# A Better FIFA Ranking: Stat 91r Final

Alli Wiggins, Sam Bieler, Nathán Goldberg
*Harvard University*

June 3, 2018

# 1 Introduction

## 1.1 Summary

Because high-level decisions in international football are made in consideration of the FIFA Men's Ranking, we suggest that FIFA adopt a more accurate ranking method at the conclusion of the 2018 World Cup. After testing several ranking methods in terms of their predictive accuracy, we developed two models that perform better than any models that have been tested on international football while also balancing predictive power and simplicity. We hope FIFA will consider our models as they decide how to improve the FIFA Men's Ranking. Here is an executive summary of the key findings that we discuss in our paper:

- We looked at 2 main model families: Elo and Glicko. The most predictive models previously applied to international football, according to the literature, were Elo-based models. We found that Glicko based models, which are more complex, outperformed all of the Elo-based models we tested.
- Our Elo-based model, **B-Elo**, is simpler *and* more accurate than the other Elo-based models.
- Our Glicko-based model, **WIGGO**, performed the best out of all the models tested.

| | | MODEL PARAMETERS | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Opponent Rating | Home Advantage | Match Importance | Goal Difference | Rating Uncertainty | Neighbor Factor | Time Decay |
| Current Model | FIFA Ranking | ✓ | ☐ | ✓ | ☐ | ☐ | ✓ | ✓ |
| Elo-Based Models | FIFA WWR | ✓ | ✓ | ✓ | ✓ | ☐ | ☐ | ☐ |
| | eloratings.net | ✓ | ✓ | ✓ | ✓ | ☐ | ☐ | ☐ |
| | B-Elo | ✓ | ✓ | ✓ | ✓ | ☐ | ☐ | ☐ |
| Glicko-Based Models | Glicko | ✓ | ✓ | ☐ | ☐ | ✓ | ☐ | ☐ |
| | Stephenson | ✓ | ✓ | ☐ | ☐ | ✓ | ✓ | ☐ |
| | WIGGO | ✓ | ✓ | ✓ | ✓ | ✓ | ☐ | ☐ |

Figure 1.1: Summary of the factors in each model's calculations.

| | | PREDICTIVE PERFORMANCE RANK | | |
|---|---|---|---|---|
| | | Binomial Deviance | Mean Squared Error | Average |
| Current Model | FIFA Ranking | | | |
| Elo-Based Models | FIFA WWR | 6 | 6 | **6** |
| | eloratings.net | 5 | 4 | **4.5** |
| | B-Elo | 4 | 5 | **4.5** |
| Glicko-Based Models | Glicko | 3 | 3 | **3** |
| | Stephenson | 2 | 2 | **2** |
| | WIGGO | 1 | 1 | **1** |

Figure 1.2: Summary of the models' relative performance.

## 1.2 Introduction and Motivation

Although the concepts underlying the FIFA Ranking's make intuitive sense (e.g. beating a stronger team gives you more points; more recent results are weighed more heavily), its implementation relies on a seemingly arbitrary mathematical formula without . The current FIFA Ranking is thus not an efficient way of measuring relative strength between national teams, which represents an obvious problem for all the federations and confederations, including FIFA, that rely on it to make decisions. Furthermore, an inefficient ranking system invites the possibility that teams might exploit its weaknesses in an attempt to improve their own ranking.
The motivation for this project is to find an alternative to the FIFA Ranking that can replace the current system after the conclusion of the 2018 FIFA World Cup, considering that some confederations, such as UEFA and CONCACAF, have already moved toward Elo-based ranking models. We look at several different ranking methods, comparing them in terms of their effectiveness at predicting match results in international football, to determine which one would be the best alternative to the current system, taking into account both simplicity and predictive accuracy.

## 1.3 Background and Literature

There is limited research available on the subject of men's international football rankings. The most relevant research has been carried out by Lasek et al. (2012), who showed that ranking methods used in other contexts clearly outperform the FIFA Ranking when it comes to predicting match outcomes[1]. Their research measured the performance of the FIFA Ranking and Elo-based ranking methods on two separate prediction accuracy metrics. Out of all the models they tested, they identified a variant of the Elo model currently in use for the FIFA Women's World Ranking and another used by the website Eloratings.net as the strongest-performing individual ranking methods.

Our analysis focuses on the two methods that proved to be most predictive in Lasek et al., the FIFA Women's World Ranking and Eloratings.net, while also developing our own Elo-based model, B-Elo, and introducing Glicko-based rating models for consideration. We consider both the accuracy and simplicity of the models in question, under the assumption that FIFA would rather adopt a new system that is powerful and simple than one that is marginally more powerful but overly complex.

# 2 Data

Our data includes the teams, final result (after Extra Time), and location of every international football match played between two FIFA member federations from the beginning of 1998 to the end of the international match window in March 2018.[2]

For our analysis, we focused only on the last twenty years because it allows us to predict on a full, four-year World Cup cycle with information from the four previous World Cup cycles, a ratio between training and testing data commonly used in statistics. Additionally, even the longest international careers of individual players rarely surpass twenty years, which means there is barely any overlap in players between the teams playing before 1998 and those playing in 2018. However, our models can easily be altered to take into account data from any year since 1872.

| Period | Years | Number of Games |
|---|---|---|
| **Train Period** | **1997-2013** | 14,796 |
| **Test Period** | **2014** | 863 |
| | **2015** | 1,055 |
| | **2016** | 927 |
| | **2017** | 932 |
| | **2018** | 195 |
| **TOTAL** | | **18,768** |

Figure 2.1: Number of games in the dataset.

[1] "The predictive power of ranking systems in association football," Lasek et al. *Int. J. Applied Pattern Recognition, Vol. 1, No. 1, 2013.*

[2] The dataset was given to us by the owner of the website landerspiel.cmuck.de.

# 3 Methodology

## 3.1 Elo Math

The two highest performing models identified by Lasek et al. – the FIFA Women's World Ranking (adjusted for the men's teams) and Eloratings.net – are both modifications of the Elo rating system. A traditional Elo system is based on the idea that skill levels among a group of teams follows a normal distribution, with most teams falling somewhere in the middle of the bell curve. A team gains points for outperforming expectations relative to its rating and loses points for underperforming expectations relative to its rating. The basic equation for an Elo update is:

**New Rating = Old Rating + K(Actual Result - Expected Result)**

- **K**: All models considered here use a K value to calibrate ratings for better accuracy. The final K value is a product of the initial K value, Match Importance weights, and Goal Difference weights.
  - **K = (Initial K)*(Match Importance Weight)*(Goal Difference Weight)**
- **Actual Result**: Game outcome mapped to a number between 0 and 1 – usually 0 for a loss, 0.5 for a draw, and 1 for a win. However, the FIFA Women's World Ranking uses different values depending on the game's Goal Difference.
- **Expected Result**: Also a number between 0 and 1, representing the probability that a team will win the game. It is calculated from a formula that takes into account the difference in ratings between two teams and the influence of home advantage, if there is one.
  - **Expected Result** $= 1/(10^{(-Diff/400)} + 1)$
    - **Diff**: The difference in rating between two teams after taking into account home advantage.
  - **Home Advantage**: All models considered here account for home advantage by giving the home team a rating boost of **100 points** in the Expected Result calculation. This translates into a 64% probability of winning for a team playing at home against an opponent that has an equal rating.

## 3.2 Glicko Math

For our implementation of the Glicko algorithm we followed the standard two step procedure as outlined below:

The first step consists of updating each team's rating deviation with the following formula. The purpose of this update is to account for the fact that after a time period we become somewhat less certain of a team's true rating/skill. In the below equation $RD_{i-new}$ is the new rating deviation for team i, $RD_{i-old}$ is the old rating deviation for any team and $RD_{init}$ is the initial rating deviation assigned to all teams and c is a hyperparameter of the model that determines the increase in rating deviations after each period of games:

$$RD_{i-new} = min(\sqrt{RD_{i-old}^2 + c^2}, RD_{init})$$

The second step involves calculating a new rating and then another new rating deviation. The purpose of this is to incorporate the results of the period into a team's rating, and to increase or decrease the certainty of a rating, depending on how the team performed relative to their expectation. In the below formula $r_i$ is the rating for team i, $E(s_{i,j}|r_i, r_j, RD_j)$ is the expected outcome of a game for team i playing against team j, and $s_{i,j}$ is the outcome of team i playing against team j.

$$r'_i = r_i + \frac{q}{\frac{1}{RD_i^2} + \frac{1}{d^2}} \sum_{j=1}^{n} K_i \cdot g(RD_j)(s_j - E(s_{i,j}|r_i, r_j, RD_j))$$

$$RD'_i = \sqrt{(\frac{1}{RD_i^2} + \frac{1}{d_i^2})^{-1}}$$

$$q = \frac{\ln(10)}{400}$$

$$g(RD_i) = \frac{1}{\sqrt{1 + 3q^2(RD_i^2)/\pi^2}}$$

$$E(s_{i,j}|r_i, r_j, RD_j) = \frac{1}{1 + 10^{-g(RD_j)(r_i - r_j)/400}}$$

$$d_i^2 = (q^2 \sum_{j=1}^{n}(g(RD_j))^2 E(s_{i,j}|r_i, r_j, RD_j)(1 - E(s_{i,j}|r_i, r_j, RD_j)))^{-1}$$

### 3.3 WIGGO Math

For the standard Glicko implementation, the $K_i$ value inside the sigma summation is always set to 1 (i.e. it does not actually appear in the equation at all). In WIGGO, this multiplier is where we incorporate the Match Importance information. There is no initial K value in WIGGO – K is simply equal to the Match Importance weights.

### 3.4 Stephenson Math

Below is the mean update for the Stephenson algrithm, which takes into account how much a team plays and how good the teams it plays are into its creation of rankings. In the equation, all variables that are shared with Glicko are equivalent. In the equation $\beta$ is a hyperparameter of the model given to add bonus for playing more games, $\lambda$ is a hyperparameter of the model designed to determine how much teams gain from playing difficult oppenets, and $\overline{r_j}$ is the average rating of the j opponents that team i faced.

$$r' = r + \frac{q}{1/RD_i^2 + 1/d^2} \sum_{j=1}^{J} g(RD_j)(y_{i,j} - E(s_{i,j}|r_i, r_j, RD_j) + \beta) + \lambda(\overline{r_j} - r)$$

### 3.5 Prediction Calculations and Error Metrics

One very important feature of all of our tested models is that due to the way that they determine the ratings of teams, it is simple to directly predict games. In order to do this one can use the Expected Result formula shown above in the Elo math section for all of our models and it then outputs a prediction for a given game. The reason that this is particularly useful is that we can use this prediction to determine how accurate each of our models is. To do this, there are two main formulas, Mean Squared Error and Binomial Deviance. The MSE formula is:

$$\frac{1}{n} \sum_{i=1}^{n} (a_i - p_i)^2$$

where $a_i$ is the actual result of a game and $p_i$ is the predicted result of the game and n is the number of games that are being predicted/tested on. The Binomial Deviance formula is:

$$\frac{1}{n} \sum_{i=1}^{n} -(a_i \cdot ln(p_i) + (1 - a_i) \cdot ln(1 - p_i))$$

For our prediction and error calculations, we used a sliding window approach starting in 2014. First, we trained the respective models on every game prior to the period we tested on (days for Elo-based models, months for Glicko-based models). Next we made the predictions for the period we were interested in and compared those predictions to the true results to calculate our error metrics. Then we would add the period we had just tested on to the training set and trained the respective model on every game prior to the next period we tested on, including the games that we tested during the preceding period. Finally, after repeating this process until the conclusion of the March 2018 international match window, we averaged both error metrics across all games to produce our final error metric averages for each model. When calculating our average error metrics, we treated every single game in the test set equally. After all, we believe that a good FIFA Ranking should be able to predict well on games that include all of FIFA's members, from friendlies between small countries to World Cup matches.

## 3.6 Update Periods

Both Elo-based and Glicko-based models are updated periodically, and we chose to use each calendar month as an update period since that is how often FIFA updates its rankings. However, for all our Elo-based models, although we set each update period to a month, we conducted game predictions at the daily level. That means that if a team played more than one match in a given month, their rating going into each match included information from the result of the previous matches during that month. The part of the month that includes the previous matches is treated as a full month period. In other words, all our Elo-models were sensitive to immediate changes in ratings.

On the other hand, because Glicko is meant to be used on a longer period basis instead of a game-by-game basis, our Glicko-based models made predictions for all games in a given month at the start of that month, without taking into account any change in rating during that period. While making daily predictions with monthly updates does increase the accuracy of Glicko predictions, we reported results from the monthly predictions, since Glicko is not meant to be updated daily.

## 3.7 Parameters

**Standard Parameters:** All of the models in this paper have ratings that are centered around **1500 points**, which is one of the standard choices for Elo. For Elo-models, in which rating changes are zero-sum, that means that the average rating is always exactly 1500. For Glicko-models, in which rating changes are not exactly zero-sum, that means that the average rating is merely around 1500. In both cases, having a defined mean is useful when interpreting any team's strength. Furthermore, for all Glicko-based models, our initial rating deviation parameter was **300 points**, which is also one of the standard choices for Glicko.

**Given Parameters:** All of the parameters in the Eloratings.net and FIFA WWR models were given, although we did have to make very slight adjustments to both of them (in Match Importance and Goal Difference values, respectively). Additionally, because our goal was to simplify models as much as it was to improve them, we developed a simple formula for Goal Difference in our B-Elo model before we did any parameter tuning and treated it as a given.

**Cross-Validated Parameters:** After setting our standard and given parameters, we still had to identify the values that would optimize each model's performance.

- B-Elo: Initial K value; Match Importance weights
- Glicko: $c$ value for rating deviation update
- Stephenson: $c$ value for rating deviation update; $h$ value for rating deviation update; neighborhood parameter; bonus parameter
- WIGGO: Match Importance weights; Goal Difference ratio; $c$ value for rating deviation

## 3.8 Cross-Validation

To find the optimal values for all of these parameters, we ran each model under thousands of different parameter combinations and then selected the combinations that minimized the model's binomial deviance. The combinations with the smallest binomial deviance were usually the ones with the smallest mean squared error as well, although that was not always the case. We also prioritized parameter combinations that we considered simpler or more intuitive, although this metric was decidedly subjective. For example, if a model performed just as well with initial K values of 19, 20, and 21, we selected 20 as the model's final K value. Here are the ranges for each parameter that we tested:

- Initial K Values: 5 – 25
- Match Importance Weights: These values had the additional restriction that each category had to have a weight that was either equal or greater than the weight of the previous category.
  - Friendlies: 1
  - Qualifiers: 1 – 4
  - Tournaments: 1 – 5
  - World Cup: 1 – 5
- WIGGO Weighted Win Ratio: $\frac{1}{2}$, $\frac{3}{5}$, $\frac{2}{3}$, $\frac{4}{5}$, 1, and FIFA WWR "Actual" Values
- $c$ value: 0 - 25
- $h$ value: 0 - 20
- Neighborhood parameter: 0 - 5
- Bonus Parameter: 0 - 5

Below is an example of how we carried out our cross-validation (Figure 4.2). The bar graphs show the binomial deviance under different combinations of Weighted Win Ratios and Match Importance Weights in WIGGO. We consistently found that a ratio of 2/3 performed better than all other ratios we tested, including the "Actual Result" values from the FIFA WWR model. Finally, we saw that increasing the Match Importance weight of World Cup matches negatively affected predictive accuracy. Ultimately, although the (1, 2, 2, 2) weight combination performed slightly better than (1, 2, 3, 3), we judged the latter to be more intuitive in a football setting (weighing tournament matches more heavily than qualifying matches). Therefore, the final version of WIGGO is the one that has three Match Importance weights as opposed to two.

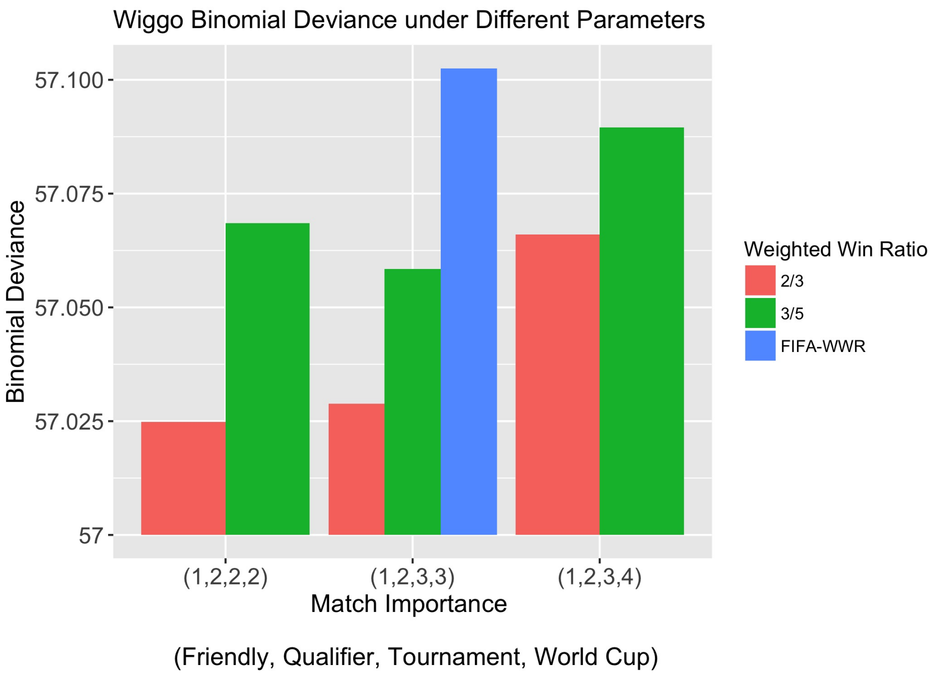


Figure 4.2

After cross-validation, these were the final parameters that we used for each model:

- **B-Elo (3 Weights)**
  - Initial K: 15
  - Match Importance Weights: (Friendly: 1; Qualifier: 2; Tournament: 3; World Cup; 3)
- **B-Elo (2 Weights)**
  - Initial K: 15
  - Match Importance Weights: (Friendly: 1; Non-Friendly (Official): 2)
- **B-Elo (1 Weight)**
  - Initial K: 20
  - Match Importance Weights: (All Games: 1)
- **Glicko**
  - $c$ value: 10
- **Stephenson**
  - $c$ value: 12
  - $h$ value: 0
  - Neighborhood parameter: 1
  - Bonus parameter: 0
- **WIGGO**
  - $c$ value: 8
  - Weighted Win Ratio: $\frac{2}{3}$
  - Match Importance Weights: (Friendly: 1; Qualifier: 2; Tournament: 3; World Cup; 3)

## 3.9 Research Outline

We will now provide an overview of each of the models we are considering, separating the Elo-based models and the Glicko-based models into separate sections. Following the overview of the models, we measure how well they were able to predict the results of matches during the period between the beginning of 2014 and the end of 2017 by comparing their binomial deviance and mean squared error metrics.

# 4 Elo-Based Models

## 4.1 FIFA Women's World Ranking

- **Initial K Value**: **15**
- **Match Importance**: **4 Levels**.

| Game Type | Match Importance |
|---|---|
| Friendly Match | 1 |
| Continental Qualifiers | 2 |
| World Cup Qualifiers, Continental Finals, Confederations Cup | 3 |
| World Cup Finals | 4 |

- **Goal Difference**: FIFA WWR uses Goal Difference to calculate the Actual Result in the Elo update equation – not as a multiplier for its K value. The table below shows the Actual Result logged for the non-winning (loss or draw) team based on goals scored and goal difference.[3] The corresponding Actual Result for the winning team is 1 minus the Actual Result for the losing team.

Goal Difference

| Goals Scored by Losing Team | 0 | 1 | 2 | 3 | 4 | 5 | 6+ |
|---|---|---|---|---|---|---|---|
| 0 | 0.50 | 0.15 | 0.08 | 0.04 | 0.03 | 0.02 | 0.01 |
| 1 | 0.50 | 0.16 | 0.089 | 0.048 | 0.037 | 0.026 | 0.015 |
| 2 | 0.50 | 0.17 | 0.098 | 0.056 | 0.044 | 0.032 | 0.02 |
| 3 | 0.50 | 0.18 | 0.107 | 0.064 | 0.051 | 0.038 | 0.025 |
| 4 | 0.50 | 0.19 | 0.116 | 0.072 | 0.058 | 0.044 | 0.03 |
| 5+ | 0.50 | 0.2 | 0.125 | 0.08 | 0.065 | 0.05 | 0.035 |

## 4.2 Eloratings.net

- **Initial K Value**: **1**
- **Match Importance**: **5 Levels**[4]

| Game Type | Match Importance |
|---|---|
| Friendly Match | 20 |
| Friendly Tournament | 30 |
| Continental Qualifiers, World Cup Qualifiers, Major Tournaments | 40 |
| Continental Finals, Confederations Cup | 50 |
| World Cup Finals | 60 |

- **Goal Difference**: Eloratings.net uses discrete values as multipliers for goal differences below 3 and switches to a formula for goal differences of 3 and above.

| Goal Difference | Multiplier |
|---|---|
| 0 | 1 |
| 1 | 1 |
| 2 | 1.5 |
| 3 | 1.75 |
| 4+ | 1.75 + [(GD – 3) ÷ 8] |

[3] We simplified the original table to set all draws equal to 0.5.

[4] Although the Eloratings.net formula distinguishes between friendly games and friendly tournaments, our dataset does not include this distinction, so we assigned all friendly games and tournaments a value of 20.

### 4.3 B-Elo

- **Initial K Value**: **15**
- **Match Importance**: **3 Levels**.

| Game Type | Match Importance |
|---|---|
| All Friendlies | 1 |
| All Qualifiers | 2 |
| All Major Tournaments | 3 |

- **Goal Difference**: B-Elo uses a simple formula to calculate a Goal Difference multiplier between 1 and 2. Draws and games decided by a single goal receive a multiplier of 1; with each extra goal, the multiplier increases by half the distance between itself and 2.

$$\sum_{i=1}^{GD} (\frac{1}{2})^{i-1}$$

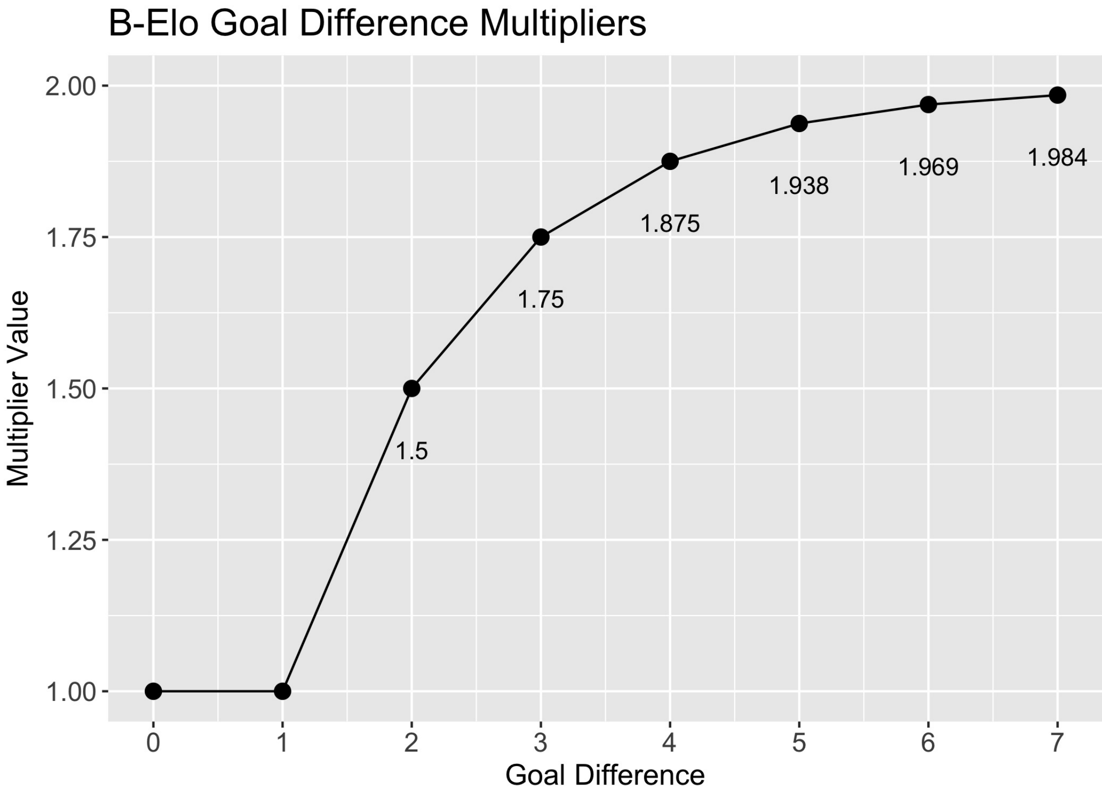


Figure 3.1

### 4.4 Results

In order to analyze these models, we tested the accuracy of their predictions each year between 2014 and 2018. First we "trained" the models on every match played between the start of 1998 and the end of 2013. Starting in 2014, we used the models to predict the outcomes of all matches on any given day. After generating these predictions, we compared them to the true results and evaluated their accuracy using Binomial Deviance and Mean Squared Error, two common metrics used to determine the accuracy of predictions. Since these metrics measure errors, smaller values correspond to better performance, although the two different metrics are not directly comparable to each other. Finally, once we calculated our error metrics, we added the true information from those games to the models before generating the next predictions, repeating that process all the way through March 2018.

Our tests showed that Eloratings.net and our B-Elo models made better predictions than FIFA WWR (Figure 3.2, 3.3). Meanwhile, B-Elo performed better than Eloratings.net in the Binomial Deviance metric, but less well on the Mean Squared Error metric. However, we also found that simplifying B-Elo further by reducing the number of Match Importance levels actually improved its prediction accuracy. Under the 2 Weight version, it only distinguished between friendly matches and official matches; under the 1 Weight version, it did not distinguish between types of games at all. Overall, we were encouraged to learn that B-Elo, which significantly simplifies Match Importance and Goal Difference calculations, was able to perform as well as or better than the FIFA WWR and Eloratings.net. Before our research, these two models had been identified by Lasek as the two best-performing models in international football.

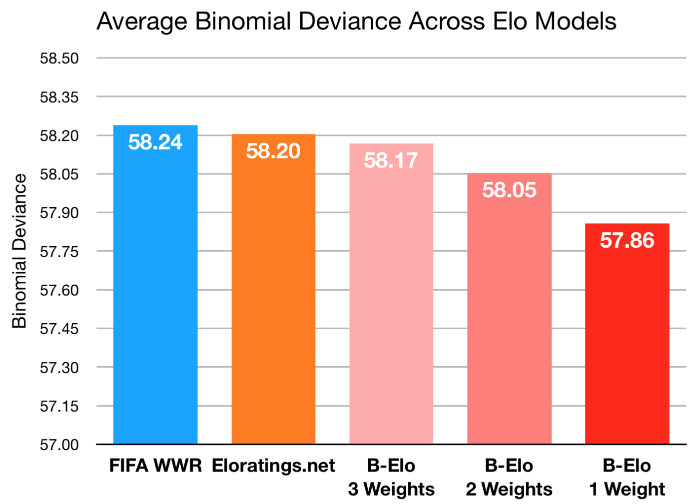


Figure 3.2

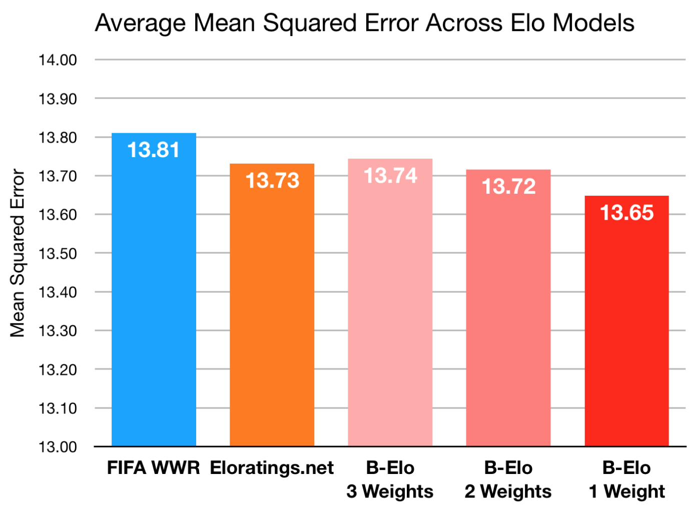


Figure 3.3

## 4.5 Conclusions

- **The B-Elo ranking model offers a simpler alternative to the current best-performing ranking models without losing any predictive power.**
- **Specifically, B-Elo is simpler *and* more predictive than FIFA WWR.**
- **If FIFA plans to adopt an Elo-based model for its Men's Ranking, we suggest they adopt B-Elo instead of an adaptation of the FIFA Women's World Ranking.**

# 5 Glicko & WIGGO Models

The Glicko model extends the Elo model by introducing a *rating deviation* to convey uncertainty around a team's rating. The rationale for introducing a measured uncertainty is that we can never know a team's *exact* strength because there are many extraneous factors that can affect it, but it is possible to say that we are more certain about some teams than others. This is especially true in international football, where teams only play a handful of games each year and rosters can change completely from month to month and the amount of games teams play is very variable.

The equation behind Glicko is similar to the Elo equation, although the introduction of the rating deviation makes the math a little more complicated. Teams that play more games have a lower rating deviation, since we are more certain of their true strength. The number of points a team wins or loses each game depends partially on its rating deviation; since we are more certain of the strength of a team with lower deviation, that teams sees their rating fluctuate less with each game than teams with higher deviations.

Finally, rather than updating after each game, Glicko is updated after a period of time (in this case after each month). Due to the uncertain nature of international football, as well as the fact that Glicko is are updated monthly, we believe Glicko-based models offer the best way to rank international football teams.

## 5.1 Glicko-1

As a baseline, we tested the regular Glicko-1 model that does not take into account any information regarding Match Importance or Goal Difference.

## 5.2 Stephenson

In addition to the standard Glicko-1 algorithm, we tested a model developed by Alexander Stephenson that builds on the Glicko model by including three extra parameters: an additional uncertainty parameter that acts similarly to the $c$ parameter in Glicko, a neighborhood parameter that shrinks a team's rating toward its opponent's rating independently of the result, and a bonus parameter that awards teams points simply for playing games under the assumption that teams improve slightly with every game they play. However, our parameter cross-validation showed that the only parameter that actually improved predictions was the neighborhood parameter, so we set the bonus parameter and the additional uncertainty parameter to zero. Adding a bonus parameter would have inflated ratings over time, making it hard to compare team strength diachronically, so not including it was also the best decision in terms of model simplicity.

To give Stephenson the best chance of success, we also tested an alternate version that incorporated Match Importance and Goal Difference information in the same way that WIGGO does once we had already optimized all of the parameters for the regular Stephenson model. While this additional information did improve the Stephenson predictions, it still did not change the final ordering of which models were the most predictive.

## 5.3 WIGGO (Win, Importance & Goal-adjusted GlickO)

In addition to the standard version of Glicko, we developed an "informed" version of Glicko, called WIGGO, that incorporates Match Importance and Goal Difference information into its calculations, much like the Elo-based models. Match Importance acts as a multiplier that takes the place of K, while Goal Difference is used to determine the value of a win.

- **Match Importance**: **3 Levels**.

| Game Type | Match Importance |
|---|---|
| All Friendlies | 1 |
| All Qualifiers | 2 |
| All Major Tournaments | 3 |

- **Goal Difference**: WIGGO incorporates Goal Difference into its calculation for the 'Actual Result' in the rating update equation – not as a multiplier for its K value (like FIFA WWR). Figure 4.1 shows the 'Actual Result' logged for the non-losing (win or draw) team based solely on goal difference. With each extra goal, the win value increases by two thirds of the distance between itself and 1.

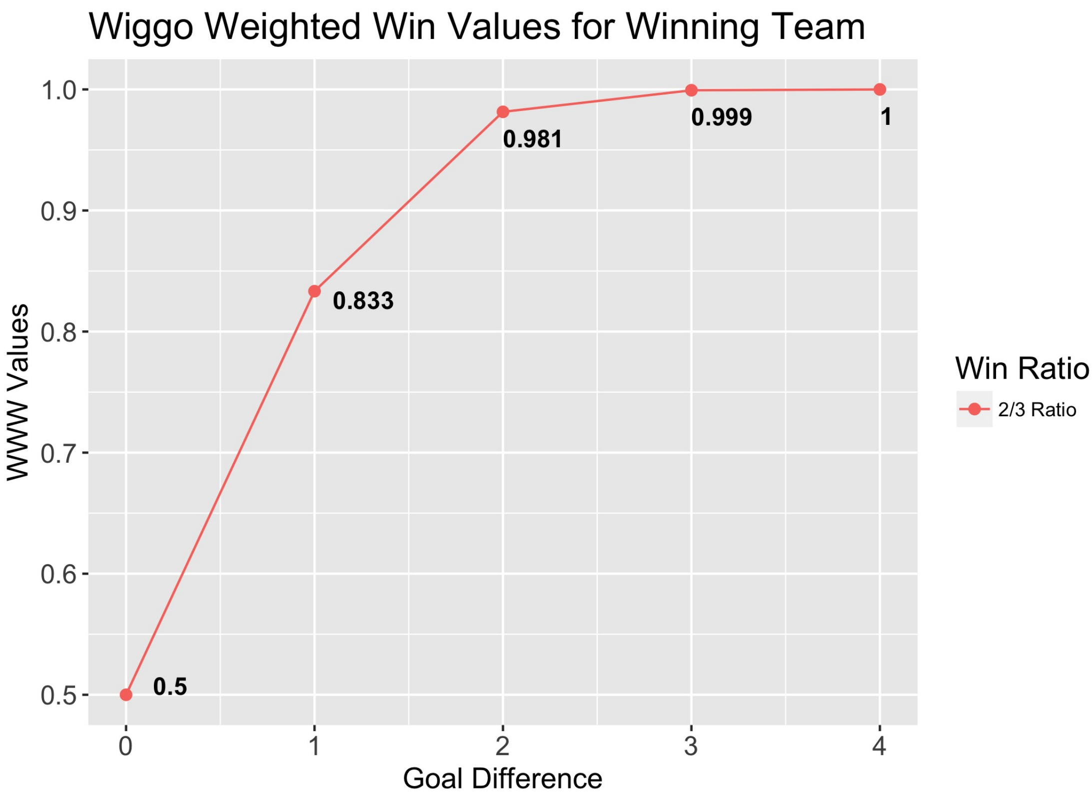


Figure 4.1

## 5.4 Results

Although the math behind Glicko-based models is slightly more complex, the result is more predictive power than the Elo-based models. Even the basic Glicko model (green), which includes parameters for home advantage, but not for Match Importance or Goal Difference, was able to outperform all Elo-based models on both predictive accuracy metrics every year (Figure 4.4, 4.5). Meanwhile, our WIGGO model (black) outperformed all models, including Glicko and Stephenson, on both metrics during every full year of testing. Thus, while only slightly more complex with its inclusion of rating deviations, an informed Glicko-based model that does consider Match Importance and Goal Difference provides the most accurate predictions.

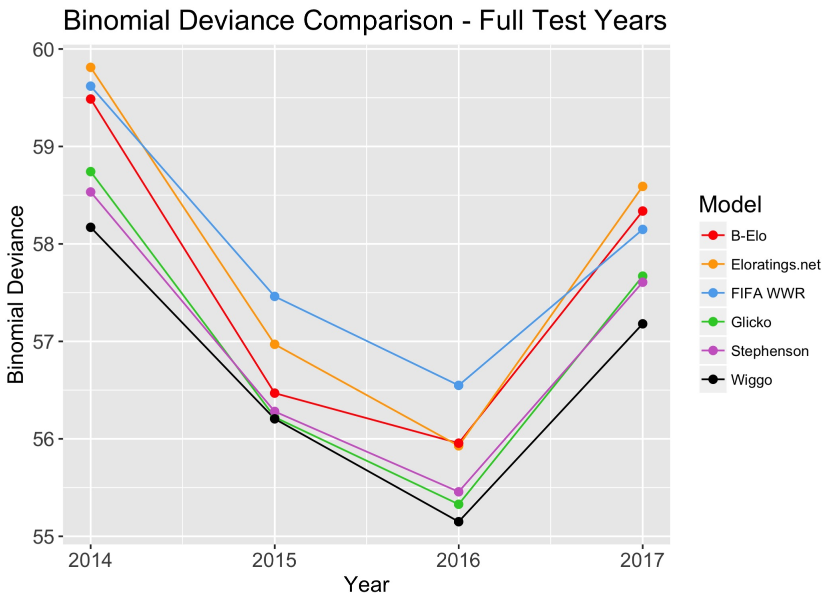


Figure 4.3

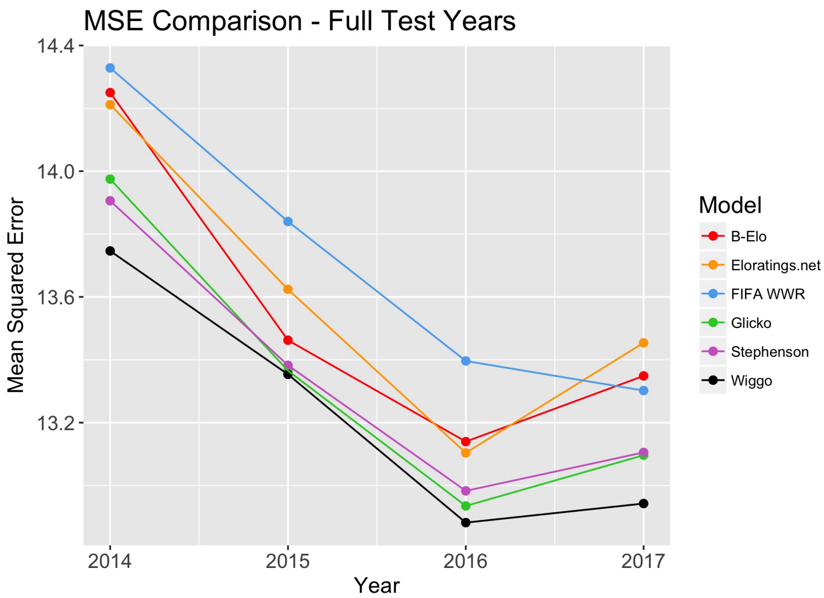


Figure 4.4

## 5.5 Conclusions

- **Glicko models include uncertainty (rating deviance) into their rating calculations.**
- **The math is more complex, but the models gain more predictive accuracy.**
- **The WIGGO ranking model offers a more powerful alternative than the current best-performing ranking models.**

# 6 Analysis of Results

Our results indicate that there exists a trade-of between simplicity and predictive power: The simpler a model is, the less predictive power it has, and the more complex it is, the more predictive power it has. We believe that our two models, B-Elo and WIGGO, strike a nice balance between simplicity and predictive power.

B-Elo is based on a simple mathematical formula with intuitive components like home advantage, Goal Difference, and Match Importance. By choosing the right parameters, B-Elo actually performs better than more complicated Elo-based models. The WIGGO model is slightly more complex but in return yields much more accurate predictions than any of the Elo-based models and even the standard Glicko. The WIGGO model incorporates the same information as B-Elo, with the addition of the rating uncertainty. Thus, while the math is slightly more complicated, the concepts remain relatively easy to grasp.

An additional advantage of these models is that their rankings are more consistent across time than the FIFA Men's Rankings. We use two metrics to explore this: the average absolute change in ranking for a team from month to month, and the second of which is the average variance of a teams ranking across all of the months that they are ranked (Figure 5.1, 5.2). Looking at these values we can see that WIGGO has both the lowest variance and average monthly change in ranking, indicating that WIGGO is the most stable ranking model. Its stability reinforces WIGGO's standing as a generally accurate ranking, as the adjustments it needs to make from month to month are minor. The FIFA Men's Ranking, on the other hand, exhibits the most variability, with more than twice as much variance and average ranking monthly change than any of the other models. These characteristics suggest the FIFA Men's Ranking is too sensitive and even erratic. Finally, the Elo-based models are more consistent than the FIFA Men's Ranking but less consistent than WIGGO; between them, B-Elo exhibited slightly less variability than FIFA WWR.

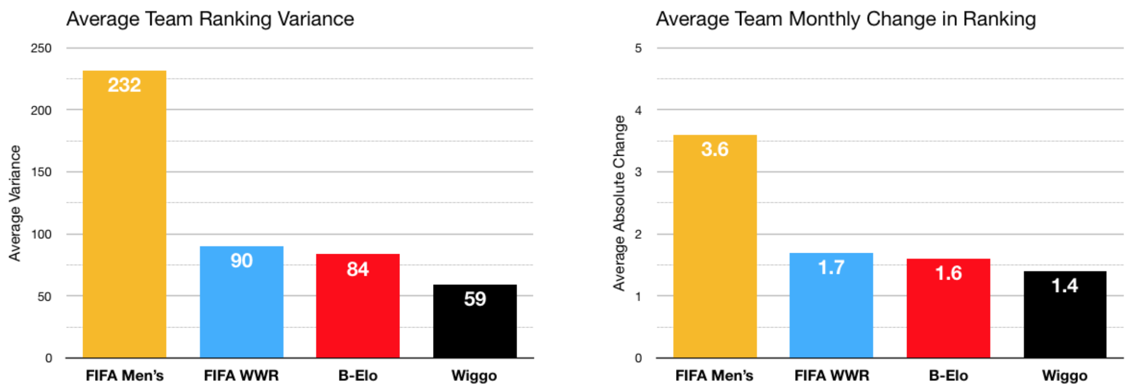


Figure 5.1

If FIFA decides to pursue an Elo-based model for its Men's Ranking, then we have developed a version in B-Elo that greatly simplifies calculations while at the same time increasing predictive power. If FIFA decides to seek other alternatives, we have developed a model in WIGGO that is slightly more complex but also more powerful than any ranking model that has ever been used in international football.

## 6.1 Conclusions

- **B-Elo and WIGGO optimize the balance between simplicity and predictive accuracy in different ways.**
- **In addition to being the most accurate model, WIGGO is also the most stable and consistent, followed by B-Elo.**
- **We suggest FIFA consider adopting WIGGO or B-Elo as its new ranking method.**

# 7 Mexico Case Study

To better understand how the different models work, we will look at the way each of them has affected Mexico's national team. The models we focus on in this section are the FIFA Men's Ranking, FIFA WWR, and our own two models, B-Elo and WIGGO. First we track Mexico's monthly ranking since the start of 2014 under all the models, and then we compare Mexico's WIGGO ranking to that of its CONCACAF peers during the same time period. Finally, we look at the effect of the 2017 Confederations Cup on Mexico's rating from both game-by-game and comprehensive perspectives.

## 7.1 Ranking Comparison Across Models

Looking at Mexico's FIFA Men's Ranking provides a good example of the variability of that ranking's variability, discussed in Section 5. Mexico's ranking changes more wildly under the FIFA Men's model than it does under any of the others, including an inexplicable drop of 17 places and recovery of 14 places in the span of two months in the summer of 2015. By contrast, B-Elo and especially WIGGO seem to fluctuate much less than the two FIFA models. Lastly, all three alternative models ranked Mexico more highly than the FIFA Men's Ranking during the entirety of the last World Cup cycle, suggesting that Mexico is underrated by the current FIFA Ranking.

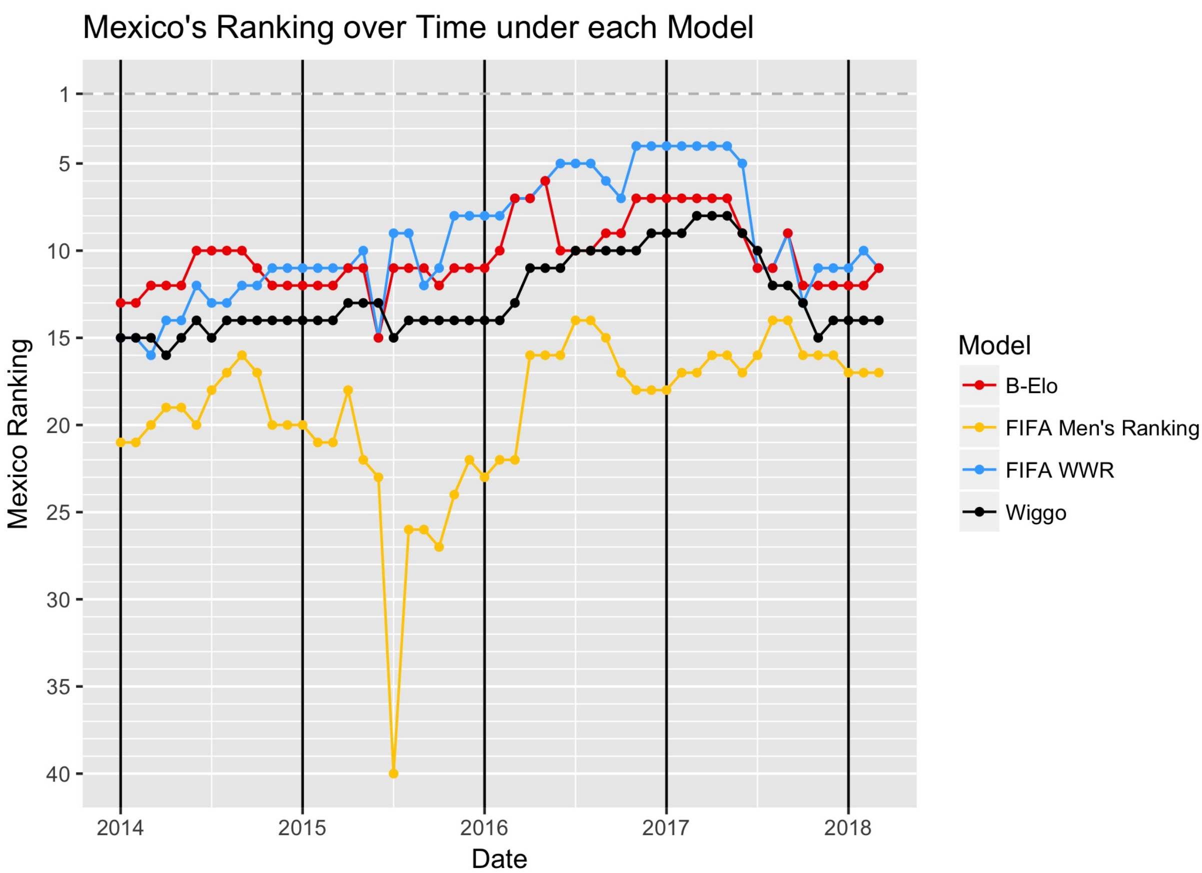


Figure 6.1

## 7.2 WIGGO Comparison Across Teams

Figure 6.2 shows Mexico's WIGGO ranking across the last four years next to the ranking of other relevant teams: two of its Concacaf neighbors (USA & Costa Rica) and the three teams that have held the top ranking spot (Argentina, Brazil, and Germany). Here are some things to note that are visible in the plot:

- **Germany** moved into the top spot after winning the 2014 World Cup and held it until the start of 2016.
- **Brazil** saw their ranking drop after the 2014 World Cup in part because of their heavy losses against Germany and Netherlands in their last two matches. They held the top spot for most of 2017 in part because of their strong World Cup qualifying performances.
- **Costa Rica** saw their ranking improve massively and immediately following their performance in the 2014 World Cup.
- **USA** has seen a large drop in their ranking from 2015 in part due to their poor performances in World Cup qualifying.

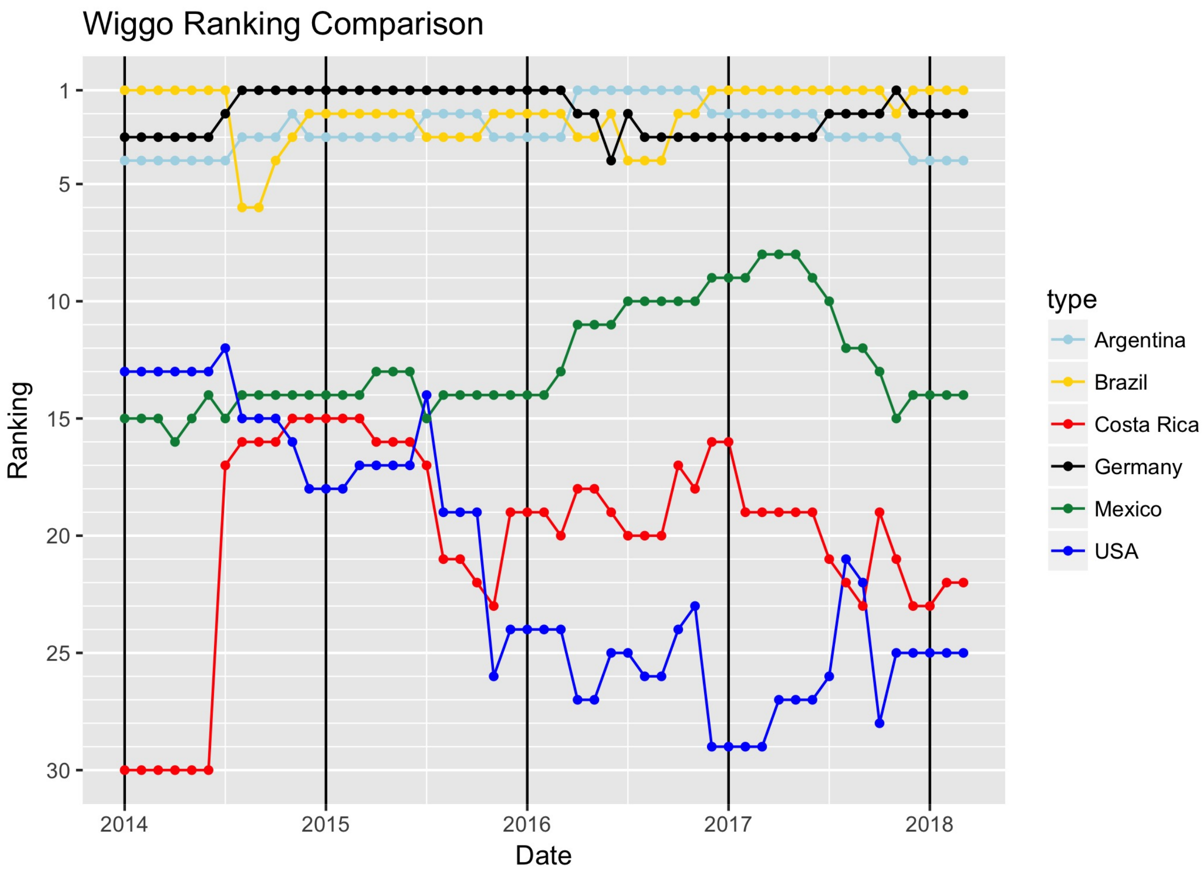


Figure 6.2

## 7.3 2017 Confederations Cup

Figure 6.3 shows the effect of each game during the 2017 Confederations Cup on Mexico's rating under FIFA WWR, B-Elo, and WIGGO. It is worth noting that, during the tournament, Mexico had a second team preparing for the Gold Cup, so the games shown include two friendly games played in the U.S. by the second team. These results sheds some light on how the models work:

- **Portugal**: Because Mexico had a higher rating than Portugal in every model, all the models slightly favored Mexico in the opening game. That is, every model assigned Mexico more than a 0.5 win probability. Drawing against Portugal, which represents exactly 0.5 wins, meant Mexico *under*performed relative to their rating and therefore lost points for drawing.
- **Match Importance**: Because of Match Importance weights, friendly wins against Ghana and Paraguay yielded less points than tournament wins against Russia and New Zealand. Every model here assigns friendlies a weight of 1 and tournament games a weight of 3.
- **Ghana**: In fact, the win over Ghana yielded almost no points under WIGGO. That is because WIGGO favored Mexico to win almost exactly by one goal, so Mexico neither underperformed nor overperformed relative its rating, causing it to remain the same.

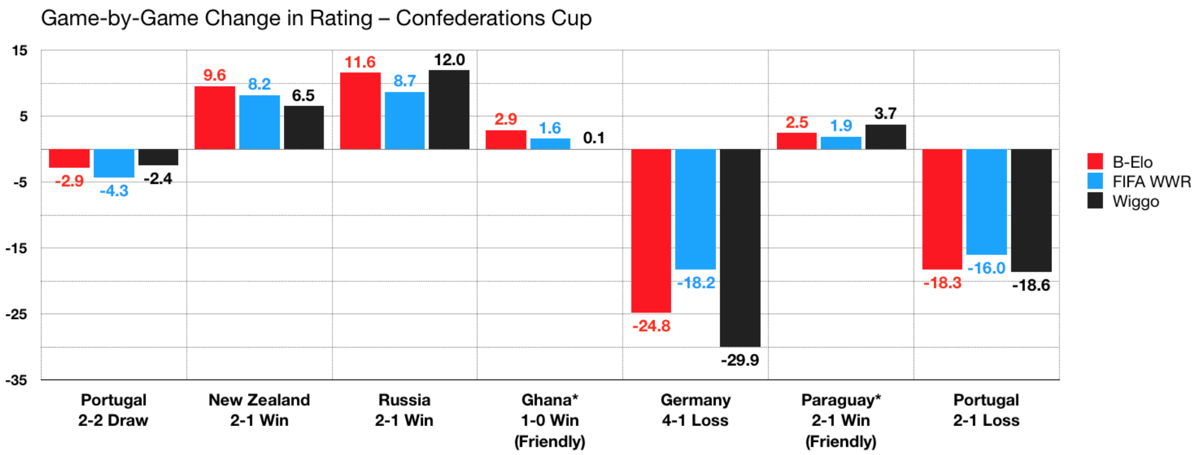


Figure 6.3

On the whole, Mexico underperformed relative to its rating during the entire Confederations Cup. They were twice favored over Portugal by every model and failed to beat them, and they lost to Germany by a bigger margin than any model expected, so they had less rating points by the end of the tournament than they did at the beginning (Figure 6.4).

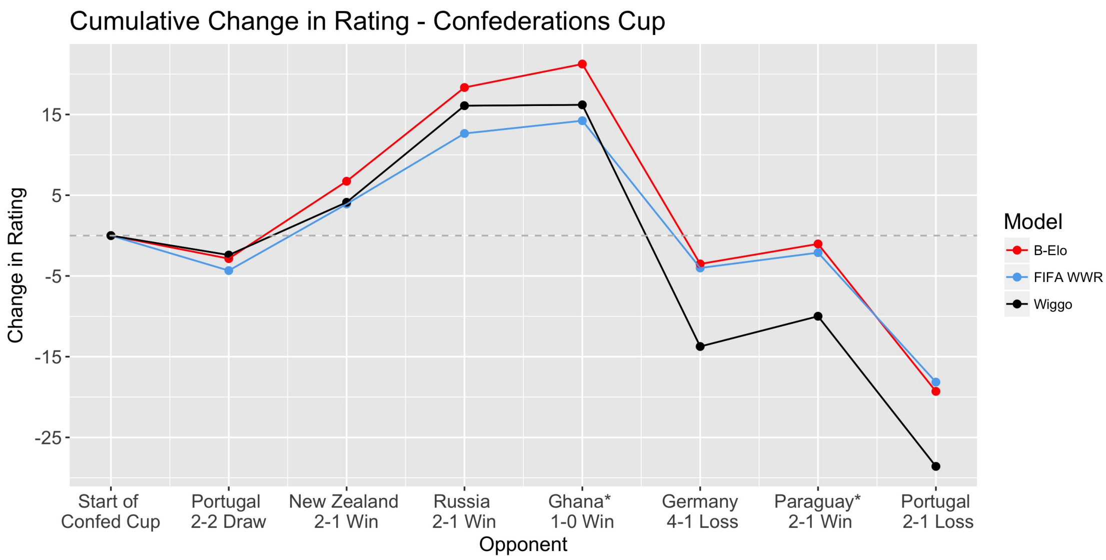


Figure 6.4

Since the FIFA Men's Ranking assigns points on a different scale and is only updated when it is released by FIFA , we cannot directly compare the change in rating across these models to the change in points on the FIFA Ranking. However, we can calculate the percentage drop in rating over the span of a month, across all the models. Figure 6.5 shows that Mexico's drop in rating was most dramatic under the FIFA Men's Ranking. The FIFA WWR model, as well as our B-Elo and WIGGO models, resulted in smaller drops in rating that were clustered relatively close together following Mexico's performance.

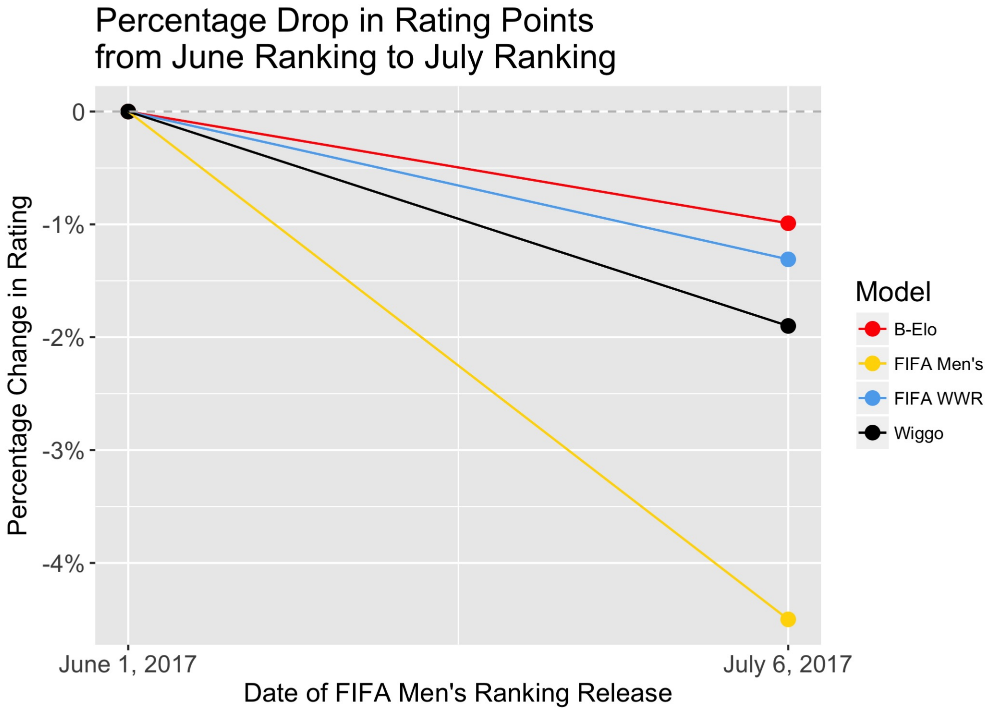


Figure 6.5

# 8 Ranking Comparisons

FIFA Ranking - March 15, 2018

| Rank | Team | Points |
|---|---|---|
| 1 | Germany | 1609 |
| 2 | Brazil | 1489 |
| 3 | Portugal | 1360 |
| 4 | Argentina | 1359 |
| 5 | Belgium | 1337 |
| 6 | Poland | 1228 |
| 7 | Spain | 1228 |
| 8 | Switzerland | 1197 |
| 9 | France | 1185 |
| 10 | Chile | 1161 |
| 11 | Peru | 1128 |
| 12 | Denmark | 1108 |
| 13 | Colombia | 1106 |
| 14 | Italy | 1062 |
| 15 | Croatia | 1053 |
| 16 | England | 1047 |
| 17 | Mexico | 1038 |
| 18 | Iceland | 1026 |
| 19 | Sweden | 1002 |
| 20 | Wales | 984 |

WIGGO Ranking - March 2018

| Rank | Team | Points | Change from FIFA |
|---|---|---|---|
| 1 | Brazil | 1926 | 1 |
| 2 | Germany | 1924 | -1 |
| 3 | Spain | 1876 | 4 |
| 4 | Argentina | 1866 | – |
| 5 | Portugal | 1858 | -2 |
| 6 | France | 1825 | 3 |
| 7 | Colombia | 1810 | 6 |
| 8 | Belgium | 1806 | -3 |
| 9 | England | 1795 | 7 |
| 10 | Netherlands | 1785 | 11 |
| 11 | Italy | 1781 | 3 |
| 12 | Peru | 1780 | -1 |
| 13 | Chile | 1778 | -3 |
| 14 | Mexico | 1776 | 3 |
| 15 | Uruguay | 1770 | 7 |
| 16 | Croatia | 1765 | -1 |
| 17 | Switzerland | 1755 | -9 |
| 18 | Poland | 1742 | -12 |
| 19 | Iran | 1738 | 14 |
| 20 | Sweden | 1733 | -1 |

Figure 7.1: Side-by-side comparison of March Rankings

# 9 Possible Uses

FIFA uses their current ranking system to determine the pools in which teams will compete in the World Cup. The pools are created to contain teams that vary in potential, thus favoring the highest ranking teams to win. When the FIFA Ranking under performs, "groups of death" occur, where top teams are placed in the same group, thus ending the tournament for some teams that should have advanced to later rounds. A better ranking system would minimize the chance of ultra-competitive groups, resulting in more entertaining, highly competitive championship rounds.

Individual teams could also attempt to maximize their ranking by scheduling games with the highest predicted rank increase. Given the probability of winning from WIGGO, and the ranking points that would be awarded, any team could plan a game schedule to maximize their expected value of ranking points. Iran unexpectedly is ranked at 19 in our WIGGO ranking due to their success in tournament qualifying games. They have not proven to be an exceptionally talented team, but because they succeeded in games with high match importance values, their ranking is high. Other teams could strategically follow in their lead.

Regardless, there are still some barriers. We still need to explore how well the prediction equation maps to results. Even WIGGO is off by about 30% per game on average. Thus, teams looking to optimize their schedule to maximize their rank may not always succeed.

# 10 Limitations and Further Research

When training our models, we included only the games from 1998 to 2018 in order to train on 20 years of data. We believed that games prior to 1998 would not contribute meaningful information to current rankings, however we were incorrect. Including data before 1998 improved the model. Thus, for further study we could find the time span for which to train the models to optimize model performance.

We also chose to train our models on all game types from friendlies to World Cup Finals. We could explore a different focus in which we create different models to predict different game types. For example, we could train our model on tournament games only and determine if the predictions for tournament results are more accurate than when we train on all game types.

The focus on World Cup games is paramount to FIFA as it is their largest event. As mentioned, we would like to compare how accurate our WIGGO predictions are when compared to real results. Thus, with the 2018 World Cup approaching, we would like to compare results in order to gauge how future teams could utilize our ranking, and also how much the WIGGO ranking changes after the event.