A System of Weighted Partial Ballots

For use at the 18th Annual Yale Mock Trial Invitational Tournament

Ben Graham* Annie J. Wang*

Version: December 6, 2013 at 07:44

Abstract

Weighted Partial Ballots (WPB) is a tabulation system that ranks teams based on point differential and opponent strength. Instead of counting every ballot as a win or loss, teams earn partial ballots based on point differential. The partial ballots earned are then weighted by the amount of partial ballots earned by the opposing team. WPB corrects the bias of traditional tabbing for narrow victories over weak opponents and instead rewards teams for significant wins over strong opponents. This short guide explains the theory behind WPB as well as its practical implementation at the 18th Annual Yale Mock Trial Invitational.

1 What do ballots represent?

Mock trial is an inherently subjective activity. A team that one judge considers polished and poised, another judge will find robotic and scripted. For any given round, observers will disagree about the even simplest measure of performance: which team won.

We can think of a team's quality as a distribution of *outcomes* along a spectrum where *outcome* is defined as the set of performances that a team gives. Every team has an average outcome, and scattered around that average, it also has alternate outcomes based on variations in performance. The best teams will have high average quality and low variance while weaker teams might have a lower average quality and higher variance. The ballot is a tool for measuring the distance between two teams. In theory, when two teams face one another, we expect that the point differential reflects the distance between the two teams.[1]

In practice, point differential rarely offers such a clear picture, and the primary reason is variance in judging. No two judges react to a round in quite the same way: the time of day, the room temperature, even the contents of lunch—to say nothing of reactions to case strategy or interpretations of the Midlands Rules of Evidence—can affect how judges evaluate performances. Moreover, even if judges could agree on the relative merits of a performance, they may not agree on what score the performance deserves. Indeed, many judges claim to score "objectively," which in practice can mean adopting low standards while assigning high scores. This results in judges

*Ben is an adviser to the Yale Mock Trial Association and a former competitor at Washington University in St. Louis. Annie is a former competitor at Yale. The opinions expressed here are solely those of the authors and do not represent the official views of the Yale Mock Trial Association or Yale University. Please do not distribute without permission. Please contact Annie at annie.wang@aya.yale.edu with any questions or comments.

[1] The caveat here is that the maximum point differential is constrained to 140 whereas the maximum distance is unbounded.

routinely scoring performances as 8s or 9s, leaving them little room to differentiate between two teams. For these reasons (and more), the point differential that we see on a ballot is both a function of the two teams' "actual" quality as well as what we might think of as random noise: variance in judging.

The goal of any scoring system is try to use a quantity we can measure—in this case, scores on a ballot—to approximate a quality that we cannot measure: the caliber of the team. At the end of a round, the point differential is the best piece of information we have about what happened in that round. The role of any tabulation system is to decide how to best make use of this information.

Here is the central assumption that this proposed system, Weighted Partial Ballots (WPB), makes about the point differential: +50 is better than +1. That is, if Team A outscores Team B by +50, we have strong evidence that Team A really did perform better than Team B in that round. If Team A outscores Team B by +1, however, we're not so sure.[2] In both cases, it is possible that Team 2 really did perform better, but it is far less likely when the point differential is high. If we extend this logic, then we can also say that if Team A goes +15 against Team B, who has (up to that point) an average point differential of +13, then Team A is likely better than Team C, which has gone +15 against Team D, which earned an average point differential of +4. In short, winning against strong teams is better evidence of a team's strength than winning against weak teams. Although this emphasis on point differential may seem radical, this type of reasoning is actually reflected in traditional, AMTA-sanctioned tabbing, in which point differential is used as a tiebreaker. Whichever team has a higher point differential wins the breaker, i.e. is acknowledged as the "better" team. In the remainder of this guide, we elaborate on the mechanics of the WPB system and contrast it to the traditional tabulation system.

1.1 Traditional Tabbing

At AMTA-sanctioned tournaments (and almost all invitational tournaments), scoring is governed by the American Mock Trial Association Tabulation Manual.[3] This tabulation system, though familiar, is a complicated and artificial mechanism. It moves from a single, binary data point (better/worse) through a series of formal constraints to arrive at a more "objective" result. These constraints are not essential properties of mock trial; rather, they are choices made by tabulation directors that determine the rules of the game.

The most obvious of these constraints is the design of the ballot. Rather than casting a simple vote on which team is better, judges are required to record twenty-eight scores. Each score is further refined into a ten-point scale, and that scale is (loosely) tied to performance criteria. Beyond the structure of the ballot, AMTA also gives judges specific instructions about how to complete their ballots. For example, the AMTA rules instruct judges to score as they go, so they are not too swayed by any one part of the trial (unlike, perhaps, real juries). Then, in the tab room, the AMTA tab manuals dictate that each part of the ballot be weighted equally, and that the results of the multiple judges be aggregated.[4]

At the simplest level, wins and losses are determined by point differential. Whichever team has more points wins the full ballot. If teams earn an equal number of points, each team wins

[2] One might argue that certain types of judges are more likely than others to use the full spectrum of points and, consequently, assign high point differentials. The problem of judges not following instructions is a problem for any tabulation system, and, ultimately, this objection goes to whether or not judges will pay attention to instructions. Our contention is that if judges are told that the magnitude of their point differentials matter, they will adjust their behavior accordingly. For more on judges' instructions, see Section 5.1.

[3] Available at `http://www.collegemocktrial.org/Tabulation%20Manual.pdf`

[4] Alternate configurations might call on the tab room to ignore outliers (for example, the highest and lowest score on each side) or to weigh witness/attorney parts unequally.

0.5 ballots. Regardless of whether a team wins by +1 or +50 or whether its opponent was First-Time Competitor F or Reigning Champion A, the team still receives the full ballot. The benefit of this approach is that, to some extent, it minimizes the impact of judging subjectivity. Even if judges disagree how much better one team is than the other, they might still be able to determine consistently if one team is better than another.

But there's a trade-off here between signal and noise, and the chief problem with traditional tabbing is that it throws away information. This binary measure of wins and losses ignores both the magnitude of the win and the quality of opponent, which can diminish its effectiveness as a mechanism of selecting the best team at a tournament. Furthermore, this type of tabbing creates perverse incentives for a team to avoid the most competitive teams so that it can maintain a strong win-loss record. After all, a team that goes 7-1 against weak teams by margins of +1 each time beats a team that goes 6-2 against top teams by average point differentials of +50. In doing so, traditional tabbing rewards teams for avoiding top competitors.

1.2 Weighted Partial Ballots

Weighted Partial Ballots (WPB) is a scoring mechanism that tries to recapture some of the information that is lost in traditional tabbing. As the name implies, WPB has two components: partial ballots and weighted ballots. Partial ballots add information about the magnitude of the win by awarding teams fractional ballots based on a ballot's point differential.[5] Teams who win by larger margins will receive a greater proportion of the full ballot, and all teams whose point differential exceeds a certain threshold (here, +14) will win the full ballot.[6] Weighted ballots incorporate information about the quality of one's opponent. For both round pairings and final rankings, the partial ballots that a team earns in each round are multiplied by the partial ballots that the opposing team earns throughout the tournament. As a result, going +50 against a good team wins more ballots than going +50 against a bad team under the WPB system.

Unlike traditional tabbing, WPB rewards teams who face strong opponents and do well against them. Round-pairings already use power-matching so that strong teams face one another. WPB merely extends that idea, so that both opponent strength and quality of win are reflected in the number of ballots earned. In many tournaments, good teams will have bad first rounds, (inadvertently) allowing them to have easier draws throughout the remainder of the tournament.[7] By Round 4, they're 4-2, and with weak opponents, they place well without ever having to face other strong teams. Under WPB, winning two ballots against a weak opponent is worth less than winning two ballots against a strong one. By the same token, losing two ballots to a good opponent is not as debilitating as losing two ballots against a poor opponent.

2 A Brief History of Weighted Partial Ballots

This document is certainly not the first time that the idea of weighted partial ballots has been proposed in mock trial. However, to the best of our knowledge, no other tournament has ever fully adopted the WPB system. Since 2008, UC Irvine's Beach Party Invitational has reported "fractional wins" and "adjusted fractional wins" on their famous Sabremetrics tab summary, but tabbing is done traditionally. Here, "fractional wins" essentially refers to the use of partial ballots

[5] Note that the traditional system also assigns partial ballots in the case of a tie. Our approach could be seen as an extension of this policy.

[6] For more on this threshold, see Section 3.1.2.

[7] We do not mean to imply that any team would purposely perform badly in Round 1. However, even good teams have bad rounds, and this sort of phenomenon has appeared on many tabulation summaries.

while "adjusted fractional wins" weights partial ballots by strength of opponent.[8] Similarly, the Macalester College Invitational Tournament uses WPB as a tie-breaker but otherwise relies on traditional tabbing to calculate round-pairings and initial rankings. The 18$^{\text{th}}$ Yale Invitational will be the first tournament to rank teams entirely using WPB.

(We said this section would be brief.)

3 How to Calculate Weighted Partial Ballots

3.1 Calculating Partial Ballots

The first component of calculating WPB requires adjusting the amount of the ballot earned based on the point differential (PD). By relying on point differential, we assume that judges will use a comparable range of scores on their ballots, but as any competitor will tell you, judges routinely ignore the instructions they are given to use the full range of scores. As a result, point differentials on two ballots can exhibit great variance, not due to the relative performance of the teams but rather the variance of judges. If number of partial ballots won were solely dependent on point differential, then teams might unfairly benefit from that rare judge who regularly hands out +50 point differentials. Hence, in addition to calculating partial ballots based on the the value of point differential, we also add a "quality win" threshold, the point differential above which a team wins the full ballot.

Ultimately, calculating partial ballots requires making judgment calls along two dimensions: first, determining the marginal value of each point, and second, determining the threshold for a "quality win."

3.1.1 The Marginal Value of a Point

In considering how to determine the marginal value of a point, we considered a variety of approaches, including linear and logistic progressions. Under a linear progression, each additional point earned increases the fraction of the ballot won by the same amount. In a logistic progression, each additional point will be worth different amounts based on where the model is centered. Ultimately, the goal is to reduce judging subjectivity while rewarding teams for strong wins.

Our intuition is that judges are more likely to differentiate between a +0 versus a +1 round than a +13 round versus a +14. Hence, we adopt a logistic approach in which the marginal amount of a ballot earned by each team is determined by the function

$$f(x) = \begin{cases} 0.5\left(1 + \log_{q+1}(1+x)\right) & \text{if } x \geq 0 \\ 0.5\left(1 - \log_{q+1}(1-x)\right) & \text{if } x < 0 \end{cases}$$

where x is the point differential and q is the quality win threshold. For each additional point above 0, teams earn a greater fraction of the partial ballot.[9]

[8] Like the system outlined here, "fractional wins" employs diminishing marginal returns to each point. However, it does not appear to include a quality win threshold—see Section 3.1.2.

[9] Some might argue that the difference between +1 and -1 is minor, and we also considered centering a logistic progression at $q/2$. However, we decided against this for two reasons: first, while it's possible that some judges don't pay attention to which team is winning, our experience has been that judges are aware of which team wins or loses any given round. Second, it's much easier to explain to judges the importance of assessing point differential if the model is centered at 0 than at $q/2$, making it more likely that judges might actually adapt their behavior to score using the full spectrum and using larger point values in general.

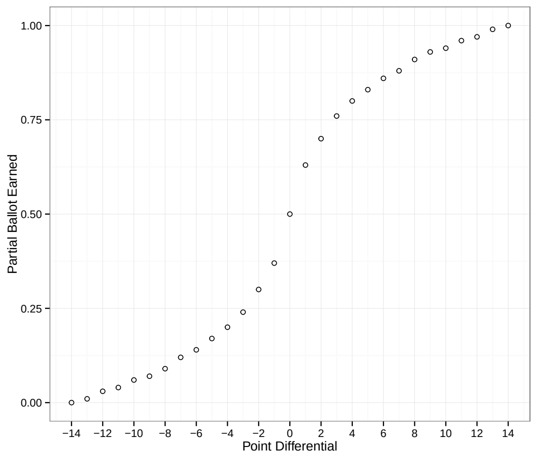

Figure 1: Logistic Model of Ballot Value

PD	Ballot	PD	Ballot
-14	0.0000	+14	1.0000
-13	0.0127	+13	0.9873
-12	0.0264	+12	0.9736
-11	0.0412	+11	0.9588
-10	0.0573	+10	0.9427
-9	0.0749	+9	0.9251
-8	0.0943	+8	0.9057
-7	0.1161	+7	0.8839
-6	0.1407	+ 6	0.8593
-5	0.1692	+5	0.8308
-4	0.2028	+4	0.7972
-3	0.2440	+3	0.7560
-2	0.2972	+2	0.7028
-1	0.3720	+1	0.6280
0	0.5000	0	0.5000

Table 1: Exact Ballot Values

3.1.2 What is a Quality Win?

Selecting the quality win threshold, admittedly, is a combination of guesswork and intuition. In traditional tabbing, the threshold for a "quality win" is simply +1, and generally, lower thresholds brings the WPB system closer to traditional tabbing. The challenge is to reward teams for particularly strong performances without exacerbating the issue of judging subjectivity.

For the Yale Invitational, we set the "quality win" threshold at +14: any PD above +14 will win the full ballot and any PD between -14 and +14 will win part of the ballot with +0 earning 0.5 ballots. Why fourteen? On an AMTA ballot, there are fourteen scores given for each team (six witness scores, six attorney scores, and two statements). We assume that if one team outscores the other team by an average of +1 on each of these fourteen scores, that constitutes a quality win.

How many ballots do we expect to be above the quality win threshold? We don't have a set number in mind, but given what we know from Nationals and ORCs, our quality win threshold is relatively high. During the 2012-2013 ORCs and Nationals, the average point differential was 7.359. Out of 144 ballots at Nationals, 27 (18.75%) were above the quality win threshold, and out of 384 ballots at ORCs, only 48 (12.5%) were above the quality win threshold.[10] As we mentioned above, lower thresholds bring WPB closer to traditional balloting, but unlike ORCs and Nationals, the Yale Invitational draws a wide range of teams, ranging from those who regularly finish top ten at Nationals to those who have yet to attend their first National Championship. This diversity is one of the strengths of the Yale Invitational, but it can also make for some lopsided matches, particularly in the early rounds of the tournament. By setting our quality win threshold at +14, we reward teams that significantly outperform their opponents without allowing a single opinionated judge to unduly influence the outcome of a round.

With the equation for partial ballots above and our quality win threshold of +14, we are able to calculate the proportion of a ballot won as a function of the point differential, as can be seen in Figure 1, and the exact value of a marginal point, as displayed in Table 1. Note that the amount of a ballot earned is reflected around a point differential of +0 (with a partial ballot value of 0.5): for any given ballot, the total number of partial ballots earned by the two teams is one full ballot.

[10] Data available at `https://db.tt/Y3MY68PB`.

3.2 Assigning Weights to Partial Ballots

After calculating the number of partial ballots earned by a team for any given round, we obtain weighted partial ballots by multiplying the amount of partial ballots the team has earned in a given round by the sum of partial ballots their opponent has won up to that point in the tournament. However, we make one adjustment before weighing scores: each team starts with 1.0 partial ballots × num. judges per round. For example, if there are 2 judges per round, 2.0 is the minimum WPB value a team can earn at the tournament.

Why do we make this adjustment? We want to compensate for teams who earn low PB values. We can use a very simple example. If Team A wins 2.0 partial ballots in a round against Team F, who wins 0.0 partial ballots throughout tournament, then Team A would earn 0.0 WPB in the round (without the adjustment). In fact, Team A maximizes the WPB it earns from the round by tying the ballot.[11]

In the spirit of competition, we want to encourage every team to win every ballot. With this adjustment, teams maximize the WPB earned from any round against any team by winning every ballot.

4 Ranking Teams

4.1 Round Pairing

Like traditional tabbing, WPB can and should be used to power-match round pairings. For the precise mechanics of pairing, we follow pages 16-17 of the AMTA tabulation manual with one significant change. Under traditional tabbing, team rank is determined in the following order: ballots won, combined strength (after round 2), point differential, and coin flip tiebreaker. Because WPB collapses measures of ballots won, combined strength, and point differential, team rank under WPB is determined by either the amount of partial ballots won (after Round 1) or the amount of weighted partial ballots won (after Rounds 2 to 3). We use partial ballots instead of weighted partial ballots after Round 1 because, with only one opponent, weighting ballots adds no additional information and may in fact detract from the accuracy of rankings.[12] Ties are broken after Round 1 by the point differential and after Round 2 and 3 by coin-flip (heads means the team nickname that is higher alphabetically gets the higher rank).

[11] We can also make this a slightly more complicated explanation. Imagine a tournament with one judge per round (for the sake of simplicity). In Round 3, Team A faced Team B. The WPB that Team A earns from Round 3 can be described as

$$\text{WPB}_A = x(1 - x + y)$$

where x is PB value that Team A earned in Round 3 and y is the PB value earned by Team B in rounds 1, 2, and 4. When we maximize this function with respect to x, we see that Team A earns the highest WPB when

$$x = \frac{y + 1}{2}$$

That is, if $y = 0$, then Team A maximizes its WPB when by earning 0.5 PB, *not* 1.0 PB. Team A only maximizes its WPB by earning 1.0 PB when $y >= 1.0$. Hence, we adjust each team's PB value upward by 1.0 PB for each judge/round.

[12] For example, a team that wins 2.0 partial ballots in a round with two judges would earn, with the 2.0 partial ballot adjustment (see Section 3.2), $(2.0 + 2.0) * (0.0 + 2.0) = 8.0$ weighted partial ballots. Similarly, a team that wins 0.0 partial ballots would also earn $(0.0 + 2.0) * (2.0 + 2.0) = 8.0$ weighted partial ballots.

4.2 Final Ranking

Final ranking is determined by the sum of weighted partial ballots earned. At the Yale Invitational, all ties will be preserved except in the case of a head-to-head matchup.[13] If two teams are tied in the amount of weighted partial ballots earned and have faced one another, the team that earned the higher number of partial ballots in their head-to-head round is ranked ahead of the other team. Note that compared to the traditional tabulation system, WPB creates fewer ties because it multiplies partial ballots by other partial ballots.

5 What to Expect at the Yale Invitational

5.1 Instructions to Judges

Like any tabbing system, the chief constraints on WPB are the quality and consistency of judging. At the Yale Invitational, we will explain to judges that we are not following the traditional scoring system and that point differential will play a large part in determining team rank. Both at the judges' presentation and on reminder sheets in the rounds, we will explain to judges that their role is to score teams relative to one another, rather than to some absolute standard of performance.[14]

Language on the AMTA ballot is unclear whether it should be interpreted objectively or relatively. On the ballot, judges are instructed that a "10" is considered an "excellent" performance while "5" is average. However, the directions are unclear as to whether that average is pegged to the universe of all possible performance or the performances in the room. The fact that AMTA lists several concrete standards for attorneys and witnesses (e.g. "made timely objections", "showed emotion appropriate to role") below the scoring guidelines suggests there are some objective standards. We plan to make clear at the Yale Invitational that a 5 should represent an average score *for that round/room* and that ballots should be treated primarily as a means of differentiating two teams from one another rather than ranking them against the universe of all possible teams.

5.2 The Tab Room at Yale

At first glance, WPB can seem like a pretty complicated system, but in many ways, it can be a much easier system to manage than traditional tabbing. At the Yale Invitational, teams should expect the same standards of accuracy and efficiency that the Invitational has exhibited in the past.

As described in the AMTA tabulation manual, each of the three ballots from a round will be checked initially for accuracy and completion. At least two persons will individually sum ballot total, calculate point differentials for each ballot, and then crosscheck their results with each other to ensure accuracy.

Unlike a traditional tab room, we will then enter these point differentials into an Excel spreadsheet. Two individuals, on separate computers and separate spreadhseets, will input these scores. Each team's WPB and rank will then be calculated on the spreadsheet, and the two

[13] Note that for final rankings, unlike rankings to pair rounds, it is not necessary for every team to have a unique rank. It may be the case that two teams finish the tournament tied for second place.

[14] In theory, an objective standard of judging is preferred to a relative one since it would allow direct comparisons of teams by point differential. In practice, even "objective" standards mean different things to different people, and it would be difficult to convince all judges of what caliber of performance a 5 should represent. Hence, for reasons of pragmatism, we advocate instructing judges to score ballots using relative measures.

individuals will cross-check their results with one another.[15] Finally, we will perform the round pairings and address impermissible pairings following standard AMTA procedure.

As in previous years of the Yale Invitational, and as laid out in the AMTA tabulation manual, the tab room will be open throughout the entirety of the tournament, except following Round 4. The tab room may also be temporarily closed at the discretion of the tabulation director (e.g. when calculating ballot totals). We invite any interested competitor, coach, or spectator to drop by the tab room with any comments or questions.

6 Overview and Examples

To summarize, WPB is calculated as follows:

After Round 1:

1. Every team's partial ballot count starts at 1.0 PB $\times$ num. judges per round. Let this value be called the *base*.
2. Calculate partial ballots earned by each team in Round 1 and add *base*.
3. Pair Round 2 using partial ballot values.

After Rounds 2, 3, and 4:

1. Calculate partial ballots earned by each team.
2. Calculate the running total of partial ballots earned by each team.
3. Calculate weighted partial ballots for each team, where

$$\text{WPB} = \textit{base} + \sum_{i=1}^{n} \left[(\text{PB of Round } i) \left(\textit{base} + \sum_{i=1}^{n} \text{PB of Round } i \text{ opponent} \right) \right]$$

4. Pair next round/determine final ranking using weighted partial ballot values.

Ex. 1 — In Round 1, Team A faced Team B. Team A went +1 on one ballot, -2 on another, and +15 on a third ballot. How many partial ballots have Team A and Team B earned after Round 1?

Answer (Ex. 1) — We'll calculate results for Team A first. Since there are three judges per round, all teams start with a base value of 3.0 PB. On the first ballot, +1 yields 0.628 ballots. On the second, -2 yields 0.2972 ballots. On the third, +15 is above the quality win threshold and so earns 1.0 ballots. Team A wins $3.0 + 0.628 + 0.2972 + 1 = 4.9252$ ballots.

Team B goes -1 on the first ballot, earning 0.372 ballots. On the second, it goes +2, earning 0.7028. It goes -15 on the third ballot, earning zero ballots. In total, Team B wins $3.0 + 0.372 + 0.0.7028 + 0.0 = 4.0748$ ballots. As an additional check, we can easily verify that the total ballots earned by Teams A and B is equal to the total number of ballots in the round: $4.9252 + 4.0748 = 9$. (Recall that each team starts with 3.0 PB. $9 - (2)(3) = 3$.

Ex. 2 — Your tournament has completed two rounds. There were two judges per round. Team C faced Team D in Round 1 and Team E in Round 2. In Round 1, Team C won 1.5 partial ballots

[15] Teams interested in receiving this spreadsheet should contact allison.durkin@yale.edu. Interested teams should expect to receive this spreadsheet in November.

(not including the base PB value). In Round 2, Team C won 0.3 partial ballots. Team D has won 0.87 partial ballots over the last two rounds (not including the base). Team E has won 3.18 partial ballots over the last two rounds (not including the base). What is the WPB value for Team C?

Answer (Ex. 2) — Since there are two judges per round, every team starts with a base of 2.0 PB. Team C has earned $2.0 + (1.5)(0.87 + 2.0) + (0.3)(3.18 + 2.0) = 8.636$ weighted partial ballots after Round 2.

Ex. 3 — Your tournament is completed. There were 3 judges in each round. Team F earned 0, 1.3, 0.5, and 2.4 partial ballots in Rounds 1 to 4, respectively. Team F's opponent in Round 1 earned, over the course of the tournament, 11.1 partial ballots. Its opponent in Round 2 earned 7.3. Its opponent in Round 3 earned 6.2, and its opponent in Round 4 earned a total of 4.3 partial ballots. These partial ballot values have *not* been adjusted with the base. What is the WPB value for Team F at the end of the tournament?

Answer (Ex. 3) — Since there were 3.0 judges per round, the base is 3.0 PB. Team F has earned $3.0+(0)(11.1+3.0)+(1.3)(7.3+3.0)+(0.5)(6.2+3.0)+(2.4)(4.3+3.0) = 3.0+0+13.39+4.6+17.52 = 38.51$ WPB.

7 Conclusion

WPB is a tabulation system that allows us to recapture some of the information lost in traditional AMTA tabbing. By incorporating data about the magnitude of a team's win and the strength of its opponents, WPB is a more accurate mechanism for determining the best team at a tournament.

That said, the adoption of WPB also signifies the end of the 8-0. Under WPB, it's entirely possible that a team going +1 on all eight (or, in the case of Yale, twelve) of its ballots has a lower WPB value than a team going +14 on seven of its ballots and -2 on one. This scenario seems to violate one of the basic principles of competition, namely, that the most a team can do is beat the teams that it is has been assigned to face. If winning all eight ballots isn't enough, then what is?

In adopting WPB, we're abandoning the assumption that ballots only give binary cues: you won, you lost. That makes everything messier (as if this activity weren't messy enough), but it also presents a more accurate picture of a team's strengths and weaknesses. Instead of saying that a +1 is enough, we challenge teams to draw more significant distinctions between themselves and their opponents. Additionally, we're arguing that a +1 is not always a +1: the value of that "win" is contingent upon the quality of the opponent. WPB is a system that rewards teams for large margins over strong teams, and if that means teams no longer hope to be paired against a weak opponent in search of an 8-0, we consider that a good thing.

8 Acknowledgments

We thank Muneeb Ahmed, Allison Durkin, Alex Feiock, Michael Nelson, Michael Polovich, Abby Walter, and Ella Wood for their comments. We are also grateful to the Yale Mock Trial Association and the teams attending the 18$^{\text{th}}$ Annual Yale Mock Trial Invitational for piloting this system.

Appendices

Revisiting the 2013 National Championship Tournament Using WPB

As an example of how WPB works, we've re-ranked the teams at the 2013 National Championships using the WPB method. Obviously, this sort of post-hoc analysis is problematic since, without pairing each round using WPB, we cannot fully achieve the goals of power-matching. Figure 3 displays the WPB value for each team using different quality win thresholds. Figure 4 is an extension of Figure 3. Instead of displaying WPB value, it displays the rank of each team. Figure 5 presents a simplified comparison of ranking under traditional tabbing and WPB tabbing with a threshold of +14. All of these calculations were done using the Excel spreadsheet displayed in Figure 2. Both the Excel spreadsheet and the raw data for these graphs can be downloaded at `https://db.tt/Y3MY68PB`.[16]

In Figure 3, we can see that teams are separated into certain "packs," and while teams move within the packs, there are rarely dramatic jumps between packs as the quality win threshold changes. In Figure 4, we notice that with a quality win threshold of +14, the ranks of most teams have stabilized within +/- one rank. In the Mueller division, the ranking of the top five teams remains startlingly consistent as the quality win threshold increases. (In fact, the rank of these top five teams are consistent with traditional tabbing.) By contrast, we see much greater movement in the Napolitano division, where the ranking within the top five shifts several times.

Again, we emphasize that these analyses are highly speculative in the absence of pairing rounds consistently using WPB. But what would mock trial be without wild speculation?

[16] An interactive visualization of this data is available at `http://bit.ly/17cNTGZ`.

Quality Win:	14	Round 1			Round 2			Round 3			Round 4		
Example Team		***Opponent***	*Opp's ∑(PB)*		***Opponent***	*Opp's ∑(PB)*		***Opponent***	*Opp's ∑(PB)*		***Opponent***	*Opp's ∑(PB)*	
Score		*Ballot 1*	*Ballot 2*	*Ballot 3*	*Ballot 6*	*Ballot 7*	*Ballot 8*	*Ballot 10*	*Ballot 11*	*Ballot 12*	*Ballot 14*	*Ballot 15*	*Ballot 16*
Ballots		*PB 1*	*PB 2*	*PB 3*	*PB 6*	*PB 7*	*PB 8*	*PB 10*	*PB 11*	*PB 12*	*PB 14*	*PB 15*	*PB 16*
UVA	**1022**	**1240**	8.71848811		**1115**	11.4123986		**1055**	11.6077034		**1398**	8.12635798	
Point Differential:		6	4	15	4	-1	-6	2	12	-3	12	27	1
Partial Ballots:		0.85928251	0.79715806	1	0.79715806	0.37202099	0.14071749	0.70284194	0.97357862	0.24404198	0.97357862	1	0.62797901
UCSD	**1055**	**1244**	7.41023642		**1356**	7.66917905		**1022**	11.4883573		**1372**	8.07777062	
Point Differential:		-1	13	9	1	8	24	-2	-12	3	12	5	8
Partial Ballots:		0.37202099	0.98726152	0.92513708	0.62797901	0.90568387	1	0.29715806	0.02642138	0.75595802	0.97357862	0.83082095	0.90568387
Duke	**1100**	**1524**	10.5733694		**1508**	8.09513856		**1455**	10.2170938		**1164**	9.03551852	
Point Differential:		-12	-5	7	7	6	3	-9	-6	-29	8	-14	-2
Partial Ballots:		0.02642138	0.16917905	0.88393704	0.88393704	0.85928251	0.75595802	0.07486292	0.14071749	0	0.90568387	0	0.29715806

Figure 2: Screenshot

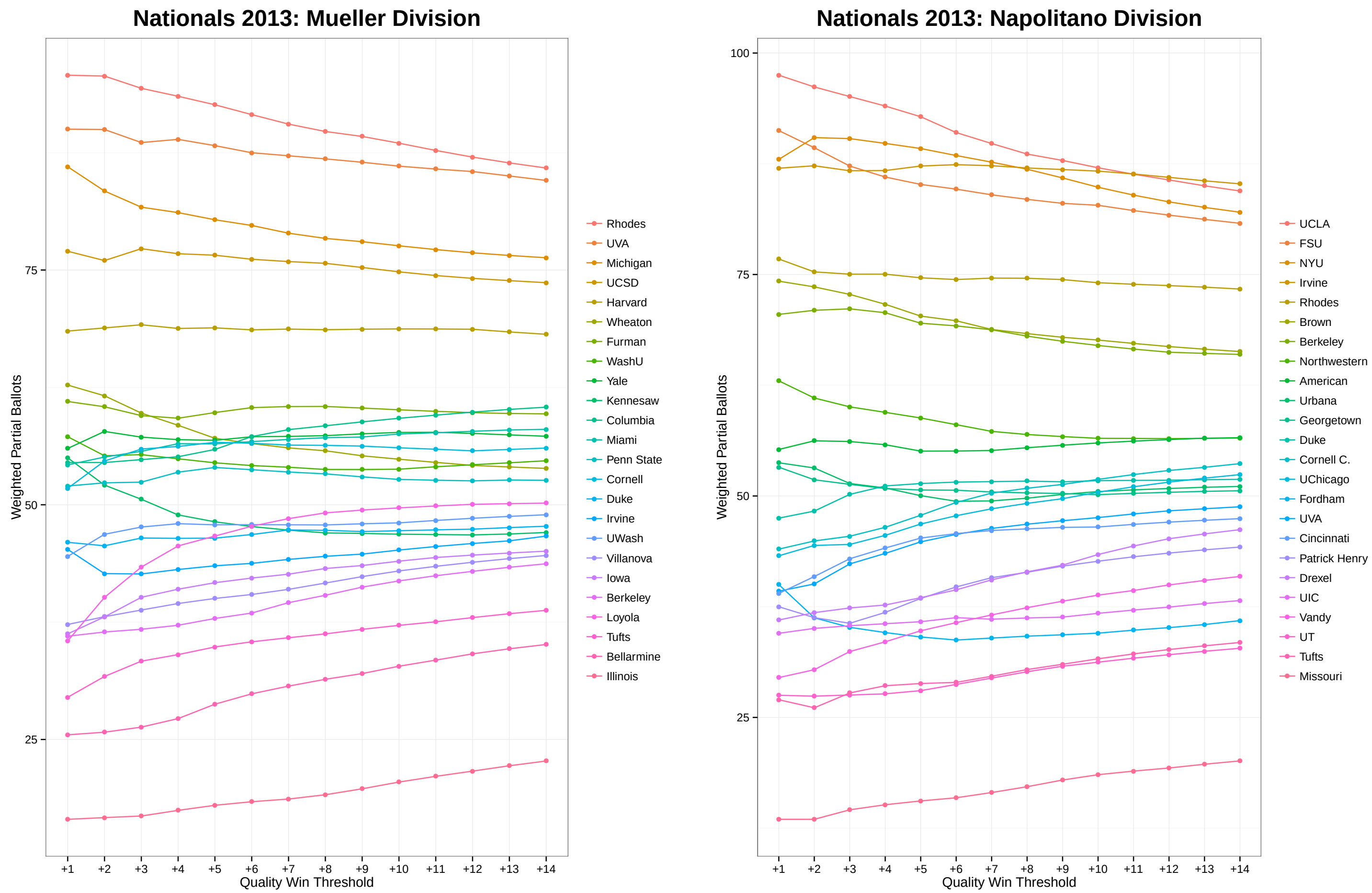

Figure 3: WPB with Thresholds from +1 to +14

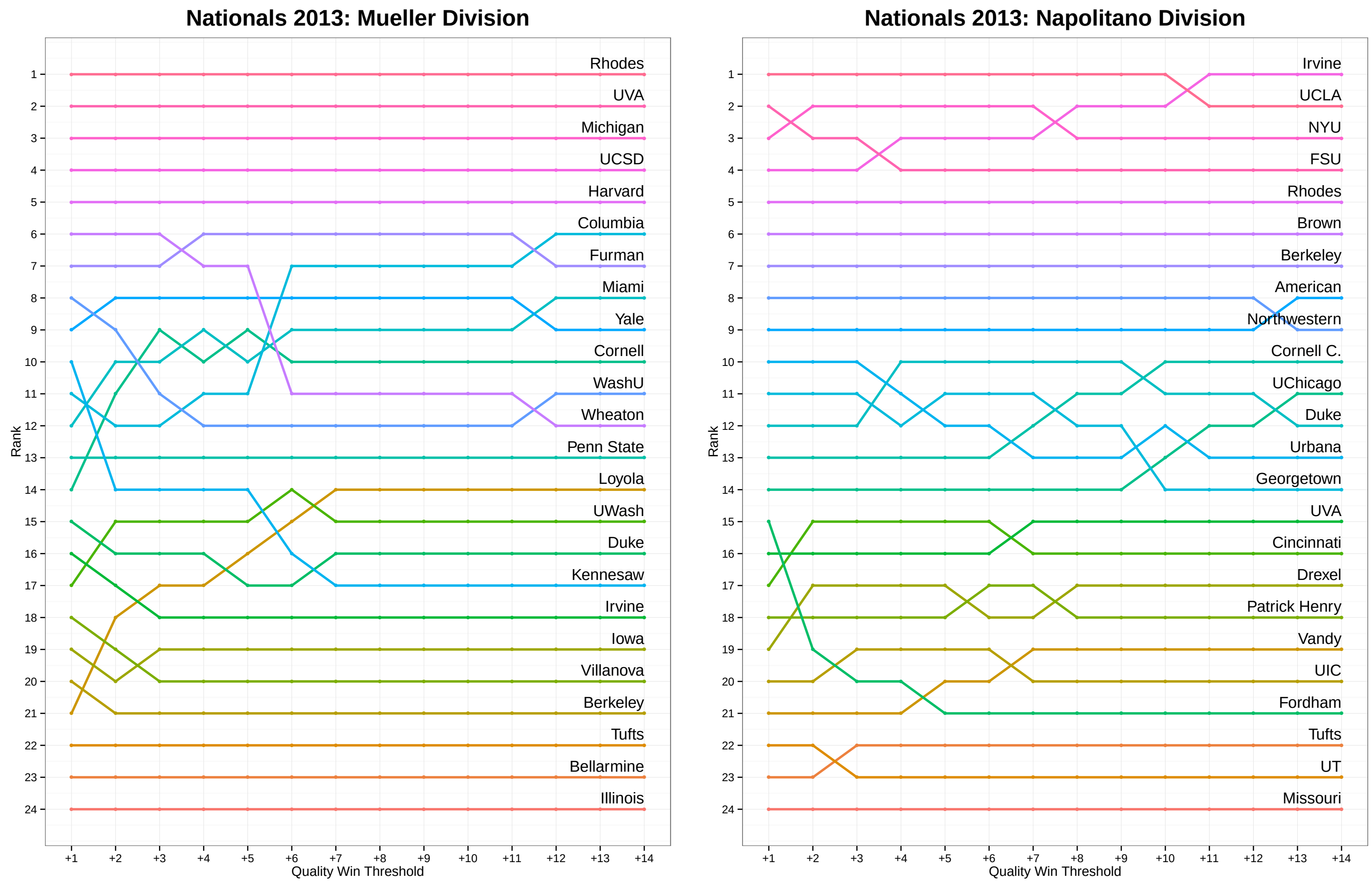

Figure 4: Team Ranks with Thresholds from +1 to +14

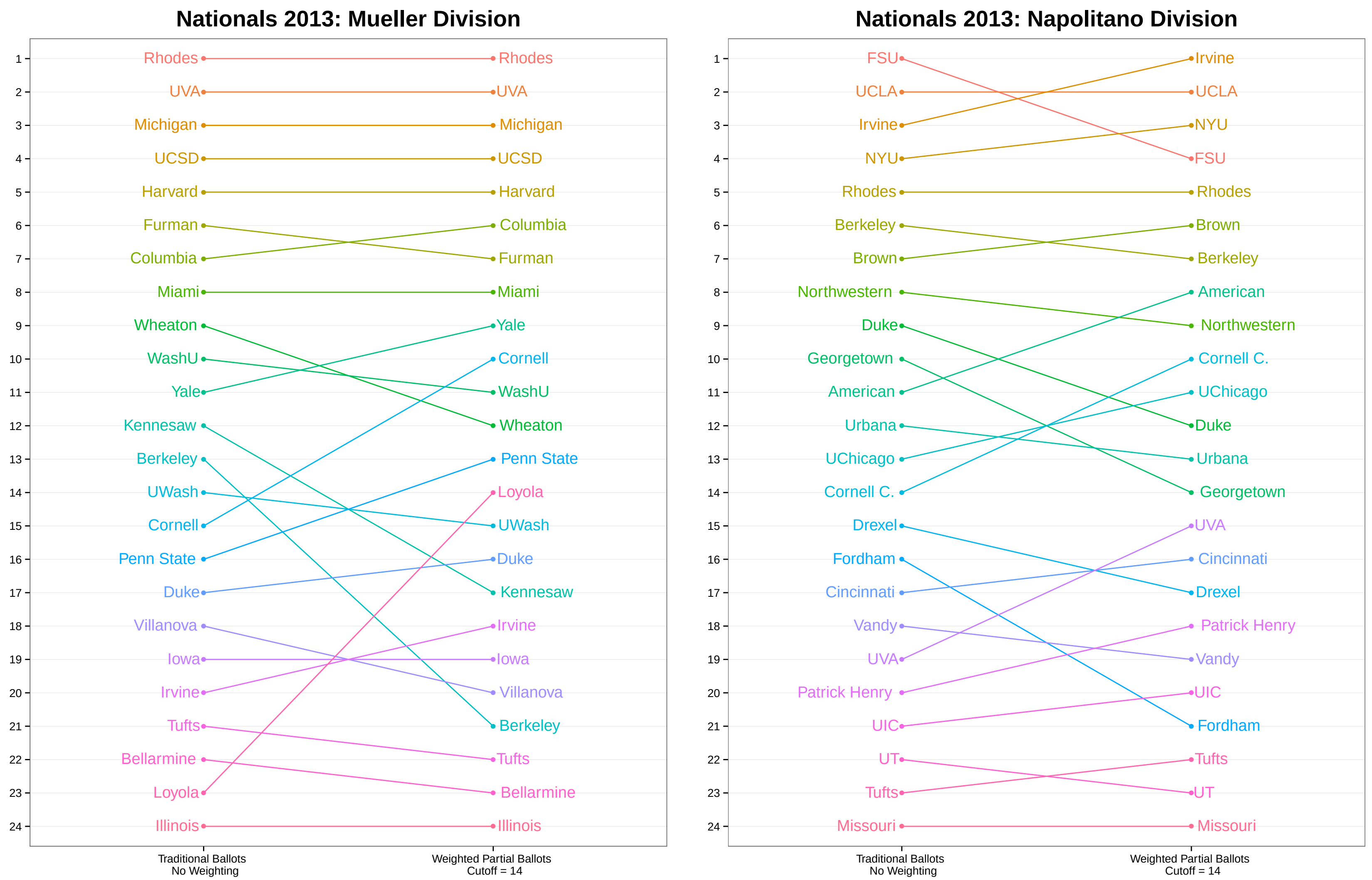

Figure 5: Team Ranking Under Traditional Tabbing and WPB with threshold +14