

December 23, 2014 Eric Barnes Volume 12. 2014 Adjudication and Motions

Proudly powered by WordPress

Monash Debating Review

An annual publication of the Monash Association of Debaters

Building the Narrative

Written by Andrew Gaulke
With excerpts from an interview with Tim Sonnreich

*

We all like to think that we're immune from being manipulated. We all think we're too smart for advertisers, or we are too smart for people who play reverse psychology on us, or whatever it might be. Actually those things exist for a reason, and we are, to a greater or lesser extent, influenced by them.

*

Debaters and adjudicators like to think that the formal logic of the argumentation is what will win debates for them. However, there remains an element of persuasiveness that is just as powerful, yet much harder to grasp. Every debater knows the experience of adjudicators who seem to totally misrepresent the arguments they made. Every adjudicator knows the experience of finding one team more persuasive without quite knowing what argument made them win. This article explores one element of that additional realm of persuasiveness, the construction of a narrative out of argumentation.

Much of what this article explores is intuitive, and much of its advice is already in practice in high level debates. My aim here is to provide some theoretical ideas as to why we debate the way we do, as well as providing some clarity to help new debaters understand case construction and persuasiveness on that level.

I want to begin by nailing down exactly what it is I'm talking about in this article. A speech in a debate can be understood two ways. The first is a philosophical understanding of how the speech functions. In this type of analysis you can extract a set of logical ideas from the speech, and those ideas tell you what the speech was trying to say. You can change the order of the

parts, you can change the specific wording in a lot of places, but the logic will remain the same. A philosophical understanding of the speech that focuses on formal logic wouldn't change its assessment of how that speech worked based on structural or cosmetic changes that do not substantially change the logic that is presented.

That is not the complete story of how human beings respond to a speech, however. I intend to introduce a second, literary understanding of how a speech operates. The structure of our argumentation is important, and should be analyzable. The words and rhetoric we use are important, and should be analyzable. A literary understanding of how a speech operates incorporates those intangible elements of persuasion. It understands the total effect of a speech on those who hear it instead of breaking the speech down into its individual logical components. It understands how each element contributes to the whole.

It takes a lot more than just good arguments. You can see that with a lot of ESL teams, where they basically make the same arguments that everybody else makes, but sometimes it's a language issue, or sometimes it's just about the way they package and present their arguments, that let them down. People take a lot of subtle cues from the way people present, in a manner sense, that hold them down. Or just the way they position themselves in a kind of rhetorical sense that can leave them out. It's not so much that they don't have the words for it, but it's that they miss the chance to build momentum and persuasion in what they're saying and just kind of jump to ploughing through the arguments.

This is why the idea of a narrative is so important. A narrative, at its simplest definition, is the way we understand how two pieces of information relate to each other. When we read a novel we understand that the events of the story are connected to each other, not only on a causal level, but also thematically and conceptually. The events can build up into a broader understanding of what a particular novel is trying to communicate.

The same thing is happening when we construct a speech in debating. When an adjudicator hearsa speech they are trying to construct a coherent meaning out of it. They want to know what the speech is about. They are trying to understand what core idea is at the heart of a particular speech because that is how we unconsciously code information. Understanding this process provides a powerful way for debaters to craft their speech for the most persuasive effect.

One of the best articulations of this comes from Tim Sonnreich's "First Principles" approach to debating. This is the idea that the most persuasive cases are constructed based on particular political ideologies that can form a core principle in the case. In other words, the persuasiveness of a case comes when the individual arguments add up to something more. They are connected by the ideology behind them, even when that ideology is unspoken.

For example, a common set of arguments in debating relate to how far the government should be allowed to control the choices of its citizens. When debaters argue that a particular policy unjustifiably infringes on people's lives they do not deploy that one argument in isolation against one policy idea in isolation. They are instead trying to construct a particular understanding about how the world works and show that, based on that understanding, we ought to set broad criteria on government intervention.

That broad understanding of the world is the narrative that unifies the meaning of the separate pieces of logic. A typical antigovernment intervention case might run as follows: first, there is an inherent value in individual freedom; second, governments are bad at making decisions for individuals; third, cultural and economic problems are solved best in unregulated environments. All of that material builds into a picture of what the world looks like. It is connected by that image of the world (in this example, a libertarian image of the world) even when those connections are not explicitly drawn by the debaters. Often those connections are not explicitly drawn, as the three themes of this broad superstructure are usually labeled based on the individual character of the topic at hand.

And when I talk about the narrative, it's about being able to give a compelling story about what the world looks like in your mind that's different than the world we live in today. And whether that's a small change or a big change, and what the reason is for that change, and how that fits with the way we currently think now, and all of that helps. Because ultimately in a debate – you know we talk about whether we proved such and such a thing in the debate. You never actually prove anything, or very very rarely do you prove anything in a debate. What you do is provide enough justification for the proposition you're making that an average reasonable person has a willingness to believe that it's probably right, without having gone to do any more thinking about it.

The reason this idea is important is because all the logical argumentation of a speech is understood based on how it relates to the core idea that the

adjudicator understands the speech to be about.
One of the simplest ways to understand this is in how an opening government team constructs the problem they are trying to solve. When a PM's introduction focuses on a particular problem an expectation is created in the mind of the adjudicator that the material that follows is intended to address that problem. When debaters flag their main themes it similarly creates expectations as to how the case being presented will understand the world. The structural signaling is something all debaters intuitively know to be important, and the reason it has grown to be so common in debating is because it provides the broad idea of the case so that adjudicators can link each individual element of logic back to that understanding.
And the reality is, particularly in a high level debate, you actually prove a lot less than you think you do, because you end up covering a lot of ground. There are a lot of arguments, and a lot of complexity to them, and just being able to explain your opponents argument, before then going on to deconstruct it takes a lot of time. So what you're really doing with the narrative is, you're giving the adjudicator the broad brushstrokes of what the final goal looks like because that helps them join the dots for you. They can see where the starting point is because you described that at the start, and they can see the end point, and then your arguments helped them to see that there's a path between them and if you do enough to show where those paths are then you win.
What often separates top level debaters is the clarity with which they communicate this hidden core idea of their case. They will use their introductions and conclusions carefully in order to make their position clear.

Understanding that this core narrative exists behind the logical argumentation is especially helpful in organizing rebuttal to a case, and in particular the strategic selection of rebuttal. Effective rebuttal is not only trying to attack a particular logical point, but also trying to break up the coherence of the world view the opposition is attempting to create, while at the same time re-enforcing the debater's own broader narrative of the world. It is impossible to effectively rebut an argument without understanding the role that argument plays in the debate. When, at a high level, arguments are well constructed and will be persuasive to a certain extent no matter how they are rebutted, it is important to see the way that the argument connects to the broader persuasive narrative the team is building. That is how you can know the extent to which you need to attack the argument in order to overcome it.
You don't really have to worry about what the logical fallacy is in the argument because again well-constructed arguments don't have logical fallacies, but no matter how well constructed the argument is, there are philosophical

alternatives to that argument because it is a debate.
This is why First Principles analysis can be particularly persuasive. When a case is constructed based on a First Principles position that case will have an internal consistency and clarity that allows for a coherent narrative. Additionally, the case will be based on often familiar narratives that the adjudicator already knows and understands. An adjudicator is likely to already understand the idea of a libertarian world view, and as a result that idea will be particularly clear to them in its construction. The type of world, and how that world functions, will be that much easier to imagine.

When a team tries to describe a libertarian world they are able to do it more efficiently because the thread of meaning that connects the individual pieces of information are easier to imagine. All cases have gaps as a simple necessity of limited speaking time, those gaps are filled in the adjudicators mind based on the narrative the speech constructs about the world, the meaning that connects the information. An argument is harder to attack when the adjudicator is doing the work for you by filling in those gaps themselves. A strong, coherent narrative encourages them to do that.

Debaters need to understand that their use of language, introductions, conclusions, and the choice of focus in their speech contributes to how the adjudicator will understand all of their material. When done well a strong internal narrative can be an excellent tool for debaters, when done poorly that can be exploited by the opposing teams whose narrative of the world is stronger.

So when I talk about narrative in a first principles sense what I'm really saying is, you can give people the image that you are trying to create, even before you get to describing the arguments for why that's a good image, so that you can then rely on and lean on that throughout the debate, to join the dots, make more cohesive what is really a bunch of singular arguments which you have chosen because given the amount of time and space you have, they're the best choices you've got to explain them, whereas in a different format you might chose a totally different way of explaining something.

When both sides of the debate are able to effectively communicate their narrative about how the world works then the clash in the debate becomes, at least in part, about the strategic choice of narrative from each side.

Both teams are trying to position themselves strategically in relation to the

overall clash of the debate. I have so far talked about building a narrative within individual speeches, but it is important to remember that the debate itself will also have an overarching narrative. When an individual makes a speech they are creating the narrative of their own case, while at the same time contributing to the narrative of the debate as a whole.

For example, an individual speech may have the core narrative “individual liberty maximizes happiness and utility”, and their opposition responds with “this problem is too complex to be solved individually”. What that contributes to is a story of the debate that revolves around a small government versus big government clash. That is the central issue the adjudicator wants resolved because it seems like the most important issue in the debate. Even if a logical analysis of the debate shows that issue to be only one equal part of the debate, the sense of its importance places it in the forefront of the mind of the adjudicator. That adjudicator is likely to preference the resolution of that issue over the resolution of other issues in the debate.

I think certainly the advantage of being the government team is you set the structure for the debate in a logical architecture kind of sense. The most common thing for every team to do is for their rebuttal points to just graft on to the method structure of the government team. Their first point was role of government so our first point of rebuttal is. Then their second point was how this affects women so that will be our second point of rebuttal and whatever. So you do get a huge advantage because most teams are too lazy or too time poor to change the frame, and so they’ve all just mapped themselves onto you where that seems like a reasonable option.

There are a number of important implications of this. Firstly, the relevance of a team to the debate is directly tied to how they are perceived in relation to that core narrative of the debate. This is because when we try to remember and understand information we are trying to give that information a coherent meaning in relation to other pieces of information we have received. When an adjudicator has already established the relationship between other pieces of information in the debate any new piece of information is automatically judged in relation to that older narrative. For an opening half team this can help them dominate the closing half by making sure that the narrative of the debate focuses on their own material. By showing that the core thread of the debate is about their arguments then closing half teams seem less relevant to the way the debate progresses. By creating a powerful and coherent narrative of the debate, the opening half has engineered an expectation from the adjudicator that that narrative will be addressed and resolved in each new speech, making it difficult for closing half teams to move away from that material.

Secondly, closing half teams need to be able to understand that that expectation is on them. If they do not feel they are able to extend adequately within that narrative of the debate they will need to be able to move the debate onto a different narrative. Introductions and points of information are particularly useful in doing this, as they serve as structural signaling points. If a team wishes to move the debate away from a clash on big government versus small government and move into a debate about , for example, consent, they need to understand how that will be understood in relation to the first clash in the debate. They need to spend more time justifying those new arguments in the debate than is strictly logical. In their rebuttal they need to show why the arguments form the opening half should be judged in relation to the issue of consent, rather than simply attacking their truth.

In narratology this is the difference between primary and supplementary events. A primary event is something that is considered core to the narrative, whereas a supplementary event is something that is tangential to the narrative. A supplementary event can have a relevance to the narrative, but it is not one of the key ideas the narrative turns on. The way we remember the events of a narrative, and assign importance to them, deprioritizes supplementary events in favor of primary events. Closing half teams in particular need to find ways of subtly placing their openings material into the position of supplementary events. People naturally do not like unresolved stories, and portraying the opening half clash as irresolvable is often a good strategy for moving the debate into another narrative that the adjudicator will inherently favor. Another tactic can be to play up a deficiency in the opening, making the adjudicators instinct for narrative resolution desire the filling of that gap. The point of this is that closing teams need to start manipulating the expectations of their adjudicator into wanting the particular contribution to the narrative of the debate that the closing team wants.

The third implication of a narratological reading of debates is that when the narrative coming from one team does not fit well with the narrative of the debate as a whole then that weakness should be exploited. If the core issue of the debate is clearly still about big versus small government while one team is talking about consent, their opposition can paint them as being irrelevant to the debate. They can use their language and how they focus their rebuttal to try and portray that team as being entirely focused on the issue of consent, even if they did address other issues in the debate. Similarly, if a team can successfully move the debate onto consent, they can re-enforce their opening as having a narrative entirely about small government. By doing that, even if opening had

material on consent, it will feel like a supplementary event from that team, and so the importance of that material is attached to closing.

Finally, when a speaker is constructing the narrative of their own case they should be doing so knowing that their opposition sill be attempting to control the narrative of the debate, and should attempt to predict what sort of clash is likely to become central to the debate. When the debate is intuitively about a big versus small government clash a team will find it easier to make that material central to the debate than arguments about consent. More pertinently, when constructing a nuanced narrative of their case they should do so in a way that engages with what is likely to be the core clash of the debate.

An un-nuanced case based on small government might be “Freedom is good”, but that only barely glances at the clash of the debate. An opposition is likely to be able to overcome that with a case constructed around the idea “freedom is good, but less important than social harm”. That more nuanced narrative doesn’t simply state a position, it also places that position in relation to the narrative of the debate. If the adjudicator leaves the room thinking that those are the ideas that sit at the heart of the two teams it is immediately obvious which one they would preference. Even before they analyze the logic of the arguments they are put in a position where they would need to be persuaded out of giving the win to the opposition. A better narrative for the small government side might be “This particular freedom is so important it can never be compromised on”, or “freedom creates the conditions for solving social harms, not government intervention”. These cases engage with the clash of the debate from the very first speech, and don’t give the opposition the advantage.

The process of understanding this in prep time is often very difficult, and is likely to be unique for each speaker. One way that can be effective is through simply acknowledging that this is a tool that is available to use for persuasion, and pushing yourself in prep time to have a larger coherence to the case that can be a strong narrative. Simply aiming for that conceptual clarity of narrative when writing and delivering arguments can give debaters the instincts to push for that narrative in their speech. A second technique is to imagine how you want the debate to end, instead of simply how you want it to begin. In three on three styles of debating, it can be helpful to imagine how you think the third speaker is going to approach the debate when they have to incorporate both sides of the case into a holistic speech. Doing this in prep time allows you to create material that deals with particular arguments in a strategic manner, instead of simply as units of logic.

Effective case construction is about understanding the status quo as a world view and understanding your vision of the world you want to live in, and then creating arguments that connect those two and give people a reasonable belief that we can and should transition in that direction, and I don't think logic alone gets you there.

The key difference between the simplistic and the nuanced versions of those narratives is that one incorporates an actual strategy to win the debate. It requires the team to understand what is at the heart of the debate and what is at the heart of their opposition's position. From there the team can think of a way to tailor their case so that they have an advantage when choosing and explaining their individual arguments.

Being able to position a case strategically requires understanding that a case builds into something greater than its logical components. Finding and exploiting that level of the debate is a crucial step in using prep time and speaking time most effectively for winning debates.

Narratological references

Abbot, H. P. (2008). Cambridge Introduction to Narrative. Cambridge: Cambridge University Press

Prince, G. D. (2004). Revisiting Narrativity. In M. Bal, Narrative Theory. Routledge

White, H. (1989). The Content of the Form. Johns Hopkins University Press

Gaulke, A. D. (2014) The Use of Narrative in Intervarsity Debate

Debating references

Sonnreich, T. (n.d.). First Principles. Retrieved from MAD Youtube: https://www.youtube.com/watch?v=aVRl8x_VlEw&list=UUsR0Bjprq48k8Saq6BLvdVA

Sonnreich, T. (n.d.). First Principles of International Development. Retrieved from MAD Youtube: https://www.youtube.com/watch?v=KnIU3S4yBto&list=UUsR0Bjprq48k8Saq6BLvdVA

Sonnriech, T. (2014, October 13). (A. Gaulke, Interviewer)

December 22, 2014 Andrew Gaulke Volume 12. 2014 Adjudication and Motions

Proudly powered by WordPress

Monash Debating Review

An annual publication of the Monash Association of Debaters

Setting Motions

Stephen M. Llano, PhD

The current movement among CA teams is to set motions that are not only interesting to debate but that attempt to educate as well. The general acceptance of videos, context slides, and information slides at nearly every tournament has liberated CA teams in their motion writing. Now that it has become a norm to provide additional information to debaters, the process of motion setting has become more imaginative, more creative, and clearly broader in scope. What was once seen as normal – looking at the daily news to set the afternoon's debate motions – now is considered lazy practice. CAs regularly set motions that focus on larger political theory and philosophy, and debaters are expected to use current events to fill in the gaps.

This is to be celebrated. It wasn't that long ago that CAs considered it appropriate to set motions based on fortune cookies. But as competition at Worlds has become increasingly competitive, the motions have followed suit. Once one could be witty, clever, and familiar with the past two days of headlines and win a lot of debates. Today one has to be much more familiar with larger trends in global affairs and the theories behind them to be successful in BP debate.

This change has some dangers. The most crucial is the risk that in our excitement to set deep, novel, and complex motions for debate we forget that debate in all aspects should be accessible to the reasonable audience. This link is what keeps BP relevant, valuable, and competitively fair. Debating should always maintain a familial relationship with public sphere discourse in some way in order to remain recognizable. Consider martial arts – a highly technical practice that appears mysterious from the outside. But placed within a real-

world context, martial arts is more than just making the moves for the approval for the master. It can serve as exercise, improving the health of the person, or it can serve as self-defense in dire situations. There are martial arts competitions held more frequently than debating competitions I would bet. And each of them preserve this balance between a fair and engaging competition that rewards the making of good moves while maintaining connection to relevance to the outside world.

I suggest a check on motion crafting that extends from the judging standard in British Parliamentary debating – the reasonable person. Although the reasonable person standard has been discussed frequently within this journal and other sources, it has primarily been considered a theory of judging. [1] I believe that the reasonable person standard should not be just for judging, but for judging motion quality as well. This extends the reasonable person standard to the ability of debaters to create arguments. I argue that the concept of the Universal Audience, created by Chaim Perelman and Lucie Olbrects-Tyteca in their work *The New Rhetoric* is the mechanism by which CA teams can check to ensure motions are set within the scope of the audience of debaters at the competition, avoiding the risk of setting a motion that although deep and interesting, might be inaccessible to those speaking simply because it isn't accessible to reasonable people.

Chaim Perelman and Lucie Olbrects-Tyteca write about argument inductively, finding the places and the means from which people generate argumentation in their daily life. Their theory is meant to help ground, expand, and improve what we might call "debate" – debates that happen in daily life as a matter of course. Within competitions we attempt to mimic this practice and create an art out of it suitable for competitive judgment. This art is often viewed as only a competition, meant to identify who is really good at it. At the same time, this competition is engaged in teaching a rhetorical relationship to the world, toward argumentation, discussion, disagreement, and toward how to engage other people about their ideas. This activity – which I call "debating" – is usually a mix of both of these ideals. Sometimes, "debating" is used to critique "debates" – what counts as good argument in the world is not viewed as such by debaters. What contemporary motion writing gets right is the idea that we should broaden our comfort zone about what we choose to debate about in order to ensure we are attentive to the entire world of potential controversy. What they get wrong is to sever this connection to public discourse nearly entirely, replacing it with their own form of the civic voice, or what seems "cool to debate." This results in two forms of motion that debating should do without.

Where Do Motions Come From?

When we consider the needs of a debating competition, motion setting is always at the top of the list. This is the opposite of reality where the decision of what to debate is what motivates the acquisition of a space, a time to meet, and an order of speakers as well as time limits or whatever other restrictions are necessary. In the world of debating, these concerns are dealt with first, and the CA team begins the motion conversation after they have been asked to serve.

I describe this process as anti-mimetic, meaning that it follows a pattern opposite its "natural" counterpart. Debating and debate have little in common beyond name, and what they do have in common could be described with the same ancient Greek word Aristotle used to describe the relationship between dialectic and rhetoric – *antistrophos*, or in the words of Jeffrey Walker, its "distant sister." He explains that the best way to view it is that, "the relation is one of systematic difference as well as similarity."[2]. Debate as a natural, public sphere phenomenon is related to but twisted away from debating, which is crisis and disagreement imposed from outside onto a group of people who have arrived precisely because they all agree that vehemently disagreeing on a few different topics for the weekend would be a great thing to take part in. They are sisters like Anna and Elsa from the film *Frozen*. The familial relationship is always present, but very distant in the way the two women engage the world.

The generation of motions leans toward the Elsa side – the creation of a world of controversy out of what is immediately present. Anna, in contrast, engages the world with what she finds in order to construct her engagement. Debating tournaments are like ice castles in the sense that they spring out of nothing, are really fantastic, and are unsustainable – their amazingness is possible due to their fragility. The competitors have not assembled to solve anything. They are not like debate attendees who are looking for a way to overcome an impasse. They are looking for rather exciting impasses to become involved in arguing about. The news is just one source, and not the best one, for the generation of debatable topics that would fit the situation. CA teams are under a lot of pressure to meet this need, and a good solution might be to put distance between the motions and the "real world." An Anna-style of motion setting would be to use what's available to solve the problem. Both have mixed results, and the film of course proves to us by the end that we are best with both approaches – a little magic and a little pragmatism.

Part of the problem with reaching this blend is well described by Chaim

Perelman and Lucie Olbrects-Tyteca in their discussion of the function of elite audiences. Perelman and Olbrects-Tyteca define the elite audience as an audience that believes the way it behaves should be a normative prescription upon all audiences. They confuse their way of thinking and believing as the norm toward which all audiences aspire. “The elite audience is regarded as a model to which men should conform in order to be worthy of the name: in other words, the elite audience sets the norm for everybody. In this case, the elite is the vanguard all will follow and conform to. Its opinion is the only one that matters, for, in final analysis, it is the determining one.”[3] Setting up a debate for the elite quickly becomes setting up a norm by which audience quality is judged.

CA teams can easily fall into believing that a motion is good because it is something the debaters “should know about.” Often this claim is ungrounded – rarely do CA teams point to a collection of literature that would be accessible and within the purview of those debating. Something that someone is writing or reading about for an advanced degree is often used as a motion with the defense that this controversy is current in the field, forgetting that most debaters do not have the ability to become familiar with that field. A paragraph on an information slide is insufficient to make debaters familiar with the controversy. Instead, speakers turn toward it as the grounding for proof instead of the grounding of the root of the controversy. The difference is between a good debate and one that the adjudicators wish they didn’t have to decide.

An example of this sort of motion was set at the Vienna IV two years ago. Before the debate, a rather long YouTube video was played that detailed how the U.K. bombed German cities after hostilities had ended in World War II. The motion was This House Believes That school children in the UK should be taught that their country engaged in war crimes. Although this is the start of a very stimulating discussion and debate, or possibly larger research project, it lacks important contextual elements that a debate should have – namely, it needs agreement on the controversy. Facts about the historical incident are not enough – to debate the motion at a depth that would be satisfactory one needs further insight. Why is this issue controversial? Who are the people involved in the discussion? The CA team believed that since people should know about this issue, it made for a good debate. What was missing was the in-depth reading, or access to debate arguments made in the world, that would indicate a number of starting points that inductively stem from the controversy. Instead, the information video and text is used as fact that becomes support for a deductive

argument about rights, state obligation, or the value and scope of education that is only tangentially related to the issue.

Another example of this sort of motion was set at Yale involving the practice of "bug chasing" where people participate in orgies with HIV positive individuals. Although the issue is worthy of reading about, controversial, and very novel, the lack of access to much of the larger controversy around it harms the debaters' ability to create arguments oriented toward a reasonable person. The surprise and shock of learning about such a practice would overwhelm the reasonable audience at first, as it would the debaters. Without access to the arguments that the practitioners might make in media to defend their choice to be "bug chasers," the debate will suffer from this lack of perspective. Again, debaters will be required to access arguments familiar to the context of the community of debaters not the groups involved in the controversy. The starting points for argument construction should be accessible.

Another concern with vanguard motion setting is the concern that because the CA team likes the motion and finds it really interesting, it passes the test for being a good motion for the competition. These motions are identifiable due to the lack of grounding in anything other than the opportunity for debaters to employ highly technical moves to access the tropes familiar to all those who debate. The motion, This House would randomly assign official first names at birth, suffers from a lack of a public sphere discussion entirely. The reasonable person, imagined as a member of the universal audience, would not recognize this topic as debatable, but more ridiculous. It would be hard for the reasonable person to see this as possibly controversial. The lack of conversation in the public sphere through accessible media make this topic hard to see as appropriately controversial, although it is clearly something that would be controversial if suggested.

These two main ideas – that reasonable audiences are the target of debaters' speeches, and that motions should be fair and accessible is not a new idea, in fact, it is the norm that we aspire to in designing our competitions. What should be clear from these two examples is that a better system of checking the quality of a motion is needed. Debatability and controversy are not enough if they are not provided within a larger context of accessibility to the debaters.

Grounding Motion Setting in the Universal Audience

The concern I have for the rift developing between debating's connection to

debate is rooted in a concern that our rhetoric is becoming overspecialized. “Argumentation aimed exclusively at a particular audience has the drawback that the speaker, by the very fact of adapting to the views of his listeners, might rely on arguments that are foreign or even directly opposed to what is acceptable to persons other than those he is directly addressing.”[4] People usually overcome this concern by attempting to offer arguments that they feel any reasonable person would find persuasive. Sometimes this takes the form of addressing a timeless audience of listeners, but we should realize that this audience is an imagined one, crafted from knowledge we have via experience about how people act and react to particular persuasive claims. Chaim Perelman and Lucie Olbrects-Tyteca identify the operation of this concern rhetorically as the Universal Audience. This is not an ontological universality – on the contrary, the Universal Audience is constructed based on concerns of context and culture. “Everyone constitutes the universal audience from what he knows of his fellow men, in such a way as to transcend the few oppositions he is aware of.”[5] The universal audience is made from the material concerns that come about from connection to society, culture, and institutions. One imagines the objections that situated people would make to one’s argument, and attempts to account for them.

The Universal Audience is the check that the rhetor uses to ensure that they are not overspecializing their argumentation. Adaptation to the audience is a good thing up until the point where the arguments work to exclude particular groups of people who the speaker may want to persuade, or more likely, groups that the speaker would like to identify with in order to make her argumentation more compelling to the immediate audience. This is the case in debating where the speaker attempts to link her argumentation at all times to the thinking of the reasonable person. “There can only be adherence to this idea of excluding individuals from the human community if the number and intellectual value of those banned are not so high as to make such a procedure ridiculous.”[5] That is, one cannot dismiss a large segment of the debaters as being ignorant because they could not debate a particular motion properly. An argument that is unconvincing might not be so because the majority of the audience is incapable of thinking. It is more reasonable to assume that the argument does not resonate with their experiences and thoughts. The same goes with motions – sometimes motions fail to produce good debates because they are not properly adapted for those who would debate them.

The use of the universal audience in motion setting would be for the CA team to think about the reasonable person standard away from judges and within the

context of argument creation. The central method of using the universal audience as a guideline is to make sure that there is enough context accessible to debaters to ensure that they can construct arguments for a reasonable person.

Reasonable Motion Setting: A Method

This process consists of three parts. First, any motion must be grounded in public deliberation. This means that there must be a test to see if reasonable, interested people could get access to a variety of sources of public debate on the topic. This is vital to access the rhetoric surrounding the controversy, which helps debaters ground their arguments within the realm of the reasonable person standard. This access should not be purely academic – the majority of reasonable people in the world do not have access to scholarly sources. Care must be ensured that there is not a lean toward such sources, considering most contemporary CAs hold advanced degrees or are studying for them. This test is most similar to the "Five Arguments" test that many CA teams employ to determine if side bias is present in a motion. This additional test of access is the same, but grounds the test outside of the competition, connecting it to the presence of such lines of argument in the public sphere.

Secondly, the team should ask if the discourse is recent enough to warrant setting the motion. CAs should check to see if the controversy is bubbling up in one form or another in ways that the reasonable person would notice. A motion could have a lot of things written about it, but if they are not circulating in current media, the reasonable person might not have an opportunity to access that controversy. There is solid and healthy conversation to be had by the CA team on this issue, as recency can have many meanings. Some topics, although not directly under robust discussion by public intellectuals or other media sources, are still things that can be assumed to be present, as they form the background of myriad arguments within states today.

One final check is related to pandering to the audience. Certainly, one should not set motions because one feels they are simple enough for debaters any more than they should set motions as a normative judgment on the quality of the debaters. There is no shortage on controversial, important, and vital issues for us to learn about and discuss. Motions should contain this spirit of the "push" toward broadening one's familiarity with the world, no question. But using this check of the Universal Audience, one might construct them as the opposite of the elite audience. This could lead to the setting of some motions

that are pedantic. How can this be avoided?

Perelman and Olbrects-Tyteca realized this might happen with their theory, since the universal audience is an imaginary judge over one's argumentation. To check against making the mistake of low-balling the average, reasonable person, one uses the undefined universal audience as a check. It is "invoked to pass judgment on what is the concept of the universal audience appropriate to such a concrete audience, to examine, simultaneously, the manner in which it was composed, which are the individuals who comprise it, according to the adopted criterion, and whether this criterion is legitimate."[7] Said another way, there are moments when a concern for accessibility might trump the presence of the actual audience, rendering them irrelevant – the arguments would appeal to a universal audience that might trump actual audience concerns or abilities. This is the moment where the CAs do a reality check, and make sure they are not overreaching in the direction of these concerns, and whether or not the debaters present can debate the motion at a quality level that preserves connection to the world while also delivering an engaging and fair competitive moment.

Let's test the motion, This House believes that the countries of the world should create and participate in a global carbon cap and trade system. The first thing the CA proposing this motion should do is some research – not about cap and trade and the arguments for or against it, but research to see where this issue is coming up in the debate world – media, public intellectuals, or other sources. This motion, like many, is unclear on this question. CAs can defend it being present due to the increasing public discourse on global climate change shifting from a *stasis* of conjecture to one of quality – "it's happening, so what should the response be?" This would be something the CA team should discuss to see if the public deliberation is suggesting this as a part of the controversy.

The recency question is also one that would need significant discussion, but if the CAs see the motion as a part of the larger discussion on global climate change, the answer is clear that this motion should be set. Passing this part of the consideration is often subjective, but checked by the CAs reminding one another that the reasonable person is also debating as well as judging – would the reasonable person find this issue controversial in a temporal sense?

Finally, the question of the undefined universal audience and that of pandering. In this case, this motion suggests a concern for meeting debaters exactly where they are. It is a debate about climate change, but also pushes

them to investigate cap and trade – something that is not appearing in the surface news sources that debaters might frequent – or it rewards those who have delved a bit deeper into the debate and not into the techniques of debating. A CA team concerned about the presence of cap and trade in the motion might choose to reword it to be about climate change – a clear trumping of the universal audience with the one that is present, and a move that could be considered pandering – keeping out the more complex argumentative possibilities over the fear that the debaters "won't get it."

Conclusion

Motion setting is the unenviable task of satisfying both one's ethical relationship to debating along with the obligation to provide the raw materials for an excellent competition. CA teams have taken on a mantle of prescribing not only motions that are good to debate about, but many motions that imply what issues debaters should be familiar with. Unfortunately, this normative push in motion setting turns debating inward, using itself as the metric of whether a motion is good for debating or not. This further isolates the competitive act of debating from real-world argumentation situations that I term debate. The debate/debating link should be preserved not only to tie value to debating, but to increase the quality of competition as well .The Perelman and Olbrects-Tyteca notion of the universal audience is the check that, if used by CA teams in motion setting, can bring more balance and less shallow debating based on information slides. The universal audience checks the motion to ensure that the reasonable person would consider this motion to be worth debating by asking if it is circulating in the collective discussion recently. It also checks CA teams from low-balling their audience at a tournament, and gives warrants to the normative push for inclusion of more complex or specialized terms in motions. Debating's value, as in martial arts, is in the application of complex moves both in the tournament and in the world. Without attention to preserving that connection, debating will become an irrelevant society of inward turned thinkers, performing what they think the vanguard will want to hear, ignoring the vast array of controversies present in the world at any given time.

1. As a starting point, see Bibby, Block, and Llano, Eds. *Adjudication: Essays on the Philosophy, Practice, and Pedagogy of Judging British Parliamentary Debate* (New York: IDEA Press, 2013). ↩
2. Jeffrey Walker, *Rhetoric and Poetics in Antiquity* (Oxford: Oxford University Press, 2000), 171 ↩

3. Chaim Perelman and Lucie Olbrects-Tyteca. *The New Rhetoric: A Treatise on Argumentation* trans. John Wilkinson and Purcell Weaver (London and Notre Dame: University of Notre Dame Press, 1969), 34. ↩
4. Chaim Perelman and Lucie Olbrects-Tyteca. *The New Rhetoric: A Treatise on Argumentation* trans. John Wilkinson and Purcell Weaver (London and Notre Dame: University of Notre Dame Press, 1969), 31. ↩
5. Chaim Perelman and Lucie Olbrects-Tyteca. *The New Rhetoric: A Treatise on Argumentation* trans. John Wilkinson and Purcell Weaver (London and Notre Dame: University of Notre Dame Press, 1969), 33. ↩
6. Chaim Perelman and Lucie Olbrects-Tyteca. *The New Rhetoric: A Treatise on Argumentation* trans. John Wilkinson and Purcell Weaver (London and Notre Dame: University of Notre Dame Press, 1969), 33. ↩
7. Chaim Perelman and Lucie Olbrects-Tyteca. *The New Rhetoric: A Treatise on Argumentation* trans. John Wilkinson and Purcell Weaver (London and Notre Dame: University of Notre Dame Press, 1969), 35. ↩

December 23, 2014 Stephen M. Llano Volume 12. 2014 Adjudication and Motions

Proudly powered by WordPress

Monash Debating Review

An annual publication of the Monash Association of Debaters

How (not) to Run Worlds: Advice from two people who needed it

When we agreed to serve as the chief adjudicators for the 34th World Universities Debating Championships (WUDC or Worlds) in Chennai, we expected to confront a wide variety of challenges – missing teams, significant delays, and even adverse dietary reactions were all within the realm of what we considered possible. The prospect that the tournaments judges would go on strike, however, was not a scenario we had entertained. Yet on the second day of the competition, we awoke to an email informing us that if independent adjudicators did not receive the travel subsidies they had been promised by the end of the day, they would refuse to judge the last three preliminary rounds and the elimination rounds of the tournament. Although the strike was averted, Worlds came dangerously close to grinding to a halt. When participants left Chennai on 4 January, the threatened judging strike and the numerous other problems meant that almost everyone saw the event as an organizational failure.

While it is comforting to treat Chennai as an aberration, its organizational difficulties were just an extreme case of a general problem. Many WUDCs have been marred by organizational shortcomings and failed to live up to their promises. The frequency of these organizational missteps is equalled by the frequency of the pledges by both WUDC hosts and the broader global debating community not to repeat the mistakes of the past. We should know: when we became the chief adjudicators for Chennais bid to host Worlds, we made many such pledges. During the bidding process and in the months leading up to the tournament, we vowed to improve the registration process, secure more reliable funding from sponsors, and house participants in a lavish hotel. We were aware that chief adjudicators and tournament organizers before us had frequently over-promised and under-delivered, but we were confident in our ability to oversee one of the most successful Worlds in recent memory. We were wrong.

So what happened? Why did Chennai Worlds fall so short of the goals we set? And why is Worlds often characterized by raised hopes at the outset and frustration during and after the tournament? This article attempts to answer these questions. Drawing on our own experiences, we reflect on some of the lessons we learned and attempt to shed light on how future hosts and the international debating community can avoid the problems that have plagued WUDCs.

What Went Wrong in Chennai?

Chennai Worlds contended with more than its fair share of organizational setbacks from tracking registration payments, to issues with getting participants visas, allocating hotel rooms, picking participants safely up from the airport, toilet paper disappearing, insufficient food provision, and dangerous dirt bike socials there are simply too many to discuss in a single article. Rather than present an exhaustive narrative of how the tournament unfolded, we have chosen to highlight a couple of incidents that illustrate some of the most serious difficulties. Unfortunately, describing some of the problems that occurred implies criticism of the institutions and individuals who organized the tournament (ourselves included). Many of these individuals worked incredibly hard and, despite the many challenges, contributed enormously to the successful elements of the tournament. Our purpose is not to disparage these individuals or otherwise point fingers nearly a year after the competition. Instead, our hope is that others will learn from our perspective and our mistakes.

Adjudicators Threaten to Go on Strike

The threatened judging strike was probably the most memorable organizational incident from Chennai. Those who were there likely remember the facts all too well. Briefly, however, here is what happened. Like all recent WUDCs, Chennai made a substantial amount of money available to help pay for experienced judges to attend Worlds as independent adjudicators. The amount in the budget for independent adjudicator travel (and for travel alone) was 40,000 euros. [1] The adjudication core made most of the decisions about which independent adjudicators to fund, and the travel subsidies we offered were quoted in euros. Aside from the figures in the budgeting documents, we never discussed the currency of reimbursement with the administration of

Rajalakshmi Engineering College (REC the institution that hosted Chennai Worlds).

At the tournament, RECs administration indicated that they wanted to pay judges in rupees, rather than euros. Their rationale was that they received income in rupees and it did not make sense to pay subsidies in a different currency. The administration also wanted to use the exchange rate that was prevalent in December 2012, rather than the one in December 2013. The rupee had depreciated during the intervening year, and so REC argued that the cost (in rupees) of providing 40,000 euros in subsidies had increased. While this was true, there had been no communication about this concern leading up to the tournament. Judges were understandably frustrated by these developments, and that frustration was made worse because the other organizational failures had already created an atmosphere of mistrust. Faced with the prospect of being underpaid in the wrong currency, and amidst a growing fear that they would not be reimbursed at all, the judges rightly, in our view, leveraged their role in the competition to force the College to pay the amount promised in euros. On day three of the competition, REC was able to pay the judges in full.

Although many participants correctly perceived that we, as chief adjudicators, supported the actions of the judges, we also bear some of the responsibility for what transpired. Confirming the precise details of the reimbursements with the administration should have been a priority, especially given that there were some early warning signs that judge funding could be an issue. In particular, the College seemed to encounter significant hurdles in trying to book flights for independent adjudicators. When we extended offers of travel subsidies to independent adjudicators, we gave them a choice: they could be reimbursed at the tournament, or in exceptional cases we could wire them the money or book a ticket on their behalf. Several adjudicators understandably opted for one of these latter options, and we began trying to facilitate the travel arrangements with REC. But despite months of back-and-forth emails between us, the adjudicators, and the organizers at REC, not a single independent adjudicator actually received a wire transfer or had a flight purchased on their behalf.

These difficulties should have been red flags. At a minimum, we should have been more transparent about our lack of control over the funding process, instead of continuing to pass on revised deadlines for when travel arrangements would be made. In one egregious case, we sent emails on five different occasions assuring an independent adjudicator that his flights would be taken care of that week or within a few days. Although transparency might

have jeopardized the willingness of some excellent judges to attend, it would also have facilitated increased pressure. By the time of the threatened strike, we had come to realize that outside pressure can be necessary to catalyse action.

From One Hotel to Three (and then Four)

A centrepiece of the Chennai bid to host Worlds was the hotel we had promised to secure for participants. The hotel, the ITC Grand Chola, was close to brand new at the time of the competition, and it is probably more over-the-top than any accommodation at a previous WUDC. Yet rather than accommodate participants at the Chola, REC assigned them to one of three hotels run by the still-luxurious, though certainly less so, Taj brand spread across several kilometers. Worse, more than one hundred participants were told there were not enough rooms at any of the three hotels for them. Tournament organizers had to scramble to find a fourth hotel that could accommodate these participants.

On one level, the administrations decision to change the hotel was understandable. REC did not raise as much sponsorship as they had intended, and the Taj Hotels cost less than 50% of what the college had budgeted for the ITC. But regardless of the financial wisdom of the decision, the change in hotel needed to be communicated to participants earlier, with an explanation as to why the ITC was no longer a viable option. The College, however, did not appear to appreciate that changing a key detail of the bid would cause participants frustration.

There certainly would have been less frustration if changing hotels had not been compounded with the failure to book enough rooms for participants. Candidly, we still do not know exactly how so many participants ended up without a room upon registration. Given that the College was in control of the finances, and it was RECs solvency on the line, we had very little insight into negotiations with the Taj hotels. We did not see how many rooms the college had booked, nor did we (the CAs and the student organisation committee) see the contract or have the opportunity to talk to the hotel before they signed.

Our experience suggests that hotels are an aspect of the bidding process where it is especially easy to over-promise and under-deliver. When participants arrive at Worlds, they will go where theyre assigned, regardless of whether the

hotel meets the promised specifications. Oversight is difficult because bidding institutions can claim that they are negotiating or have an agreement with a hotel which is hard to verify yet can also genuinely state that they will not (and should not) sign a contract with a hotel until after a bid is won and ratified. After ratification, little can be done to change a hosts decision about hotels.

Sources of Organizational Failures

From our perspective, two causes lie at the heart of the organizational problems at Chennai and other WUDCs: first, the host institutions lack of experience at putting on large debating competitions, and second, a misplaced belief that an experienced adjudication core can compensate for the hosts inexperience.
Institutional Inexperience

Within the past fifteen years, no institution has hosted Worlds more than once. And even if they had, the organizational team would likely have been vastly different the second time. To at least some degree then, each WUDC host has been unprepared for the responsibility. Worlds is too large an undertaking an institution has to be responsible for more than 1,000 participants for eight days to master every detail the first time around. But experience hosting large debating competitions matters. No matter how well-intentioned a host institution may be, overcoming a lack of familiarity with large debating competitions will prove daunting.

Prior to Chennai Worlds, neither the REC administration nor the key members of the local organisation committee had run a competition of any meaningful size. In fact, the debate program at REC was recently formed and participants had not attended many WUDCs. That kind of inexperience manifests itself in several ways in preparing to host Worlds. On a practical level, there is a tendency to underestimate the time and resources it takes to successfully run a competition like Worlds if an institution has not gone through a similar ordeal. For instance, the REC administration (although not the student organisation committee) believed that they needed no more than 40 volunteers to run the event. Similarly, the administration undervalued, in our view, the importance of conducting extensive practice runs in the days leading up to the competition.

On a less tangible, but perhaps more consequential level, an inexperienced host institution lacks the kind of intuitive familiarity with debating competitions

that only comes with years of participating in the community. There is a certain rhythm to debating competitions, and a set of expectations, that can be difficult to explain to individuals who have not spent many of their weekends during university traveling to IVs. For instance, anyone who had attended multiple WUDCs would probably have understood the value of housing participants in one hotel; the time participants spend interacting with debaters from across the globe back at the hotel is one of the highlights of Worlds. But from the perspective of an inexperienced institution, the downsides of using three hotels might seem worth the financial savings.

Relatedly, a familiarity with debating competitions can help host institutions understand and anticipate the kinds of sore spots that will most antagonize participants. Readers who attended Worlds in Botswana will remember the difficulties that the organisation committee had procuring meals that complied with some participants dietary needs. One preliminary round had to be delayed nearly two hours so that vegetarian attendees would not have to debate or judge on an empty stomach. Yet even more frustrating to some at the time was what many perceived as the organisation committees nonchalant attitude toward this failure; the committee seemed genuinely caught off guard by the participants strong reaction. Similarly, the REC administration appeared to us to be taken aback by the level of outrage over judging subsidies. While on the one hand these kinds of frustrations are relatively easy to anticipate not feeding participants in accordance with their dietary restrictions and failing to pay judges the full amount they were promised would strike most people as unacceptable we think a host institutions slow reaction time often reflects a gap in understanding that experience hosting competitions and greater exposure to the global debating community would fix.

To be fair, REC recognized that their lack of experience could be problematic, and they made a sincere effort to guard against the mistakes inexperienced hosts are prone to make. For instance, REC brought a large delegation to Worlds in Berlin. Senior administration officials, as well as 10 members of the organizing committee came to Berlin and made a genuine effort to understand the logistics of hosting Worlds. The organizers also put on what was largely considered to be a successful social, giving us and many participants confidence in how the tournament would be run. While the trip to Berlin was valuable – and something we would recommend for all future organising committees – in hindsight it just was not enough. And in some ways, the large College presence at Berlin and the relative success of Chennai Night were actually counterproductive. These experiences gave the REC administration

confidence that they understood Worlds and would be able to run the competition without much trouble. This confidence made the administration less willing to heed the advice and wishes of the student organization committee, external organizers, and the adjudication team.

Additionally, REC took the significant step of funding an external organisation committee to help handle logistics at Worlds. This committee consisted largely of experienced European tournament organizers, several of whom held senior positions in the Berlin Worlds organizing committee. Without this external organisation committee, it is questionable whether Worlds would have been run at all. However, as the College had never worked with the external committee before, the College was reluctant to trust them or allow them to make independent decisions. Meetings between the external committee and the administration often descended into shouting matches. While external experience is valuable, it cannot replace institutional knowledge unless external organizers have independent authority to make decisions. But host institutions will be understandably reluctant to hand over that kind of authority when it is their money and their reputation on the line.

Misplaced Faith in the Adjudication Core

When we were campaigning for Chennais worlds bid, several country representatives we spoke with expressed concern about whether REC could handle the organizational responsibilities. Our response was generally to sing our own praises even if REC was an inexperienced institution, we argued, the two of us would be actively involved in overseeing the preparations for Worlds. With our collective experience, the competition would run smoothly. We were wrong. Although we dedicated significant time and effort to following through on the promises we made about Chennai Worlds, we found ourselves far less capable of influencing the preparations than we had thought. Based on the conversations we have had with previous Worlds adjudication core members, this is a common mistake.

It is not just the adjudication core members themselves who over-estimate their influence on Worlds; the global debating community similarly places too much faith in a bids adjudication core. To some degree this is understandable the members of the adjudication core are often the most well-known and experienced individuals associated with a bid, and so they end up being the metric that participants use to calibrate expectations. But we hope this article

can help debunk the myth for future adjudication core members and participants alike that the adjudication core has control over the logistics of Worlds.

The principal reason this myth is unfounded is because final decision-making authority almost always rests with the host institution (or the organization committee). This was particularly true in Chennai, where the Colleges administration was actively involved in the organisation process, and therefore wanted its administrators to have the final say on all decisions. While the administration was occasionally happy to listen to our advice, we had limited ability to implement changes on our own.

For the influence we did have, we felt a need to marshal carefully. As the adjudication teams organizational influence is derived from the colleges willingness to listen, our tendency was to be diplomatic rather than confrontational. We felt there was a risk of poisoning the well with the administration if we attempted to micromanage from abroad, jeopardizing our working relationship before we arrived in Chennai. Avoiding that outcome meant relying on the representations from the organization committee and the administration, pushing back only when we felt it was necessary. In hindsight, we erred too far on the side caution. For example, we should have placed substantially more pressure on the college when it came to choice of hotels in the lead up to the competition. We should have pushed to see signed contracts, rather than accepting it will be signed soon as a sufficient explanation.

At some point, however, criticism and scepticism do more harm than good. Every adjudication core needs to be able to rely on the organisation committees representations and visa versa in order for meaningful collaboration to take place. Chief adjudicators should not, in our view, set themselves in opposition to the host institution. However, this further reduces the ability of a CA team to influence the running of the event.

Suggestions

The World Debating Community has a strong interest in not repeating the organisational problems that Chennai encountered. That is much easier to say than to do it is hard to completely eliminate the risk of an organisational catastrophe. We hope that some of the suggestions below can reduce that risk.

Improve the Bidding Process

In a perfect world, the best way to avoid the challenges that plague Worlds would be to more reliably select hosts that will put on excellent competitions. Unfortunately, it is very difficult to know two years out which of the promises in a bid will be fulfilled and which ones will not. Compounding the problem, institutions bidding to host Worlds have every incentive to promise the moon to secure the bid because, once theyve been selected, the debating community is locked into that bid. The costs of transitioning to another host or, in the worst-case scenario, cancelling Worlds, are simply too high to constitute a credible alternative. While we believe most promises during the bidding process are made in good faith, we also suspect that institutions would be far more conservative about what they promise if there were a mechanism to hold them accountable.

Without such a mechanism, our best advice to anyone who has a say in selecting a Worlds host is to decide on a bid based on the individuals who put it together, not on what those individuals promise to deliver. As best we can tell, this is the opposite of how most institutions and countries currently make decisions about which host to select. Cost, understandably, tends to be the first factor, but the lavishness of the hotel, the amount of free alcohol, and other similar perks are also high on the list.

To the extent that the individuals associated with the bid receive any scrutiny, that attention tends to fall on the named members of the adjudication core. This too is understandable as we have explained, members of an adjudication core usually have deep ties to the debating community, and they of course have a vital role to play in the tournament itself. But when deciding between bids, the quality of an adjudication core cannot meaningfully counterbalance a lack of institutional organizational strength.

The debating community should therefore focus on who the members of the organization committee are and why their institution is bidding to host Worlds. Key organizers should be prepared to discuss their prior experience putting on large events, how they plan to divide up responsibilities for the competition, and how they will interact with each other, their universitys administration, and the adjudication core. Members of the debating community should also ask tough questions about why an institution wants to host Worlds, what it stands to gain, and how it has demonstrated a commitment to debating. If an institution seems like it is prematurely vying for the chance to host Worlds, a

healthy dose of scepticism is warranted.

In some cases there is only so much that you can find out through questioning prospective hosts directly. It may be useful to allow other institutions on the prospective hosts national (or regional) debating circuit to comment on their perception over the hosts suitability – even if only privately. While that information may be biased and coloured by inter-personal relations, their comments may still be valuable.

Strengthen the Audit of Host Institutions

Arguably, the single worst mistake we made as chief adjudicators was failing to travel to Chennai in the months leading up to the competition. A visit to the campus, and in-person conversations with RECs administration and the organization committee, would likely have allowed us to spot several of the areas that would later become trouble spots. We could then have spent the months before Worlds trying to strengthen some of the key interpersonal relationships and focusing our efforts on the logistical hurdles that would prove to be the most problematic, such as increasing the number of volunteers. Yet even if we had made such a visit, and we had spotted problems in the making, our ability to correct them would have been limited by the need to preserve a good working relationship with REC.

The complicated relationship between an adjudication core and a host institution is one of the many reasons why we support the decision of Worlds Council to send a small team of independent auditors to evaluate a host institutions preparedness three to six months before Worlds. These auditors Councils resolution requires two or three must publish a report on their visit within two weeks of returning. In theory, this new requirement should provide the debating community with much-needed transparency. A hosts preparations for Worlds have generally been a black box, and attendees often do not know what to expect until they arrive. And unlike the adjudication core, institutional constraints should not limit the auditors ability to be critical.

We were pleased to see that the audit report for the Malaysia bid was candid about the deficit Worlds will likely run this year. But we were disappointed that the report came out more than a month later than the deadline set in the Council resolution (our understanding is that this was not due to any fault of the auditor). Conducting such an audit within the timeframe set by Council is

critical because that maximizes the leverage attendees can exert. Ideally, the audit report for future years will be published before attendees have submitted their last round of payments. As we saw first-hand in Chennai, sometimes outside pressure is necessary to catalyse action. We would also like to see the audit report cover more ground. Auditors should describe conversations with the host universitys administration, meetings with key third parties especially hotels and the status of important contracts. That kind of detail would empower attendees to apply pressure to the host on the issues that are most likely to flare up at the competition.

Pass on Organizational Knowledge

Many of the challenges we have discussed in this article could be avoided if hosting Worlds was something other than a one-shot game. If the debating community professionalized and monetized the responsibility of putting together the WUDC every year, we are confident that there would be a dramatic rise in the quality of the competitions organization. For now, that goal is unrealistic. As an alternative, hosts should look for ways to avoid re-inventing the wheel every year.

One way to do a better job of passing on organizational knowledge and experience is to treat organizational documents the way Worlds adjudication cores have come to treat adjudicator and debater briefings. Each year, an adjudication core starts with the previous years adjudicator and debater briefings and then makes the edits they see fit. Such a system provides for significant continuity the majority of the briefing remains unchanged while allowing for flexibility to add clarity to contentious issues or respond to changing norms on the international debating circuit.

Given that each host will face a unique set of organizational challenges (different numbers of attendees, different hotels, different costs, etc), there is clearly less room for continuity on the organizational side than there is on the adjudication side. That being said, there is no reason not to try to standardize certain aspects of running Worlds. Registration, both before and at the tournament, would be a strong candidate. The website and spreadsheets organizers use to keep track of which institutions have registered for Worlds should be passed down from organisation committee to organisation committee. The same goes for the spreadsheets organizers use to assign participants to particular hotel rooms and the process for checking in

participants at the start of each morning.

Worlds Council should require hosts to make these and similar documents available to future hosts. Admittedly, every host wants to put in place their own new system for improving how Worlds is run; that is how we felt, and we have talked with future hosts who have similar ambitions. But there is value in continuity Worlds will run more smoothly if repeat attendees are familiar with past systems, and participants will be well served if hosts avoid the temptation to test-run their ideas, like a brand-new check-in system, at Worlds. If future hosts can more easily implement a procedure that previous hosts have successfully used, the variance in the organizational quality of Worlds from year to year will decrease.

Conclusion

Deciding which institution gets to host Worlds will always involve a significant degree of uncertainty. A hosts motivations for bidding may be opaque, and the international debating community will never have perfect information about a hosts ability to live up to the promises in its bid. Athere is a risk that If so,

The experience of Chennai, coming so soon after Botswana, should ideally catalyse the international debating community to avoid this outcome. Although we have not discussed the challenge of drumming up more bids, we hope this article will help those tasked with voting on bids scrutinise bids more carefully. At a minimum, there should be an expectation that a Worlds bid prove itself by virtue of past organizational success, even if such a norm may be unrealistic in the immediate future. In the long run, even more is required. It is vital that the debating community find effective ways of monitoring the preparations that hosts are making and create mechanism to pressure organization committees to live up to their promises. Without such reforms, the frustrations participants experienced in Chennai, Botswana, and elsewhere will recur.

1. As a practical matter, this depreciation was offset by the fact that the adjudication core had only allocated 32,000 euros of our travel budget, rather than the 40,000 euros that was promised at ratification ↩

Monash Debating Review

An annual publication of the Monash Association of Debaters

Transgender exclusion in debating: A case for pronoun introductions

Crash Wigley

This article is a condensed and amended version of an article that was published online in June 2014, which can be accessed at googl.com/zxUgtM. In this piece I argued that the debating circuit should establish a policy in which debaters are asked for their name, speaking position and preferred pronoun at the beginning of all rounds.

The reaction of the circuit to the policy has generally been reassuring. Since the original article's publication, it has now become common practice in the IONA circuit to institute pronoun introductions at competitions, and most competitions in IONA have used the policy outlined in this document or variants thereof. Pronoun introductions have also been used in competitions on the US circuit, in continental European Competitions (such as the Ljubljana IV) and at EUDC 2014. I am grateful for this response – to be clear, I would not have been able to continue with debating if the situation had carried on as it was, and this was the driving force behind the original article – and as ever, welcome any questions or suggestions about the policy. In England, the National Universities Debating Council has set up a working group on the best way we can formalise pronoun introductions as a circuit (through formalising a national policy, sharing best practice etc.). Other national circuits (including Ireland and Scotland) have also formally instituted pronoun policies. In the UK, we are also developing a policy for introducing pronoun introductions at schools-level competitions, recognising the importance of supporting transgender young people and countering transphobia at school along with the unique challenges of communicating this information to teenagers.

For reference: I use the words **trans** and **transgender** interchangeably to refer to people whose gender doesn't match up with the gender they were assigned at birth, and/or who have genders other than female or male. This includes trans men, trans women and people with non-binary identities (e.g. genderqueer people, gender neutral people etc.). I use the words **cisgender** and **cis** to refer to people who aren't transgender.

The deal with misgendering

During debates speakers use gendered language to talk about the other people in the debate, whether that's saying 'sir' or 'madam' when offering a POI, or using pronouns to refer to what previous speakers have said (e.g. "Speaker X said this, but what he doesn't understand is…"). Similarly, judges refer to speakers using gendered pronouns during the deliberation, when calling speakers up to speak and when explaining the call and giving feedback.

In debates, trans people put themselves at risk of being publicly misgendered (i.e. referred to by the wrong gender) which can be upsetting. It is unpleasant if you are, for example, a transgender man and people repeatedly call you 'she' or 'madam' during a debate. It can make people feel self-conscious, or like their gender isn't being respected. This is so prevalent that without pronoun introductions I would expect to be misgendered at every debating competition I attend. This makes debating an exclusive space that discourages trans debaters from participating.

In most cases, people would like to know how to avoid being making these mistakes. Even in the private setting of the judges' deliberation, judges who care about respecting trans people's genders should want to know how people would want themselves to be talked about. It therefore should become standard practice to do pronoun introductions at the start of each round in debating competitions. It is already common in many trans circles for people to say what pronoun they prefer when introducing themselves, and it is generally considered polite to ask for somebody's pronoun preferences if you don't know.

How pronoun introductions work

At the start of debates, chairs already have to find out which order the speakers

in each team are speaking. When doing this chair judges should also ask which pronouns each speaker prefers.

This is the sort of exchange that occurs:

Chair: So can I check, who is speaking first for opening government?

Kate: Kate.

Chair: And what is your preferred pronoun?

Kate: She.

Chair: And speaking second?

Crash: Crash, and I don't mind being called either they or she.

Everyone makes a mental note of this information, and then the chair proceeds to ask the speakers on the other teams in a similar and polite fashion.

Before or after this spiel, the chair or panel might also want to give their names and pronouns. Individuals may choose to specify a pronoun (such as she, he, they or any alternative pronouns) or to say that they don't mind or have no preference.

If somebody says they want to be called 'she' or 'he', it is fair to assume that they are also happy to be called 'madam or 'sir' or 'Madam/Mister Speaker' accordingly, unless they say otherwise. However, if somebody asks to be called 'they', it is sensible to avoid using gendered terms altogether, and finding gender-neutral replacements (e.g. saying 'On that point' rather than 'sir' to offer a point of information).

Two things are important. Firstly all speakers should introduce their names and pronouns to the entire room rather than just write them on the ballot, so that all other speakers know. Secondly, these introductions should happen in all rooms. If you're cis (i.e. not trans) and don't get misgendered on a regular basis you might think this is unnecessary for you, but unless this happens for all speakers in all rooms then it puts a lot of pressure on trans debaters to personally request to introduce their pronoun in each of their rooms. That can be intimidating and make people feel unwelcome. It requires trans debaters

who want to be referred to in a specific way and are at risk of misgendering to effectively repeatedly identify themselves as transgender and ask for special treatment. It pushes the burden on transgender people to make themselves the odd one out, rather than recognising that in this circumstance we don't need to make assumptions about anyone's gender, and can create a space where everyone can self-define rather than be labelled.

All speakers and judges should listen carefully to which pronouns people prefer, and endeavour to use those in their speeches, and in life more generally. If speakers realise they've made a mistake, the best response is to quickly apologise, correct themself, and then move on. If a speaker doesn't realise they have misgendered somebody during their speech, the chair's place should be to remind speakers to use the pronouns that other speakers have asked to be referred to by at the end of the speaker's speech. People should not be referred to the Equity Team for accidentally using the wrong pronoun provided they apologise if they make a mistake. Pronoun introductions remove the need for cisgender speakers to guess the preferences of speakers in their debate, and so make mistakes less common.

This system has the added benefit of judges and other speakers knowing each others'names during the debate, and allows judges to listen to how individuals pronounce their own names.

Competitions who want to introduce this system will need to explain it to all speakers and judges. It is important that people understand the reasons why pronoun introductions have come about, to stop it from becoming something we do 'just to be politically correct'. That said, in many ways the system is straightforward, and it is sensible to not to make more out of it than needs be. On top of all this, in the separate judges'briefing, the system of asking at the start of the round should be explained, as well as the importance of asking in every round. If individuals make fun of or mock the system of pronoun introductions, they should be referred to the equity team – if trans debaters are to feel welcomed and not just 'tolerated'people need to take pronoun introductions seriously.

Potential concerns and alternative policies

In discussions, many people have raised concerns about the effect of this policy on people who aren't comfortable making public declarations about their

gender for whatever reason. In this context it is important to note that pronoun introductions are not an affirmation of gender or identity – they're an instruction about how you would prefer others to talk about you in a specific context. Pronoun introductions allow people who have a complicated relationship with gender to experiment with different pronoun use in a respectful space as they feel comfortable. Furthermore, it is perfectly legitimate for individuals to reply that they do not mind or have no preference if this is what they prefer.

In addition, fears that people would feel like they are 'betraying themselves' by asking to be referred to by 'birth gender' pronouns when they're closeted are overstated. These people (whose situation I have been in myself) live in a world where they constantly have to make decisions where they present themselves in a way that doesn't match with their identity to protect their own security. They are best placed to make these decisions for themselves. What this system does do is give people more control over the language that other people use to describe them, and that's helpful for people who have a complicated relationship with gender. Finally, being closeted when trans is unpleasant full stop. To me, it is a greater priority that the circuit can be seen to be valuing, welcoming and supporting the participation of out trans debaters, and in doing so it can increase the confidence of people coming to terms with their gender identity. This policy does not 'require disclosure'. It merely recognises the fact other speakers are going to have to refer to individuals by some pronoun, and gives them the opportunity to have a say in which one they choose.

At Zagreb EUDC 2014 Council, additional concerns were raised about the effect of this policy on ESL speakers.[1] Nevertheless, while the circuit must be reactive to the needs of ESL speakers and the background of individuals from different circuits when considering how this policy should be introduced to tournaments by equity teams, and while recognising speakers' different backgrounds in English might encourage participants to be more tolerant of mistakes made by speakers (especially if an individual has requested people to use a pronoun other than 'he' or 'she'), ESL status should not be a carte blanche for speakers to ignore individuals' preferences about their gender. That would have the effect of maintaining a debating circuit that is exclusive of trans debaters (many of whom will be ESL).

The alternative policy suggested by the Equity Team was to promote the use of gender-neutral language.[2] This policy is insufficient. Firstly, it expects that speakers within debates will call other speakers 'they' rather than 'he' or 'she'

to avoid misgendering. Nevertheless, this is a much harder norm to enforce and to inform people about, as unless people are being very consciously listening out to the pronouns used, uses of 'he', 'she', 'sir' or 'madam' will go unnoticed. As a result, at best this policy would result in '*calling everyone who looks like they abide by conventional sex-gender binaries he or she, and anyone who looks trans, queer or otherwise non-conventional 'they'*". Furthermore, the main concern expressed by ESL speakers during EUDC was that it was harder to use 'they' as a singular pronoun as its use in that context in English is rare, and so it doesn't overcome the concerns about language difficulties.

Consequently, such a policy would continue to make debating exclusive for many trans people. It would ignore people's preferences to be referred to as 'he' or 'she' and not 'they'(which is important given that many trans people will have fought for the right to be respected *as a man* or *as a woman*). Indeed, it could single out trans people to such an extent that they end up being the only ones referred to as 'they', implying that trans women are not seen as 'real women' and trans men are not seen as 'real men'. It would also disrespect the preferences of 'conventional-looking' speakers who might strongly prefer the use of a pronoun that wouldn't be anticipated. Finally, because of the concerns about norm-enforcement, trans debaters would have little confidence that participants were even going to call them 'they', rather than use inappropriate gender-specific pronouns. Recognising these issues, EUDC Council implied that it would be happy for individuals with such specific preferences to announce them to the room at the start of debates, a so-called 'don't ask, do tell policy'.[3] This was effectively the status quo when I decided to first advocate for pronoun introductions, and it was what motivated me to do so – otherwise I would not have carried on debating. That policy permanently forces additional burdens onto transgender people and makes debating exclusive, as explained previously. As a result, requiring all speakers to say at the start of the debate their name, their speaking position, and which pronoun they would like others to use when speaking about them is an undemanding solution that is needed to welcome trans debaters.

The Bigger Principle

There are many reasons to think that there are advantages to pronoun introductions in everyday life. Whenever you're asking a group of people to tell each other their names you could ask them to give their preferred pronouns. It takes the pressure off trans people to individually tell everyone how they want

to be talked about; it makes cis people aware of the potential existence of transgender people and it shows that we care about how people identify rather than just applying whatever label we decide fits best. In the same way that people often choose their own name or nickname, we care about giving people the authority to decide how they are referred to by other people, and this is an extension of that principle.

Pronoun introductions are especially useful for people who are in the process of transitioning, or have a non-binary gender or an unusual gender expression because even going on names or how people dress isn't going to be sufficient to work out people's preferences. They also give people who might want to experiment with, for example, being called 'they' the opportunity to do so without it having to be such an intimidating step. They also provide space for people to define their relationship to gender in other ways (e.g. some people like being referred to as 'they' for ideological feminist reasons, because it limits the extent to which we are gendered unnecessarily).

As such, I would like pronoun introductions to proliferate more generally in society and I hope that our debating circuits can be a bit pioneering in this respect. I strongly believe pronoun introductions are also a bare minimum needed for at least most trans people to feel in any way comfortable at debating competitions.

Recommendations

Competition organisers (CA teams, equity teams, convenors) should put in place pronoun introductions in their competitions, as set out here. Information about the policy should be given in a such a way that firstly, participants understand the reasons for the policy and know what is expected of them and secondly, the wellbeing of trans debaters and trans people in society is considered. (i.e. if organising teams explain the policy poorly, the policy could make debating a less rather than more welcoming place for trans debaters).

Debating societies should use pronoun introductions where appropriate at internal events (e.g. when new members will be meeting each other for the first time). This will make debating more inclusive of trans debaters from the ground-level, and will familiarise debaters with the concept before they attend competitions.

1. Zagreb EUDC 2014 Council Minutes, p. 8. ↩
2. Ibid. p.7. ↩
3. Ibid. pp. 8-9. ↩

December 23, 2014 Crash Wigley Volume 12. 2014 Commentary and Critique: the Functioning of Tournaments

Proudly powered by WordPress

Monash Debating Review

An annual publication of the Monash Association of Debaters

An Evaluation of Four-Team-Per-Contest Swiss (Power Paired) Tournament Structures Using Computer Models in Python

Neil du Toit

Abstract

In this paper we present the results of an analysis of the structuring of four-team-per-contest, Swiss (power paired) / elimination tournaments. We create models for teams and tournaments using Python. Team scores are sampled from normal distributions. We estimate the mean and variance parameters of the distributions based on a statistical analysis of the tab of the Chennai World Universities Debating Championships 2014. We provide a discussion of the appropriate methodology for selecting evaluation measurements. We then provide an overview of the more common measures of rank correlation, and rank disorder. We run one thousand iterations of the model of each tournament structure. For each model, the iterations are performed once under the assumption of no team variance, and once using samples from the distributions. The results provide accurate estimates for the population means of the chosen metrics. The no variance iterations isolate the inherent fairness, and suggest the inherent competitiveness, of the tournament structures. The iterations with estimated parameters suggest how fairly the tournament will perform in real world applications. By comparing the performance of the tournament structures, we suggest answers to the following questions: Which bubbling procedure is most fair? Which intra-bracket match-up procedure is most fair? How many rounds should a tournament have? How many randomised rounds should a tournament have? How influential are these

decisions on the competitiveness and fairness of a tournament? How fair and competitive are power paired tournaments?

Introduction

The Swiss tournament system was first used in a chess tournament in Zurich, in 1895. Since then, FIDE(the World Chess Federation) has officially recognised five different Swiss tournament structures.[1] Originally,preference was given to ensuring board fairness (the equivalent of ensuring that debating teams speak in each position a similar number of times). Over time, more emphasis started being placed on ensuring competitor fairness.

The Swiss tournament structure has a number of attributes that make it an incredibly desirable format for debating tournaments. It can be completed in significantly fewer rounds than a Round Robin. Round robin tournaments also pair up the weakest teams against the strongest, which can be undesirable. In comparison to elimination tournaments, the Swiss system has the advantage of allowing everyone to compete inall of the (prelim) rounds.

However, in porting the Swiss system to British Parliamentary debating, new problems have been introduced. The fact that ordering problems exist is common knowledge. We do not, however, know the severity of the problem; and the precise nature of its causes is often confused.

In this paper we will be looking at the prelim stages of a Swiss/elimination tournament. The fact that there is an elimination phase after the prelims, is important in so far as it requires us to look at the 'break' ordering of the prelims. We will not, however, be looking at the elimination stage *per se*. We begin by investigating some of the more relevant differences between chess and debating, and the problems which they cause.

Causes of Power Pairing Failure

The Monotonicity Problem

One of the FIDE rules, which apply to every tournament structure, is that no two players may face each other twice. The primary reason for this is rather simple: if teams repeatedly face each other, then they will be taking too many

points off each other. The result is that lower ranked teams can easily 'catch up', and,therefore, the difference between teams on the tab will no longer have any relation to the true difference in skill between the teams. In effect, the tab is "compressed".[2] Again, it is well known that the middle of the tab often 'catches' the top rooms, when teams in the top rooms are constantly taking points off each other. This may appear to make the tournament more exciting, but it is not too difficult to imagine the injustice that may result. Consider a team just below the break who repeatedly catches a team just above it. The higher rated team may win several of these encounters. However, the lower rated team only needs to win once(at the end), in order to break, above the better team. This is the case no matter how much better the higher rated team actually was. By forcing monotonicity, chess tournaments allow the gap between teams to widen, until they reflect the true difference in ability between the teams.

The Stability Problem

Arguably the most serious concern with power pairing, is its instability. This refers to the fact that the tab doesn't converge to any particular order. After settling, it fluctuates, quite significantly, around the correct order. This is true even in the complete absence of any upsets. The source of this problem lies in the fact that four teams compete in each BP debate. By awarding more than one point for a win, power pairing enables teams to 'jump' over brackets, without having had to face any team in that bracket. By 'bracket',we mean a group of teams in a tournament that are on the same number of points (the WUDC Constitution uses the word 'pool' to mean the same thing). For example, should one bracket have 4 teams on n points,and another bracket have 4 teams on n+1 points, 2 of the teams on n points will end up above a team who started on n+1 points. Necessarily.

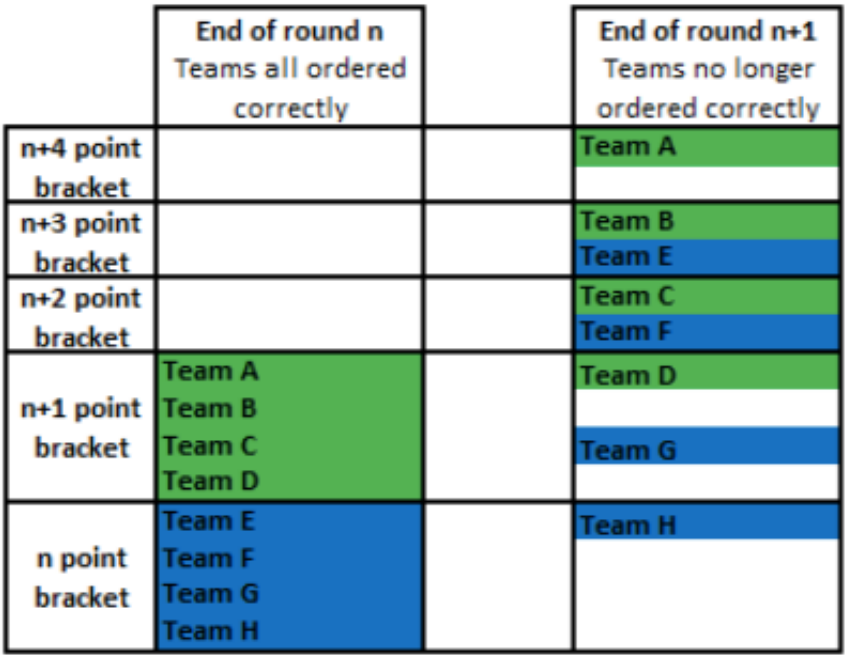


Figure 1: An illustration of the stability problem.

In the initial stages of the tournament, this is not much of a problem. However, once a portion of the tab has settled, then running another round will disorder that portion. Since different parts of the tab settle at different times, a significant portion of the tab is always going to be getting more disordered each round. The result, is an upper limit to how ordered the tab can get, before the order starts fluctuating, and the degree of order levels off towards an asymptote. This does, of course, only happen when teams are close together. The monotonicity problem described above ensures this compression. In that sense, these two problems are mutually re-enforcing

Creating A Metric for The Disorder of Tabs

There is a considerable amount of literature on the subject of disorder, and several definitions from which we can choose. As Paul Collier notes,[3] one should always try and use criteria set by other researchers, so as to avoid the temptation to define your hypothesis to be correct. None the less, many measures of disorder are inappropriate for tournaments.[4] We therefore offer a brief discussion of how we selected our criteria

Interpretability

In this paper, we are primarily interested in practically significant differences. To that end, we would like our metrics to be interpretable. That is to say that

they should have an obvious meaning. Measures such as the Kendall τ coefficient and Goodman and Kruskal's gamma, are popular, and well suited to hypothesis testing. However, they don't offer any insight into the absolute disorder of a list.

Robustness

In tournament evaluation, we must fully account for outliers. If a team is severely disadvantaged by a tournament, it will be no consolation to the team that this was a rare event. Tournament structures need to be designed to ensure that *every* instantiation meets some minimum criteria of fairness (in the absence of variance attributable to the teams). Therefore, we use some statistics which are noticeably volatile.

Consideration for The Break

It is again well known that power pairing has a preference for the extremes. That is to say that the top and bottom few teams in each tournament will be relatively better ordered than the middle. An important part of the preliminary stages of a tournament is the ranking of the break teams. This is unique to our purposes,and traditional measures of disorder will not take this into account

Metrics Used

Preliminary Definitions:

- •A team's "rating" is where they should have placed in the tournament
- •A team's "ranking" is where they actually placed
- • "The break" refers to the top ranked 16 teams

Measurements on The Entire Tab

Spearman's Footrule Distance: Spearman's Footrule Distance is the sum of the differences between the ratings and the rankings of the teams[5]

Spearman's ρ Distance: The Spearman's ρ distance is similar to Spearman's footrule, however, it exaggerates outliers, by squaring the distances before

summation[6].

Measurements on The Break

Measurements Relating to The Correctness of The Break

Break upsets: Break upsets is the number of teams that should have broke, but didn't. Equivalently, it is the number of teams that shouldn't have broke, but did.

Break-loser: The break loser is the top rated team to not break in a tournament. Ideally this is the team rated 17th. If the number is significantly lower than this, then it will indicate that a strong team has been severely disadvantage in that tournament

Measurements Relating to The Ordering of The Break

Spearman's Footrule Distance on The Break: We re-rate the teams who have made the break from 1 to 16. The Spearman's Footrule Distance is then calculated as normal.

The Models

How the Models Calculate Intra-Bracket Match-ups and Bubbles

Bubbling

Bubbling is the procedure whereby the tabbers adjust the brackets in order to make each bracket consist of a number of teams that is divisible by four. The WUDC Constitution, Art 30(3)(c), states that "If any pool (The Upper Pool) consists of an amount of teams equivalent to a number that is not divisible by four,then teams from the pool ranking immediately below that pool (The Lower Pool) may be promoted to the Upper Pool…" This is a somewhat cumbersome provision.[7]. However, it is clear that bubbling must consist of 'pull ups'. We investigated three different ways of selecting the teams from the lower bracket that need to be bubbled up:

- **Low**: Teams in the lower bracket will be bubbled starting from the bottom, in terms of rating. See Figure 2.

• **High**: Teams in the lower bracket will be bubbled starting from the top, in terms of rating

• **Random**: Teams in the lower bracket will be bubbled randomly. Note that this is currently what the WUDC Constitution requires

The WUDC Constitution, Art 30(3)(d) provides that “Once the pools have been adjusted in accordance with 3(c) then the pools are divided into debates of four teams”. We investigated three different ways in which this can be done:

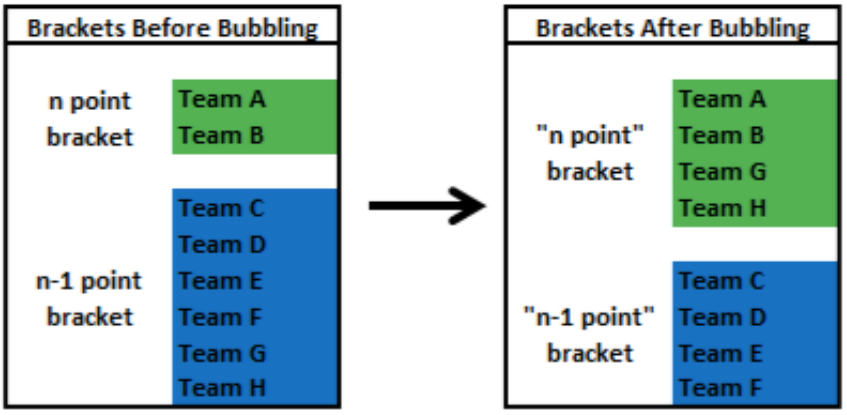


Figure 2: An illustration of low bubbling.

• **Splitting:** Teams in a bracket will be paired in the same way as teams 17th to 48th in Art 30(5)(b)of the WUDC Constitution, adjusting for bracket size. See Figure 3.

• **High-High**: The top four rated teams in bracket will form a room, continuing as such through thebracket

• **Random**: Teams in a bracket will be paired randomly. Note that this is currently what the WUDC Constitution requires.

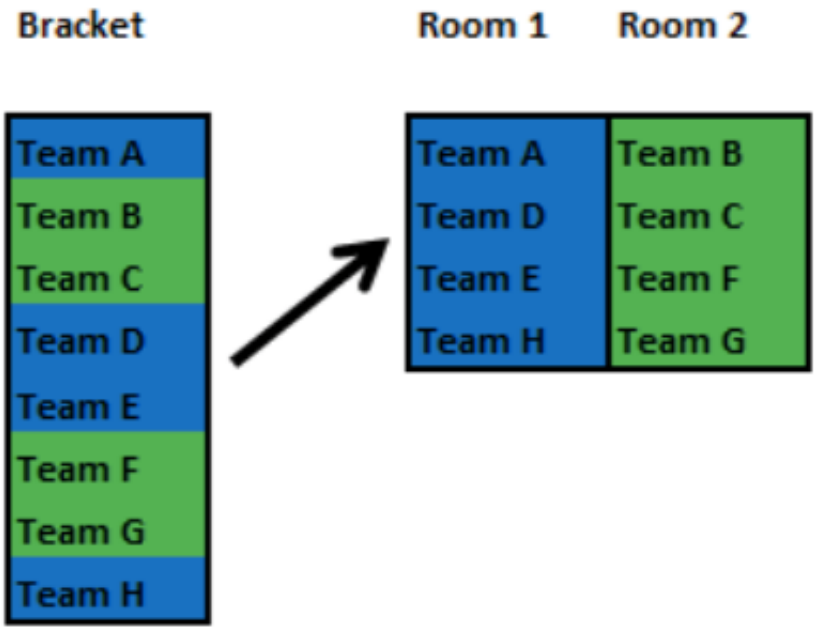


Figure 3: An illustration of splitting brackets.

The Model Without Upsets

We rate 366[8] teams from 1 to 366. The top rated team in every room will always win, followed by the next highest rated team, and so on. The only differences in outcome, over the various iterations of the model, are due to the first round, which is completely randomised, as in Art 30(2)(g) of the WUDC Constitution.

Model With Upsets

An upset is any debate result in which a team places higher than a team which was ”better” than them. This happens when variance is introduced to the team’s performances. We wish to investigate how well the different tournament structures tolerate variance. I.e. are the results still reasonably accurate, when a couple of upsets occur? It must be stressed, however, that too much weight should not be afforded to these results. Teams have been modelled based on the 2014 Worlds tab, with the average speaks of each team in each round being used to estimate the mean and variance of each team’s speaks.[9] From these populations, the model will sample scores for those teams in each round. Unlike in the case of no upsets, we now have to determine what the ratings of the teams ”should” b. This question is not a trivial one. The first question is whether the estimated population means, or the actually attained means, should be used for comparison purposes. We decided that the latter would be more appropriate. An iteration of the model which samples,on

average, higher or lower than the population mean, models a case where a team performs better or worse at that given tournament than they would normally be expected. It is only right that they should thus place higher or lower, respectively. The second question is more tricky. It concerns whether a team with a lower mean score than another team might actually be "better" in some sense than that team. The best estimate,of course, is that they are not. However, see section 7.5 for a discussion of this problem.

Results

Figure 4 and 6 tabulate the results of the different bubbling and pairing[10] procedures, for a 9 round tournament with 1 random round. Figure 4 shows results without upsets, figure 6 shows results with upsets. Figures 5 and 7 are plots the five number summaries of the Spearman's Footrule distance results from tables 4 and 6. "High-High" pairing permutations are on a separate axis, because the values of their five point summaries are all orders of magnitude above those from the summaries from the other permutations. Figure 8 shows how the mean of the Spearman's Footrule distance decreases as the number of rounds in a tournament increases. The tournament structure used is one with random pairing and bubbling. Note how the graphs level off towards an asymptote. Figure 9 shows the effect on the mean of the Spearman's Footrule distance, when more rounds in the tournament are allocated to completely randomised pairing (as in Art30(2)(g) of the WUDC Constitution). The total number of rounds remains constant, at 9. The tournament structure used is one with random pairing and bubbling.

Measures of Disorder Results From 1000 iterations of 9 Round, 1 Random Round Power paried Tournaments, with No Upsets in Any Debates

		Spearman's Footrule			Spearman's Rho			Break Upsets			Break Loser			Spearman's Footrule On Break		
Pairing Procedure	Bubbling Procedure	Mean Obtained Value	Minimum Obtained Value	Maximum Obtained Value	Mean Obtained Value	Minimum Obtained Value	Maximum Obtained Value	Mean Obtained Value	Minimum Obtained Value	Maximum Obtained Value	Mean Obtained Value	Minimum Obtained Value	Maximum Obtained Value	Mean Obtained Value	Minimum Obtained Value	Maximum Obtained Value
Splitting	Low	2.87E+3	2 404	3 086	4.08E+4	30 976	46 168	2.00E+0	2	2	1.50E+1	15	15	2.03E+1	20	26
	High	3.38E+3	2 554	4 108	5.84E+4	36 218	82 418	2.02E+0	0	5	1.10E+1	4	17	2.28E+1	4	44
	Random	2.99E+3	2 264	3 494	4.67E+4	27 438	66 052	1.07E+0	0	3	1.46E+1	7	17	2.76E+1	6	54
High-High Pairing	Low	2.43E+4	23 822	24 880	2.81E+6	2699316	2928342	5.00E+0	5	5	1.20E+1	12	12	3.40E+1	34	34
	High	2.46E+4	24 076	25 210	2.84E+6	2719312	2987390	8.61E+0	6	10	6.83E+0	6	10	5.02E+1	32	62
	Random	2.45E+4	23 840	25 192	2.84E+6	2719246	2990608	6.07E+0	4	8	1.01E+1	5	12	4.18E+1	18	68
Random Pairing	Low	3.16E+3	2 536	3 752	5.83E+4	41 700	82 412	1.51E+0	0	4	1.47E+1	11	17	2.25E+1	6	36
	High	3.53E+3	2 854	4 240	7.07E+4	45 704	107 176	1.88E+0	0	5	1.14E+1	4	17	2.73E+1	2	64
	Random	3.35E+3	2 682	4 132	6.55E+4	46 200	97 656	1.36E+0	0	4	1.40E+1	7	17	2.95E+1	4	54

Figure 4: Tabulation of bubbling and pairing procedure results, without upsets. Note that the expected values of Spearmans Footrule, Spearman's Rho, and Spearman's footrule on the break, are 3.76E+4, 6.32E+6 and 8.50E+1 respectively(J Marden *Analysing and Modelling Rank Data* (1995) page 77). A perfect tournament has no break upsets. The break loser in a perfect tournament would be 17.

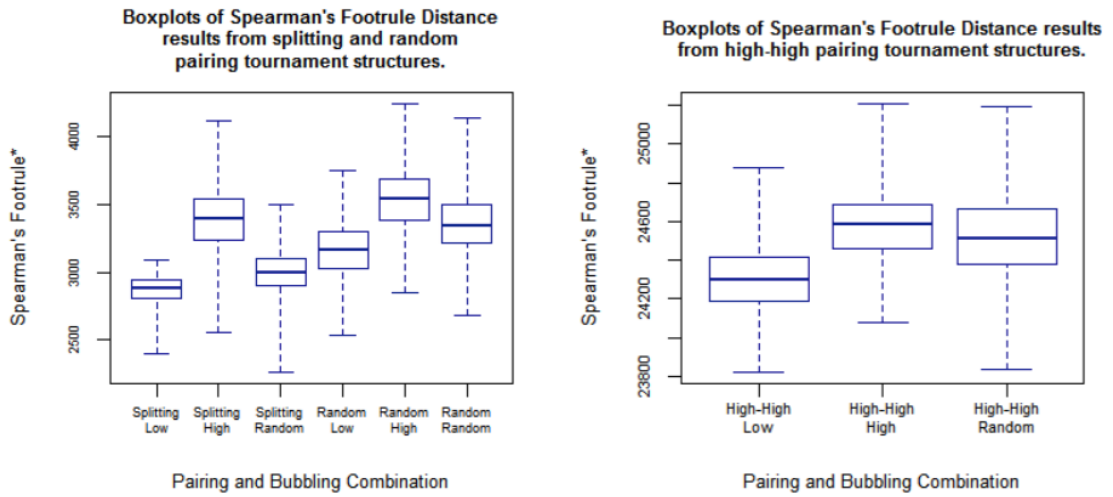


Figure 5: Box and whisker charts of Spearman's Footrule result, for different bubbling and pairing procedures, without upsets. ⋆ Note that vertical axis does not start at zero.

Measures of Disorder Results From 1000 iterations of 9 Round, 1 Random Round Power paried Tournaments, with Scores Sampled From Normal Distributions with Estimated Team Means and Variances

		Spearman's Footrule			Spearman's Rho			Break Upsets			Break Loser			Spearman's Footrule On Break		
Pairing Procedure	Bubbling Procedure	Mean Obtained Value	Minimum Obtained Value	Maximum Obtained Value	Mean Obtained Value	Minimum Obtained Value	Maximum Obtained Value	Mean Obtained Value	Minimum Obtained Value	Maximum Obtained Value	Mean Obtained Value	Minimum Obtained Value	Maximum Obtained Value	Mean Obtained Value	Minimum Obtained Value	Maximum Obtained Value
Splitting	Low	5.88E+3	4 358	7 116	2.00E+5	114 282	289 176	3.08E+0	0	8	1.02E+1	2	17	4.07E+1	6	76
	High	6.45E+3	5 000	7 820	2.39E+5	145 496	338 600	3.89E+0	0	9	7.66E+0	2	17	4.96E+1	8	84
	Random	6.09E+3	4 568	7 366	2.14E+5	126 508	304 922	3.25E+0	0	7	9.48E+0	2	17	4.44E+1	10	80
High-High Pairing	Low	2.08E+4	19 118	22 210	2.12E+6	1806400	2405186	7.47E+0	4	11	6.07E+0	2	12	4.93E+1	24	88
	High	2.10E+4	18 966	23 068	2.17E+6	1865470	2563062	7.97E+0	5	11	5.38E+0	1	10	6.07E+1	32	90
	Random	2.09E+4	19 198	22 554	2.16E+6	1889776	2484222	7.73E+0	4	13	5.78E+0	1	11	5.35E+1	24	92
Random Pairing	Low	7.06E+3	5 962	8 472	2.81E+5	204 342	411 934	3.45E+0	1	7	9.31E+0	1	16	4.19E+1	12	76
	High	7.62E+3	6 006	8 990	3.26E+5	213 684	444 944	4.27E+0	1	9	7.13E+0	1	16	5.22E+1	14	90
	Random	7.32E+3	5 966	8 930	3.01E+5	205 730	453 740	3.61E+0	1	8	8.73E+0	2	16	4.65E+1	10	80

Figure 6: Tabulation of bubbling and pairing procedure results, with upsets. Note that the expected values of Spearmans Footrule, Spearman's Rho, and Spearman's footrule on the break, are 3.76E+4, 6.32E+6 and 8.50E+1 respectively(J Marden *Analysing and Modelling Rank Data* (1995) page 77). A perfect tournament has no break upsets. The break loser in a perfect tournament would be 17.

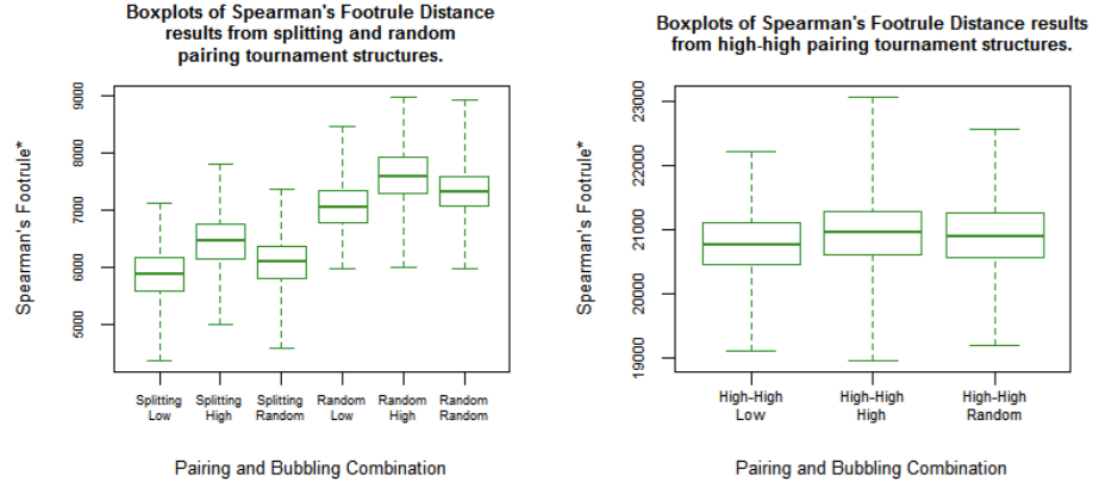


Figure 7: Box and whisker charts of Spearman's Footrule result, for different bubbling and pairing procedures, with upsets. ⋆ Note that vertical axis does not start at zero.

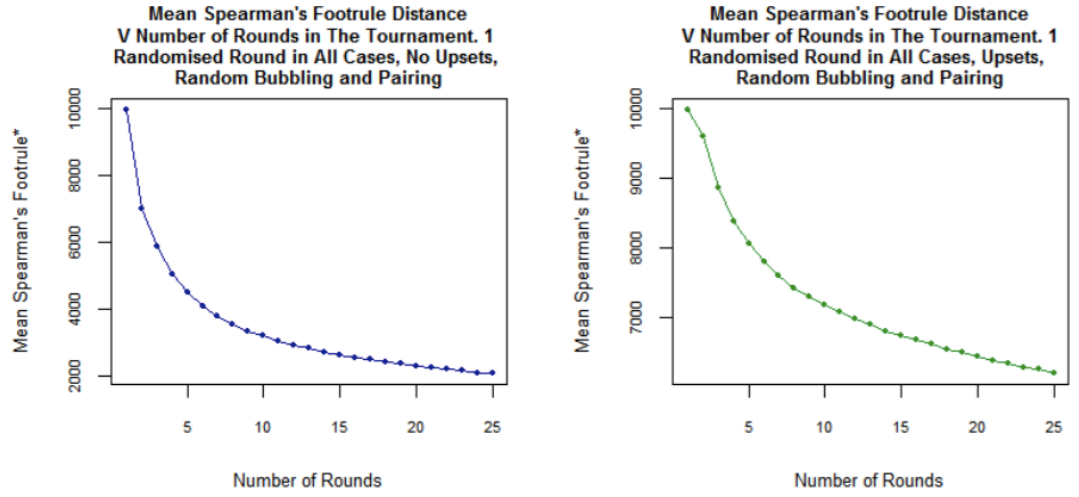


Figure 8: The effect of the number of rounds in a tournament on the mean Spearman's Footrule Result. ⋆ Note that vertical axis does not start at zero.

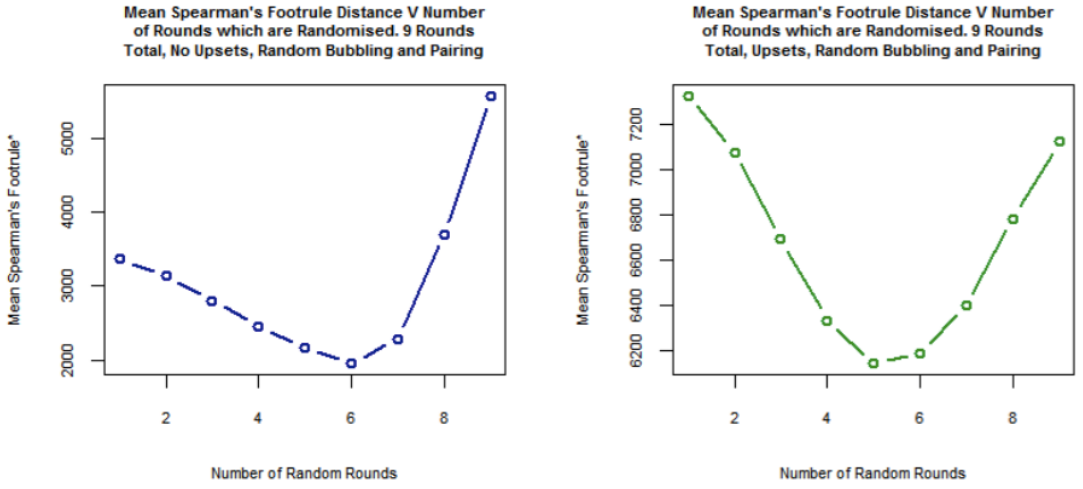


Figure 9: The effect of the number of completely randomised rounds in a tournament. ⋆ Note that vertical axis does not start at zero.

Interpreting the Results

A Note on Speaker Scores

Any non-randomised tournament structure will necessarily need to compare teams who are on the same number of points. The most obvious way to do this is by using team average speaker scores. It might therefore be worth discussing some issues raised in English and Kilcup's Article, *Abolish Speaker Tabs*.[11] First, it must be noted that the team's average speaker score does not suffer from all of the problems of individual speaker scores, described in English and Kilcup's article. Second, if speaker scores are used in determining match-ups or bubbling, I would recommend not using the total speaks up to that round

(which is what our model used). Rather, use the speaks from only the previous round. This will both make each round more competitive, and will ensure that outlier speaker scores only affect a team once. Third, it must be noted that there are many alternatives to speaker scores, which can also separate tied teams. Chess systems have had to develop measures of relative strength based only on wins, because you don't get a score in chess, you only win or lose. For example: the Buchholz System takes the sum of the points of each of the opponents faced by a team. "Direct Encounters", splits players (teams) based on who performed best when they faced each other. "Number of games played as black", is self-descriptive, and is used because black is considered more difficult. This could find an analogue as "number of debates as Opening Government", or whichever position has been weakest. There are several other systems as well,[1] and new ones could be created for debating (such as WUDC Constitution Art 4(a)(iii)).

Fairness and Competitiveness

Fairness is what we have been directly measuring with the model. It concerns whether the better teams in a tournament do actually do better, and if not, how evenly teams are prejudiced. A related, but not equivalent concept is that of competitiveness. Competitiveness refers to the extent to which a tournament incentivises teams to perform at their best. In any tournament where future round match-ups can be both predicted by the teams, and affected by them, there may arise an incentive to perform poorly. In theory, a tournament that is fair will not be uncompetitive. In practice, unfair tournaments can be competitive, and vice versa. For example, elimination tournaments are very unfair, have predictable future rounds, and yet are very competitive. This is because teams can't affect who they face in future. They either face whoever gets assigned to them, or they drop out. By contrast, round robins are the most fair tournament structures,and yet they often become highly uncompetitive. This is because teams who do badly early on start taking the tournament less seriously. It is apposite to mention here the analogous effect of dropping blind rounds. When teams reach a point where they either have enough, or too few points to break, it will affect their performance, if they are aware of the fact.

How Do the Tournament Structures Support Competitiveness?

Broadly speaking, randomisation supports competitiveness through unpredictability. Splitting brackets, and bubbling low, support competitiveness by creating incentives to score high. Pairing high, and bubbling high, do not

support competitiveness at all, because they create incentives to score low. Note that these considerations only apply to speaker scores, not points.

How Do the Tournament Structures Support Fairness?

We submit that the only way for a tournament to be more fair, is to order the teams better, and minimize outlier teams. It could be argued that a tournament which does a worse job of ordering teams, is in fact more ”fair”, if it prejudices teams on a random basis. This is fallacious reasoning. If a team ends up being severely disadvantaged, due to a combination of randomly being bubbled up more often than other teams and/or randomly drawing the strongest teams in the bracket more often than other teams, it will not be any consolation that this was a rare event, nor will it help that all the other teams in the tournament had stood an equal chance of being so disadvantaged.[13] That random can be unfair, is perhaps even more evident when considering completely randomised rounds, in figure 9. More completely randomised rounds aren’t even less fair by the Spearman’s Footrule metric, they are more fair, and yet one would undoubtedly still be very cautious of having too many completely randomised rounds. The reason is simple, some teams might be disadvantaged too much in having to face very strong teams. These same sorts of black swan events can happen within brackets. Evidence of this can be found by looking at the maximum values of the metrics in figures 4 and 6. In particular, the Spearman’s Rho metric, which exaggerates outlying teams within a tournament. Notice how the Spearman’s Rho maximum for ”Random Random” is more than double that for ”Splitting Low”, in the no upsets table. If one can appreciate that completely randomised rounds may be unfair, then it should not be too much of a stretch to imagine that randomised pairing and bubbling, which do worse than other tournament structures, by the Spearman’s Footrule metric, may be unfair as well.

Understanding Bubbling and Intra Bracket Pair-Ups

Pairing teams high-high, and bubbling high teams, was historically a popular method. The argument for this system is probably based on the mistaken assumption that, because power pairing pairs off teams on the same or similar points, it should pair off teams on the same or similar speaks. Two differences between speaks and points make this reasoning erroneous. Firstly, the tab doesn’t weight speaks equally to points. Ranking is by points first, and then by speaks. The primary objective of any debating tournament is there-fore to ensure that teams get the correct points, not the correct speaks. By bubbling

and pairing high,it is the teams in the bracket that should have got the most points from a round, who are now the most disadvantaged.[14] The second difference, is that points are zero sum. This means that, *ceteris paribus*, the other teams in your room will determine how many points you get. Speaks, however, are largely (though admittedly not entirely) independent of the other teams in the room. Therefore, if the tournament is also interested in getting a good speaks ranking, it doesn't matter much where teams are.

There is also an interaction between bubbling, and the stability problem described in section 2.2 above. Recall that some of the teams in a bracket will end up outranking teams in the bracket above them, when each new round is run. Bubbling high means that it won't even be the best teams in the lower bracket who "jump" in this way. It will be the teams below the best teams; the best teams having just bubbled up.

Low bubbling is perhaps counter intuitive. The logic behind it is that the teams in the bracket who were originally expected to lose, should be the ones who are disadvantaged. Bubbling is always going to be a problem for someone. With low bubbling, in the absence of upsets, things will go almost the same way they would have, if no bubbling had taken place (but see section 7.5). Perhaps most importantly, low bubbling increases the stability of the tournament. When the top teams in a bracket "jump" over the teams in a higher bracket, some of those teams that they jump over will now be teams that had bubbled. I.e. teams that they would normally beat anyway.

Splitting brackets is analogous to low bubbling. It affords the teams who are better on average the greatest chance of winning.

Last Note on Variance and Upsets

One could always simply rank teams by their total speaks. If speaker scores were completely reliable, then this tournament would, by one rating method, be perfectly fair. However, even assuming that speaker scores are completely reliable, there are reasons why such a tournament would not be preferable. Tournaments should give teams a chance to recover from bad rounds, to do well when it counts, etc. In this way, a team may rightly be considered "better" than a team that outscores them. Similar reasoning reveals a problem with low bubbling. At a given stage in a tournament, the best estimate is always that the bottom teams in a bracket "should" generally lose the next round (if there are no upsets). Yet no-one would suggest giving these teams an automatic fourth.

The purpose of having a full tournament in the first place, is to give them a chance to do better. By extension, it cannot be correct to make it unreasonably difficult for them to do well. Some amount of outright upsets in a tournament may even be considered healthy. The same problem also presents itself when splitting brackets. However, it does so to a much smaller extent. We maintain that a tournament should, at a minimum, seek to be fair in the absence of any upsets. However, the importance of variance in speaker performance must not be overlooked.

Conclusion

It is apparent that the preliminary rounds of debating tournaments cannot be considered to be particularly fair. The two largest reasons for this problem, are probably the lack of monotonicity, and the stability problem, caused by art Art30(3)(h)(ii) and (iii) of the WUDC Constitution. As a result, we suggest the following:

• The prelim rounds should only be seen to ensure that at least a large portion of the top 16 teams in a tournament will continue to the break rounds. The tab should not be considered to have any further value.

• As a direct consequence of the above, all Swiss tournaments must have break rounds. Tournaments structures such as that used at the South African WUPID qualifier, 2014, which consist of only power paired rounds, must not be used again. If it is desired that every team is able to speak in every round, then a round robin format must be followed.

• Tournament organisers, and the WUDC, should seriously investigate the possibility of placing at least some upper bound on the number of times that teams may see each other

• It has been shown that splitting brackets for match-ups, and bubbling low teams, significantly reduces the unfairness of tournaments. However, the bubbling structure is much less influential, and low bubbling presents its own concerns, which might outweigh its benefits. Splitting brackets, however,should be seriously considered as an alternative to randomisation.

•No tournament should ever match up teams high-high, or bubble up high

teams.

1. http://www.fide.com/component/handbook ↩
2. The other reason is obvious, tournaments are also just more fun when teams get to see a larger number of different teams. ↩
3. Collier, P The Bottom Billion (2007) page 18 ↩
4. For example, many of the measures of the disorder of a list, are tailored to measuring the number of CPU cycles required for a computer to sort the list. ↩
5. Diaconis, P & Graham, R. L *Spearman's Footrule as a Measure of Disarray* Journal of the Royal Statistical Society. Series B (1977) Vol.39(2) page 262-268 ↩
6. Diaconis, P *Spearman's Footrule as a Measure of Disarray* ↩
7. Why not just: "…consists of a number of teams that is not divisible by four…"? ↩
8. There were 340 team at Worlds 2014. We drop four, because not all teams completed in all the rounds. ↩
9. We recognise that speaker scores are somewhat unreliable. However, they are certainly the best estimate available for the absolute strength of teams, for the purposes of modelling upsets. See also section 7.1, for more on this. ↩
10. The word pairing doesn't accurately capture the fact that there are four teams in a debate. But we use it for convenience. ↩
11. M English & J Kilcup Abolish Speaker Tabs Monash Debate Review (2013) ↩
12. http://www.fide.com/component/handbook ↩
13. The FIDE chess rules, in addition to monotonicity, require that a team may only bubble once per tournament. ↩
14. It also has the effect of inflating the scores of the teams who don't bubble. However, low bubbling also has the effect of deflating the remaining team's scores. Thus, between low and high bubbling, this a moot point. ↩

December 23, 2014 Neil du Toit Volume 12. 2014 Commentary and Critique: the Functioning of Tournaments

Proudly powered by WordPress

Monash Debating Review

An annual publication of the Monash Association of Debaters

It Actually Has a Real-Life Function: Debating as a Pedagogical Tool in Singaporean Education Introduction

Competitive debating is often lauded as a means of promoting critical thinking in students. In the published literature, debate is referred to as having the capacity to act as “an intense learning laboratory” that “is to language arts what calculus is to mathematics” (Hooley 18). It is therefore ironic that debate enjoys limited emphasis in Singaporean education. Besides a minority of schools which prize debating as a niche co-curricular activity, most schools do not see their debating societies as a major part of their branding efforts. Debating is even less visible in the classroom, as national examinations are largely written in nature, and oratorical abilities take a backseat. This article examines the viability of debating as a pedagogical tool for high-schoolers, in the context of the teaching of General Paper (GP) in Singapore. GP, a compulsory ‘A’ Level subject, requires students to craft argumentative essays on real-world topics, and demonstrate comprehension of given passages. I posit that the skilful adaptation of conventional debating formats and strategies can improve the teaching of not just GP, but classroom teaching in general.

Characterising the Singaporean Student: Bridging the Gap between Expectations and Reality

Singapore’s education system aims to help students develop three 21st Century Competencies: “Civic Literacy, Global Awareness and Cross-Cultural Skills; Critical and Inventive Thinking; (and) Communication, Collaboration and Information Skills” (21st Century Competencies). However, the reticence of most Singaporean students means that this expected readiness to engage in intellectual expression does not often materialise in reality. This can be credited to the “monologic, transmission-oriented mode of teaching that has been found to characterise teaching in Singapore” (Teo) which can be resolved by enabling “space for dialogue…to be expanded in classrooms” (Teo), with one form of such dialogue being “dialogue as a debate” (Teo). Indeed, debating, and

its ability to promote expressive and cognitive skills, might contribute towards the attainment of these competencies.

Debating and the Teaching of General Paper

Given that GP is meant to "develop…the ability to think critically, to construct cogent arguments and to communicate their ideas" (General Paper), there are definite parallels between the subject and debating. As a GP teacher at Meridian Junior College, I tested out the applicability of debating to the classroom. Most encouragingly, students were generally enthusiastic towards the concept of debating. However, there was little actual knowledge of what debating entails. There was a need to explicitly provide instruction on the smallest details, to prevent the debate from deteriorating into aimless banter.

I attempted a modified British Parliamentary debate in two classes of 21 and 26 students respectively. To render it more manageable, speeches were reduced to 5 minutes, and students could share the speeches, with a maximum of 2 students co-performing one speaker role. Points of information were retained, to promote responsiveness. I considered having students offer quick-fire rebuttals to opponents post-debate instead, but this might become disorganised and rowdy.
In retrospect, a comprehensive debrief would have been valuable, but I could not conduct one due to time constraints. However, I found that the viability of the activity lay in how I could constantly evoke relevant aspects of the debate to better their academic learning. For instance, substantive points paralleled the structure of their essays. Also, the onus on Closing Teams to distinguish themselves from Opening Teams helped students understand how to negotiate the Application Question, a particularly difficult exam component where they were expected to write a short essay in engagement with a passage. Blindly rehashing the author's original arguments is frowned upon as it does not evidence any value-addition in argumentation. Cross-referencing to the idea of a debate extension allowed students to understand what is meant by value-addition, something many struggle with. The interactive nature of the debate enabled instant recall when I referenced aspects of the debate, even months after.

In addition, I encouraged debate on controversial issues, such as the execution of Van Tuong Nguyen, and to justify their stance. In each discussion, every student was given a green and a red card, with the former indicating

agreement and the latter, disagreement with the motion. To allow for open-mindedness, students were allowed to switch stances if they wished, but had to explain why.

Lastly, I sought to link debating strategy to argumentative skills for essay-writing. Students adopted assigned profiles (different ages, genders et cetera) and considered how this would affect their stances on certain government policies. This demonstrated how characterisation of an issue or stakeholder could affect persuasiveness. On another occasion, students debated a hypothetical motion to vary jail terms based on prisoners' incomes. After an informal debate, groups penned justifications for their positions on the whiteboard. A specific writing format was mandated, with questions like "Why is this the case? What is the outcome?" Finally, a selected speaker would present the answers in a cogent speech. While mirroring the delivery of a substantive point, the writing format assigned also resembled the structured teaching of paragraph-writing that had gone on throughout the term. The debate influence behind the activity made this structure extremely intuitive, because students were thinking in terms of what bases they had to cover to be persuasive, rather than seeing it as a formula to be memorised.

Insights for Classroom Education in General

My takeaways are not unique to Singaporean education. Indeed, given the widely-acknowledged value of critical thinking, classroom debating can be useful across many contexts, and to any discipline that prizes argumentation and diverse views, especially the humanities and social sciences. However, certain considerations are needed for effective implementation.
How do considerations that accompany classroom debating differ, one might ask, from that of competitive debating? For one thing, the participants' objectives differ. For debaters, debate mastery is an end in itself. However, classroom debating is a means to the end of larger educational and assessment outcomes. Hence, teachers must clearly explain the linkage between the debates and subject learning, rather than expect organic skills transference. Secondly, teachers work with larger class sizes. My classes ranged in size from 16 to 26 students, as opposed to the Worlds Schools team size of 5 members. Also, unlike large debate clubs, non-speaking students could not be left to watch and track debates independently, because non-speaking students are unused to the length and rigour of full debates and are more likely to tune out. Hence, teachers must consistently play an active facilitative role to prevent

disengagement. For example, teachers can give watching students a part to play when the debate is ongoing – the role of Scribe or Questioner can keep students busily engaged during debates. Alternatively, an interactive debrief, where students know they will be questioned on the debate, would incentivise them to focus.

Thirdly, crucial guidelines for classroom debates might be deemed unnecessary, or even excessive in competitive debating. Being strict about debate etiquette is vital. In some debate circuits, such as the Asian varsity scene, the ability to mock an opponent's case without resorting to personal attacks, or to make witticisms at each other's expense, can be easily dismissed as stylistic flair or playful banter. In class, however, where public speaking may be stressful for shyer students, explicit boundaries are needed to establish a safe space.

Some other considerations should be taken into account. Tumposky raises reservations about the reductive nature of classroom debating, saying that "by setting up issues as dichotomies, debate…ignores the multiplicity of perspectives inherent in many issues" (Tumposky 53). Furthermore, she suggests "a confrontational classroom environment" (54) would alienate certain participants. She draws on research that shows that "very few women are comfortable with adversarial argument" (54) and that "cultures that value social harmony rather than individualism also are likely to prefer pedagogies that seek harmony" (54), citing "African-American, Latino, Native American and Asian students" as examples. (54).

In response to Tumposky's critique, I posit that given the importance of learning to process and justify ideas in this complex information age, students, regardless of background, should be trained to consider the logical validity of ideas, rather than avoiding open argumentation simply because argumentation is inherently combative. Also, inclusiveness can be generated if teachers consciously keep the activity from domination by the same few voices. Lastly, students who greatly prefer collaborative learning can undertake research or preparatory roles, and participate without the stress of delivering speeches.

Furthermore, regarding Tumposky's concern about binary argumentation, debating can actually scaffold, rather than detract from, the attainment of multifaceted thinking. Learning to grapple with two divergent opinions is the first step towards negotiating a greater variety of views. Additionally, pluralism taken to extremes has its downside – individuals may avoid holding definite

opinions “in the interest of embracing “difference””, taking the easy way out and “(seeking) refuge from the pluralist storm in that crawlspace provided by the expression “I don’t know” (Fine).

In summary, teachers who attempt to introduce debating into their pedagogy should take four key criteria into account:

- Do the students feel safe sharing their thoughts?
- Is this activity inclusive?
- Can the teacher act as an effective facilitator and moderator, enabling free student-directed exchanges, but also intervening if things get hostile?
- Is the debate relevant to the demands of the academic subject, and is this linkage visible to students?
- These criteria would allow the cognitive merits of debating for youths to be incorporated into education, while mitigating obstacles that classroom implementation may result in.

December 22, 2014 Huiyi Lu Volume 12. 2014 Commentary and Critique: the Functioning of Tournaments

Proudly powered by WordPress