Pablo Rankings FAQ

Preface: This FAQ is written for the Pablo ranking used in the NCAA women's game. Middle Hitter has made a few modifications when referring to the source of the match scores.

Send comments and questions to **Pablo**. He'll answer questions about methods and take suggestions. However, he won't respond to "Your crazy, how can X be ranked where they are!"

Frequently Asked Questions:

I. HOW IT WORKS

1. Is this a computer ranking system?

No. This is *my* ranking system. It's my algorithm. The criteria and the approach were created by me. The computer is only used to carry out the calculations. In principle, it could be done with pencil and paper, but it would take a lot longer (an abacus could be used to speed things up).

2. On what basis do you rank the teams?

The underlying model for the ranking system is a probabilistic approach, which says, essentially, that the larger the difference in abilities between the two teams in a match, the more likely the better team is to win. For example, if we could clone a team and have it play itself, such that the teams were perfectly evenly matched, the chance that one side will win is 50%. On the other hand, in a total mismatch, the probability that better team will win approaches 100%. Somewhere in between, there will be teams whose differences result in a 75% chance of winning, or some other value.

What do these probabilities mean? I interpret them in light of any good probability: as the number of trials gets very large, then the fraction of wins will approach the probability. Well, that's all well and good, but it helps us little. Unfortunately, we don't have a very large number of trials, and we need to make assessments of the differences between two teams on the basis of only 1 or 2 (or sometimes 3) head-to-head matches. Thus, the challenge of evaluating teams.

3. How do I interpret the ratings?

In the end, the ratings are set on an absolute scale, with the #1 team set to 10000 points, but the absolute values don't matter. What matters is the difference between two teams. The difference between the two teams indicates the probability that the given team will win. The relationship between difference and probability is actually an integrated normal distribution (about diff/1000), and is fairly easily handled in Excel (this is a change from 2002, when the function was a lot more complicated). A table that gives a flavor of the relationship is shown below:

Difference	Probability
0	0.5
100	0.540

250	0.599
500	0.691
750	0.773
1000	0.841
1250	0.894
1500	0.933
1750	0.960
2000	0.977
2500	0.994
3000	0.999
4000	0.99997

Thus, the way to read this is that if two teams are separated by 1000 points, then the favorite has about an 84.1% chance of winning the match, a team that is rated 2000 points higher than its opponent wins 97.7% of the time, etc. Overall, the higher rated team won 85.9% of the time in 2001. However, the performance of the model is not based on how well it accounts for matches that have already been played, but on how well it predicts upcoming matches. This will be addressed more below.

4. How do you determine the probabilities?

If team A beats team B in a single match, then that indicates that team A is a probably better than team B. The more lopsided the match, the larger the difference between the two teams on average. Thus, each outcome is scored according to how lopsided it is. In the past, I used to break the outcomes into 11 different scenarios:

Normal 3 gamer
Lopsided 3 gamer
3 Game Blowouts
Close 3 gamer
Very Close 3 gamer

Normal 4 gamer
4 Game Blowouts
Close 4 gamer

Normal 5 gamer
5 Game Blowouts
Close 5 gamer

I would allow each of these outcomes to be scored differently (although they would not be required to be). However, recently I have discovered through match simulation that there is a very fundamental relationship between the probability of scoring points in a match and the probability of winning the match. Amazingly, the relationship between point probability and winning percentage is exactly the same as the relationship between Pablo ratings and winning probability. What this means is that there should be

a linear relationship between point probability and Pablo ratings. Unfortunately, we don't really know the point probability in each match, so I estimate it by using the point percentage. Therefore, each match is scored based on the percentage of points the team scored. It is a simple conversion, with

Game score = 27700*(point% - 0.5)

Therefore, if a team scores 55% of the points in a match (e.g. wins by scores of 30 - 25, 30 - 25, 30 - 24), they will be rated 1350 points ahead of the losing team (teams that have a 55% probability of scoring win 91% of the time). A team that scores 52% of the time (e.g. 30 - 25, 25 - 30, 30 - 25, 30 - 25) is rated 630 points ahead of the losing team (winning about 75% of the time). The maximum score for any match is always 2500, which corresponds to a point percentage of about 0.59. Thus, all matches better than e.g. 30-21, 30-21, 30-21 are scored the same. 2500 points indicates a win probability of about 99.4%. The system doesn't recognize any probabilities higher than that.

An interesting conclusion of this is that it is more important to know how many points are scored than how many games are won, and a team that outscores their opponent but loses will be rated higher. However, this is the same conclusion that I had drawn previously when the match weights were determined empirically, so it is not just a consequence of the modeling. In fact, the average ratings that I find for the different types of games in the old breakdowns are essentially the same as the ratings that I obtained when they were determined empirically. Therefore, the point % relationship that I discovered through simulation does not just work in theory; it is confirmed by actual results.

To get the rankings, I just vary the team ratings so to minimize the deviation between the actual game results and the calculated ratings differences. There's more to it, of course, but that's the basis for it. In a perfect model, the rating for each game would match the rating differences between the two teams, with a deviation of 0. In reality, because of the natural variation of performance and the nature of probability itself, it is impossible to create a perfect model.

5. Is that all there is to it?

It used to be. Past versions of Pablo rankings were based solely on the points approach. However, this year, after extensive testing, I have modified things slightly. The rankings now are based on a model that combines this "Pure Points" approach with a "W/L" approach, in which only wins and losses are taken into account, and not the match scores. Although it is possible to run each model separately, in the current version they are carefully integrated together in a single package. Tests have shown that the accuracy of the combined model is similar to that of the pure points approach. The W/L model alone performs significantly worse. Performance issues are addressed in the next section.

II. PERFORMANCE QUESTIONS

1. Do you account for the home court?

Yes. Home court advantage is an adjustable parameter in my system. The home court advantage varies, but is usually worth 200 points. I have heard suggestions about improving the home court advantage model, taking into account that, for example, not all neutral sites are really neutral, and that travel distance can affect the home court advantage. Although these are probably correct observations, I don't think that they will have as much impact on the quality of the rankings as many other factors. I may try to consider them later, after a few other issues are addressed.

2. Is it correct to include the "lopsidedness" of the match when calculating probabilities?

I know it is all the rage in NCAA football to disregard ranking systems that include the point spread for a game, on the grounds that "winning is all that matters." However, the motivation behind such a movement is suspect. In fact, Jeff Sagarin has had major conflicts with the BCS (football ranking system) because of this very issue (however, I see recently that he has not dropped out as he had originally claimed he would). Despite the extent to which people complain, Sagarin points out that not including the point spread in his ranking system leads to a *less accurate prediction* of future games. Therefore, on the grounds that the goal of a ranking system is to predict which team is better, and which team will win, then it is better to include the point spread in the ratings analysis. Personally, it looks to me that the real reason people object to including point spread in ranking systems is because doing so gives results different from opinion polls.

I have done a lot of testing of potential models and fitting procedures, using 2002 and 2004 data. In the end, the pure points model that I described above gives the best overall performance in terms of being able to predict upcoming matches, but the combined model is very similar. However, models that rank based solely on W or L, or 3/4/5 games <u>consistently</u> (as in, so far, always) perform worse than the full model or combined models.

3. But my team wins by a lot of blowouts and uses a lot of the bench, making the matches closer than what they would be if it was just the starters. Doesn't that bias the results?

It might, but to what extent I don't know. If it was a consistent feature, that happened every game, it could make a difference, but it's not clear that it happens that much. Regardless, it probably is too rare of an occurrence (affecting only a couple teams a year, at most) that it does not justify changing the model to accommodate it. Some have argued that the Ballicora model, where the match is only evaluated on whether it goes 3, 4, or 5 games, is better because it does not penalize the team for subbing (provided that it doesn't cost the team a game). However, I have tried the model where I only consider whether it was a 3, 4, or 5 game match and found it to perform on the whole (much) worse than the full model (this is somewhat surprising because BCR does much better than my version based only on whether it is a 3, 4, or 5 game match; I address this aspect in section III-3). Basically, it comes down to the following: Do I want to use the best model overall? Or should I sacrifice 15 correct matches a week (and 150 a

year) so that I get a theoretically more precise ranking of the top five (with no guarantee that it will make any difference)? I choose to the model that performance the best for all teams over the course of the whole season. As with all the aspects of my program, I am trying to think of approaches to use to handle these types of issues. I actually have something in mind for this right now and might test it this fall.

4. There's no way that (insert name here) is the XXth best team!

OK, it's not a question, but it is a common complaint. The answer to this is highly dependent on the specifics. However, there are a couple of points that I need to emphasize:

a) Oftentimes, people are surprised to learn that, in fact, the team in question is not as bad as one believes. This is especially the case when the team is one without a reputation for being good. My first step when I hear this complaint is to look up the record of the team in question. Often I find it is better than expected. Interestingly, I do not have any easy access to that type of information in my spreadsheet (I do not keep track of w/l records), and would have to count it up by hand if I wanted it. I use Rich Kern's website *(Men's scores are be provided Middle Hitter)*.

b) As with any probabilistic model, it does a better job with more data. Therefore, I don't put too much stock on rankings that come out early in the season. As the season progresses, the rankings become better.

c) One of the problems early in the season comes in the disparity in schedules. Because I put a limit on the team differences, a lot of wins in lopsided matches are not very useful from an evaluation standpoint. Therefore, even a dozen games into the season, the ranking can be overly dependent on one or two games. My belief is that the system does best when teams play a lot of games against teams that are close in ability. The biggest errors show up when looking at teams with lots of lopsided matches (up or down). Florida A&M is particularly notorious in this regard. They are a very good team that plays in a very poor conference. Hence, a very large number of their matches end up being blowouts against very weak competition, which are basically ignored in the ranking (see section II-5). When this is the case, their ranking even at later stages of the season should be viewed with caution.

d) Teams are ranked solely on what they have done, not what they were predicted to do, or what fans or coaches expect them to do for the rest of the season. However, there is some predictive value in the rankings.

e) Teams are ranked only based on their scores logged at **www.richkern.com** *(Men's scores are be provided **www.middlehitter.com**)*. Therefore, missing scores will obviously affect the teams' rankings. This is not much of an issue at the DI level, but still is very important for DII and DIII.

5. To what extent to you account for strength of schedule?

I don't have a specific "strength of schedule" parameter in my rankings, but it is

included indirectly. A team that consistently plays and beats teams that are ranked within the top 50 will be ranked a lot higher than a team who only beats up on teams that in the bottom 100. Wins against good teams are duly rewarded, whereas wins against the worst teams do not help one much.

At some point there are diminishing returns. Matches can be only so lopsided, while differences between teams can be very, very large. Therefore, I do have a limit to the differences between the two teams that I consider, beyond which I don't pay attention to the actual value. Therefore, if the two teams are separated by 3000 points, but the limit is 2500, then I will use the 2500 value in my fitting. Currently the limit is 2500.

6. My team beat that team. How can they be ranked higher?

The system tries to find the optimal values for ratings to give the best fit to the data, and it uses all the available data (except, to an extent, those data points for extremely lopsided matches, as noted above). While head-to-head data is useful, it is not clear that a single head-to-head is more informative than 5 matches with common opponents, or lots of matches against a common field.

7. Is it fair compare teams by comparing games against common opponents? There's too much variation!

Transitivity is a tough game, and gets even more challenging when you get to three and four opponents removed. But it's not bad. I did an exercise a couple of years ago where I looked at the first half of the Big 10 schedule, and tried to predict the outcomes of head-to-head matches based solely the differences with common opponents. I found that I could predict about 85% of the head-to-head results using a common opponent approach, which is essentially the same as what I get from my rating system. As expected, there were upsets (BTW, don't try this exercise with the 2001 Big 10 football season; that was one whacky season with lots of upsets).

8. My team's best player missed a match for some reason. Doesn't that affect their ranking?

Sure does, which is one place where my system will run into problems. I assume a constant level over the entire season. If certain players are missing, that will affect the outcome and hence the rating. However, I prefer this approach, wherein I integrate all games over the entire season.

9. Do you count more recent games more heavily?

Yes. I have implemented a procedure where more recent games are weighted more heavily. I have found that the best approach is to weight games played 42 days ago as half of those played today. While a one month half-life sounds appealing, don't read too much into it. I'm not sure physically what it means.

10. I was a big fan of the Ballicora Computer Rankings (BCR). How does your system compare to that?

In fact, BCR and Pablo rankings are fundamentally very similar. They both use the same type of probabilistic algorithm as the basis for the rankings. This is not surprising,

though, because it was the Ballicora rankings that motivated me to develop Pablo in the first place. I really liked Ballicora's probabilistic approach, and thought it would be fun to do something like it on my own. I didn't ever intend to replace BCR as the "computer ranking system," and initially my goal was just to do something fun hoping that someday I might be almost as good as BCR. However, as I continued to tweak my system and start evaluating it, I discovered an interesting fact: in terms of doing what I wanted to do, Pablo was actually as good as BCR or even better!

11. What do you mean "better"? How good of a ranking is Pablo anyway?

In order to determine how good a ranking system is, we first have to clearly define how we are going to evaluate the systems. As stated in Section I-3, the goal of my system is to create a ranking that can be used to predict the outcomes of upcoming matches. If I wanted to, I could choose a different focus, such as a true reflection of who has won, as opposed to who will win. In fact, I can and have done that type of ranking, creating a ranking that has a very high success rate in accounting for wins and losses in matches that have been played. However, my preferred approach is to make a system that can predict results for teams that have not played. We already know who won matches that have been played. The more interesting question is what is going to happen next week.

To that end, I have compared the performance of Pablo in predicting the outcomes of matches with that for other common rankings. The procedure I use is the following: at the end of each week, I compare the outcomes of all the matches that I can with the predicted outcomes, and just add up the score. The table below shows the results obtained in 2003 for comparing Pablo, BCR, and just for fun, RKPI, which is Rich Kern's attempt to replicate the RPI, the system the NCAA uses.

	Rankings	Next Week Test		
Week	**Date**	**Pablo**	**BCR**	**RKPI**
2	9/07	79.6		
3	9/14	80.6		
4	9/21	84.0	77.0	
5	9/28	80.7	80.4	79.9
6	10/5	83.7	83.3	76.4
7	10/12	85.0	84.4	81.3
8	10/19	82.6	78.3	73.7
9	10/26	78.9	78.3	75.5
10	11/2	78.4	78.0	71.6
11	11/9	83.1	80.1	77.2
12	11/16	80.5	77.3	74.7
13	11/23	80.0	81.4	81.4
	Average	81.3	80.2	77.6

The BCR rankings in the table include a 60 point home court adjustment because that is what gave the best results. The values in the columns under Pablo, BCR, and

RKPI are the percentage of matches that each method predicted correctly in the week following the ranking.

The table shows that over the course of the season, Pablo typically performed better at predicting upcoming matches than did BCR or RKPI. The difference between Pablo and BCR is certainly not large, but is consistently observed (outside of one week). The average deviation is about 1.1%, which amounts to something like 30 – 40 matches over the course of a season that Pablo did better. Granted, in a data set of 3000 matches it isn't large, and historically, I have never made a big deal of promoting Pablo as better than BCR. In my opinion, the important question is whether Pablo works as well as BCR, which I consider to be the "Gold Standard," and this table indicates that indeed it does. RKPI trails well behind, missing more than 100 additional matches each year. Yet, this is the ranking (or at least a version of the ranking) that the NCAA chooses to use when making its tournament selections. Sigh…

I have also done a comparison with the AVCA rankings, where possible. Assuming that teams ranked higher in the AVCA are the ones predicted to win, I found Pablo finished about 2.5% better over the course of the 2003 season, correctly predicting 293/336 matches. Using the AVCA poll lead to 8 extra losses (285/336). BCR with a 60 point HCA had 291 of those matches correct. However, the comparison with AVCA is dicey because, for one, there isn't a Home Court advantage that can be applied, and second, it's not clear what the real purpose of the AVCA rankings are in the first place. As far as I can tell, the AVCA rankings are basically what you get when you take last week's ranking and make adjustments based on whether the team lost or not, and therefore probably has some aspect of being a "reward" for past performance as opposed to a real analysis of who would likely win if they played. If you recognize this and keep it in mind when considering the AVCA rankings, it is not a real problem. But if your team is going to play another team and you want to consider how they should do, you are better to look to something like Pablo than to the AVCA poll.

12. Why is BCR almost as good as Pablo?

After doing all this work, the question is no longer "Does Pablo work?" but is instead, why does BCR work almost as well as Pablo? I have noted above that when I try the same approach as Ballicora, using only the information of 3, 4, and 5 game matches, I get significantly worse results than are obtained with BCR. So what is it about BCR that made it work in the first place? After contemplating this question for a long time, I think I have it figured out. By breaking the analysis down to a game by game basis, BCR effectively weights close matches more heavily than non-close matches. Thus, a 5 game match is given 5/3 the impact of a 3 game match. Pablo doesn't do this. Apparently, weighting the closer matches more heavily is a good thing to do. I am trying to think of a way to do this in Pablo rankings to see if it helps. This may be a way to improve Pablo even more.

III. PERSONAL ISSUES

1. How long have you been doing this?

I started building my ranking program at the beginning of the 2001 season. I was familiar with the Ballicora system, and thought that I could handle doing something like that. I started on a simple level, but then developed it into a complete package by middle of the season. However, I continually modify the software to make it easier to work with and to learn more about how it is working.

2. What are your qualifications for doing this?

My main qualification is that I know how to use Excel, albeit not always elegantly (although that is improving all the time). However, much of my insight into doing this comes from an understanding of phase space theory, a probabilistic scientific model for describing physical systems. The model I use is based on phase space theory, and I have programmed the algorithm into Excel. Recently I have developed (with help from others) a visual basic script to automate the fitting process. Other than that, my only qualification is that I am a volleyball fan.

3. What are you working on now?

My goals for this fall are to try to see if I can work on that weighting issue (Section II-12) and to try a couple of things to address the possibility of subs in blowouts, but the big new project is to extend Pablo rankings to other divisions. Now that Rich Kern ***(Middle Hitter)*** has started logging scores for DII, DIII, and NAIA matches, we should be able to produce rankings for those divisions as well. The biggest problem that we currently face is in the fact that the database is still very sparse. Modifying the Excel spreadsheet to accommodate other divisions is not difficult, and takes about 5 minutes. However, it doesn't do any good if the scores aren't there. I have spent countless hours already this fall inputting DIII scores into the database, but some scores are just really hard to find. We could use all the help we can get on this project, so if you would be willing to contribute, please contact **Middle Hitter** at **mhpr@middlehitter.com** and tell him you would like to help add non-DI scores to the database. I anticipate putting out a couple of DIII rankings this fall (2004), but it all depends on the quality of the data.

4. How can I contact you to get more information?

I am a frequent contributor to the **VolleyTalk forum** during the volleyball season, where I typically post under the name Pablo, or with some pithy Family Guy reference (currently, though, it is p-dub). I also have an email account at Excite, at the address volleyball_Pablo@excite.com. Unfortunately, I don't check that email very often so if you send something there, please don't expect an immediate reply.