Measuring US Chess Rating Uncertainty

Scott Sussex and Mark E. Glickman
Harvard University

April 22, 2019

1 Introduction

The Ratings Committee (RC) was tasked by the US Chess Executive Board (EB) on May 19, 2018, with exploring the computation involved in constructing a measure of variability/uncertainty in a player's rating. This task was welcomed by RC members who responded to the task, acknowledging that providing a measure of variability would be useful for players and tournament directors to characterize the reliability of a rating. Given that many online implementations of chess servers and other online gaming leagues have developed and adopted rating systems that are accompanied by uncertainty measures (including, for example, Microsoft's TrueSkill rating system, among others), US Chess would be joining a forward-thinking group of gaming organizations that would be using data science ideas in the rating system, one of its important services to tournament players. The goal of the work was to produce a summary measure that would accompany a rating when published. Rather than a player seeing their rating appear as, say, "1593," they would see their rating appear as "1593 (65)," where the value in the parentheses would indicate a measure of uncertainty such as a standard error for the rating. Larger standard errors indicate greater uncertainty.

Mark Glickman, the chair of the RC, agreed to explore the computation involved with

measuring ratings uncertainty and brought along one of his students, Scott Sussex, to help with the development. This document summarizes the results of the work.

2 Guiding principles

Our approach to measuring rating uncertainty followed the following three guiding principles.

1. The greater the frequency of games played, typically the less uncertainty in the rating.
2. The more time that has elapsed since the last time a player has competed, all else being equal, the greater the uncertainty in the player's rating.
3. The more a player's rating increases, decreases, or generally varies over time, the greater the uncertainty.

Our approach requires determining certain parameters of the algorithm up front, but once these parameters are determined the rest of the algorithm is straightforward to run on an ongoing basis. Furthermore, the approach is easily implementable, and should not be difficult for US Chess to incorporate into the ratings code.

3 Basic Approach

A typical measure of the uncertainty of an estimate is the *standard error*. For the most basic setting of reporting the uncertainty of a sample mean, the standard error can be understood as a two-step calculation: (1) calculate the average squared deviation of each observed value from the sample mean, and then (2) divide this value by the sample size, and take a square root of the final value. This method assumes that the observation values are obtained

independently and at random. Our proposed approach to determining the standard error of a rating is similar, but acknowledges that the observations, i.e., a player's ratings, are not independent of each other; a player's most current rating is very much dependent on their previous rating. Also, the recency of ratings is relevant for measuring uncertainty. If a player competes in several events over the last month or two, we can assume that the ratings provide more information about current ratings uncertainty than if the player has competed in several events most recently two years ago.

Our basic approach is to compute the equivalent of a standard error for a player's rating based on their recent history of computed ratings, but accounting for the correlation between ratings and for the timing of ratings. We assume that we have for each player a history of rating values that usually are recorded after each rated event, and we keep the most recent three years' worth of ratings (though in production this number of years may change). For each rating over the past three years, we compute a weight (from a formula explained below) that equals 1.0 if the rating were computed today, and gradually decreasing close to 0.0 for ratings three years ago. To obtain the standard error for a rating, we compute the (1) time-weighted mean rating, and then the (2) weighted standard deviation of ratings around the time-weighted mean rating. This weighted standard deviation is then divided by a factor that accounts for the correlation of ratings, which generally inflates the weighted standard deviation. Finally, to convert to a weighted standard error, the weighted standard deviation is divided by the square root of the sum of the weights.

4 Algorithm Specification

The following description explains the computation to obtain an estimate of the current standard error of a rating. Consider a player who has competed n events in the last (say) three years. Assume for the moment that $n > 0$, though below we consider when $n = 0$. For

event i, $i = 1, \ldots, n$, let t_i be the number of days before now/today when event i took place; larger values of t_i correspond to events occurring longer ago. Let r_i be the post-event rating for event i.

Define w_i to be the weight of event i in the computation. We let

$$w_i = \exp(-\gamma t_i) \tag{1}$$

for event i, where γ is a tuning parameter. When $t_i = 0$ (i.e., the event was today), the value of $w_i = 1$. For $t_i > 0$, the value of w_i is less than 1, which reflects the lower relevance of event i in the computation of the current standard deviation. Based on our work, we have set $\gamma = 0.001$, though this value will be revisited at implementation time. This value of γ can be estimated through cross-validation on a set of actual player ratings to minimize the average squared deviation between a current rating and its time-weighted estimate.

We now define the weighted mean rating, and the weighted variance of the rating. We first describe the procedure in its raw form, but below explain an important modification to stabilize the computation. The weighted mean rating over the past three years is given by

$$R = \frac{\sum_{i=1}^{n} w_i r_i}{\sum_{i=1}^{n} w_i}, \tag{2}$$

and the weighted rating variance is given by

$$V = \frac{\sum_{i=1}^{n} w_i (r_i - R)^2}{\sum_{i=1}^{n} w_i}. \tag{3}$$

The mean and variance in (2) and (3) are the average rating and the variance of ratings over the past three years, but weighted by the events' recency.

The standard error, accounting for auto-correlation C among successive ratings, is given by

$$S = \sqrt{\frac{V}{(\sum_{i=1}^{n} w_i)(1 - C^2)}} \tag{4}$$

where the auto-correlation C is a tuning parameter. The value we determined for C was 0.73, though again we will revisit this value at implementation time. This value of C came from averaging the auto-correlation among a large set of players' ratings. The value of S is the final standard error to report.

Because some players may have very few (or no) games in the past three years, we regularize the computation in (3) in the following manner. We assume that a phantom game was played at the earliest point in the three year time window, assuming no games were played. If at least one game was played, we assume the phantom game occurred one day prior to the earliest game. Let t_0 be the number of days ago the phantom game occurred, so that the weight of the game is $w_0 = \exp(-\gamma t_0)$. We assume that the contribution of the game to (3) is to average in a squared deviation corresponding to the initial uncertainty in a player's rating. We assume this squared deviation to be $d^2 = 300^2$. The revised formula is therefore

$$V = \begin{cases} d^2 & \text{if } n = 0 \\ \frac{w_0 d^2 + \sum_{i=1}^{n} w_i (r_i - R)^2}{w_0 + \sum_{i=1}^{n} w_i} & \text{if } n > 0 \end{cases} \tag{5}$$

This is followed by applying the revised version of (4), that is,

$$S = \sqrt{\frac{V}{(w_0 + \sum_{i=1}^{n} w_i)(1 - C^2)}} \tag{6}$$

It is worth noting that when games are played during the three year window, the impact of the phantom game is minimal when many games are played.

5 Simulations

We demonstrate the proposed method on two simulated players to demonstrate the behavior of our approach. In each simulation scenario, a player has competed in 150 1-game events over a span of five years. The player's rating is updated according to the core Elo rating updating algorithm, which is a simplified version of the US Chess formula (but with similar behavior). Opponents' ratings are generated by simulating from a normal distribution with standard deviation of 200 centered at the player's current computed rating, which varies over time based on previous game results. Binary game results and the implied rating changes are simulated using the Elo winning expectancy formula based on the player's true ability as specified in the simulation scenario.

In the first simulation, the player is assumed to have a constant ability of 1500 across the 150 games. The results are shown in Figure 1. The rating changes, as shown in the bottom panel in red, vary between 1460 and 1540, which might be expected of a player who is playing consistently at the level of a 1500 player. The standard error measure in the top panel starts out large (given no previous information), but quickly reduces to a standard error that generally stays around 15-20. The small increase in the standard error around Day 1000 corresponds to the period over which the player's rating has been decreasing steadily from Day 200 (with a rating of around 1550) to a rating of 1450, and then up slightly, corresponding to some variation in rating over time.

The second simulation involves a player whose true ability is 1500 until game 75, at which point the player's ability increases linearly to 1700 by game 100. During the last 50 games the player's ability is at 1700. This simulation might reflect a player whose ability is increasing gradually over time. The results of this simulation are shown in Figure 2. The player's rating begins to climb significantly after Day 1000, and the increase is reflected in an increase in the standard error over time in the top panel of the figure starting Day 1200. The larger

standard error reflects the added uncertainty in the player's true ability towards the end of the 150 games.

6 Performance with Actual Players

We selected three actual players with established ratings who have competed actively between May 2014 to May 2017 to demonstrate the performance of our approach. The three players differ in average rating; the first has ratings that start around 1900 but decrease to around 1820, the second player has ratings that stay between 1960 and 2010, and the third player has ratings that stay near 850. We show the results of our method on games from May 2014 to May 2017. Games played before May 2014 (starting at May 2013) are included in our standard error calculations, but are not shown in the accompanying figures.

In Figure 3, the player's rating seems to be stable at around 1920 for awhile, but then starts to decrease. In the top panel, the standard error decreases over this time, reflecting the consistent ratings. When the player's rating drops to 1830 and then increases, and then reverts back to a level of about 1820, the standard error begins to increase.

Figure 4 shows a player whose rating is generally jumping around between 1960 and 2000, which is a fairly confined range. The standard error reflects this consistency, and mostly decreases over this range. This is a player whose performances are fairly stable.

The final player is represented in Figure 5. This player has not played as many events as the previous two. Because the player has not competed as frequently, the standard error in the top panel of the figure is a bit larger in magnitude (starting at 180, and decreasing to around 60). The player's performances are generally consistent in the 800 to 900 rating range, and this is reflected in a standard error that is decreasing somewhat down to around 60.

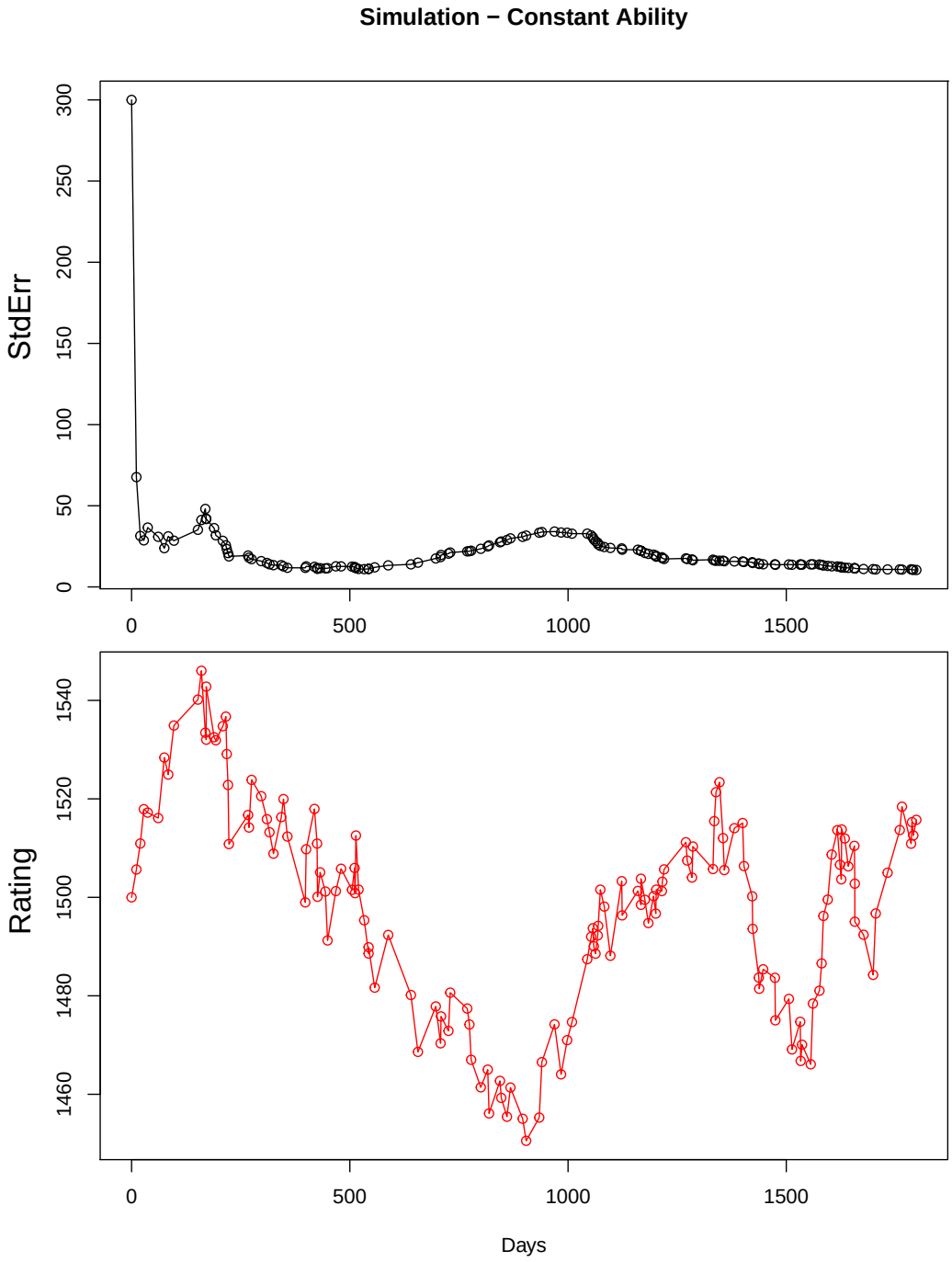

Figure 1: Player with fixed ability playing 150 games over a 5-year span. Top panel is the standard error computation over time, and the bottom panel is the player's computed rating over time.

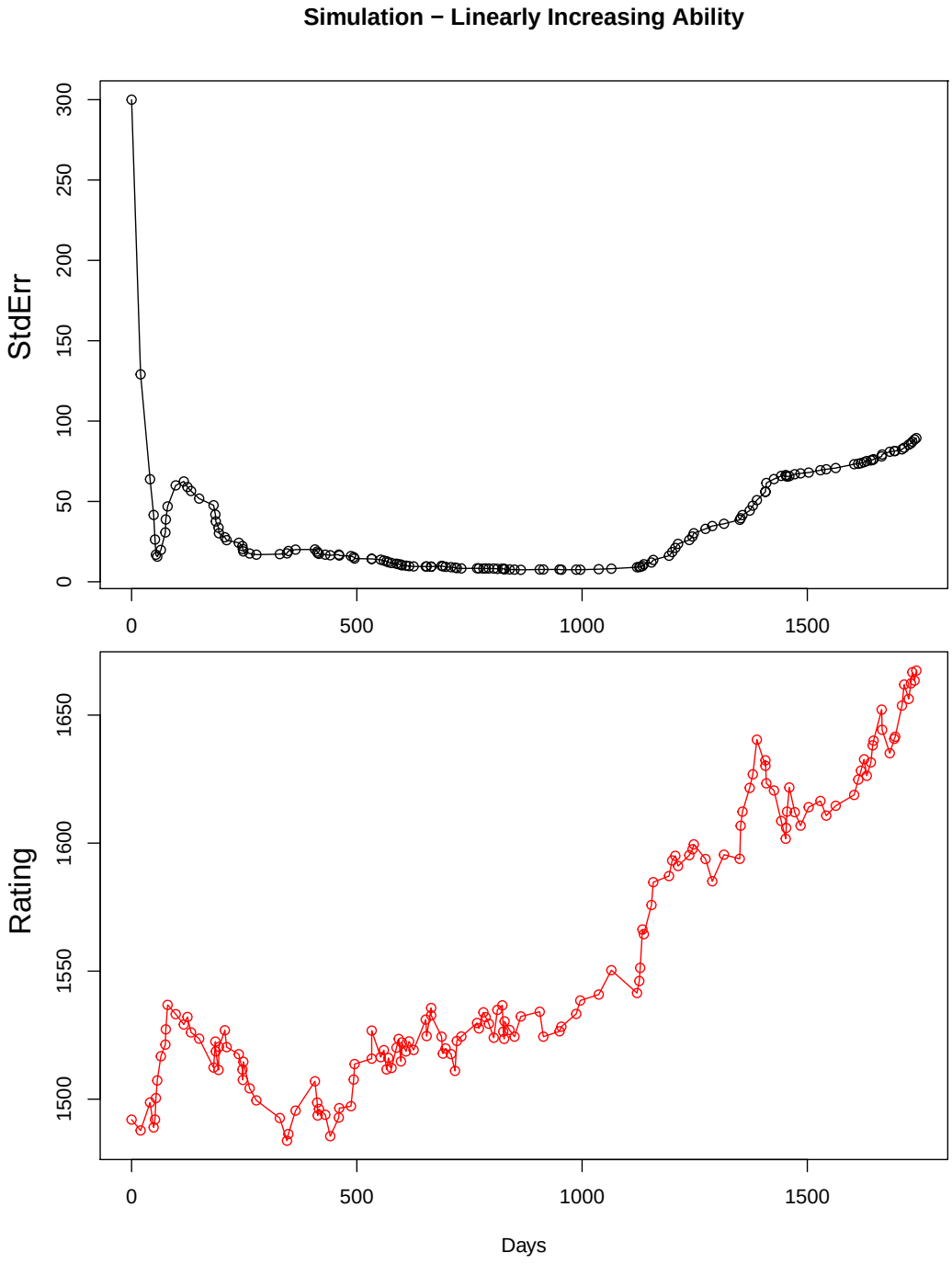

Figure 2: Player with ability fixed at 1500 for the first 75 games, then increasing linearly to a 1700 ability until the 100th game, and then fixed for the remaining 50 games. Top panel is the standard error computation over time, and the bottom panel is the player's computed rating over time.

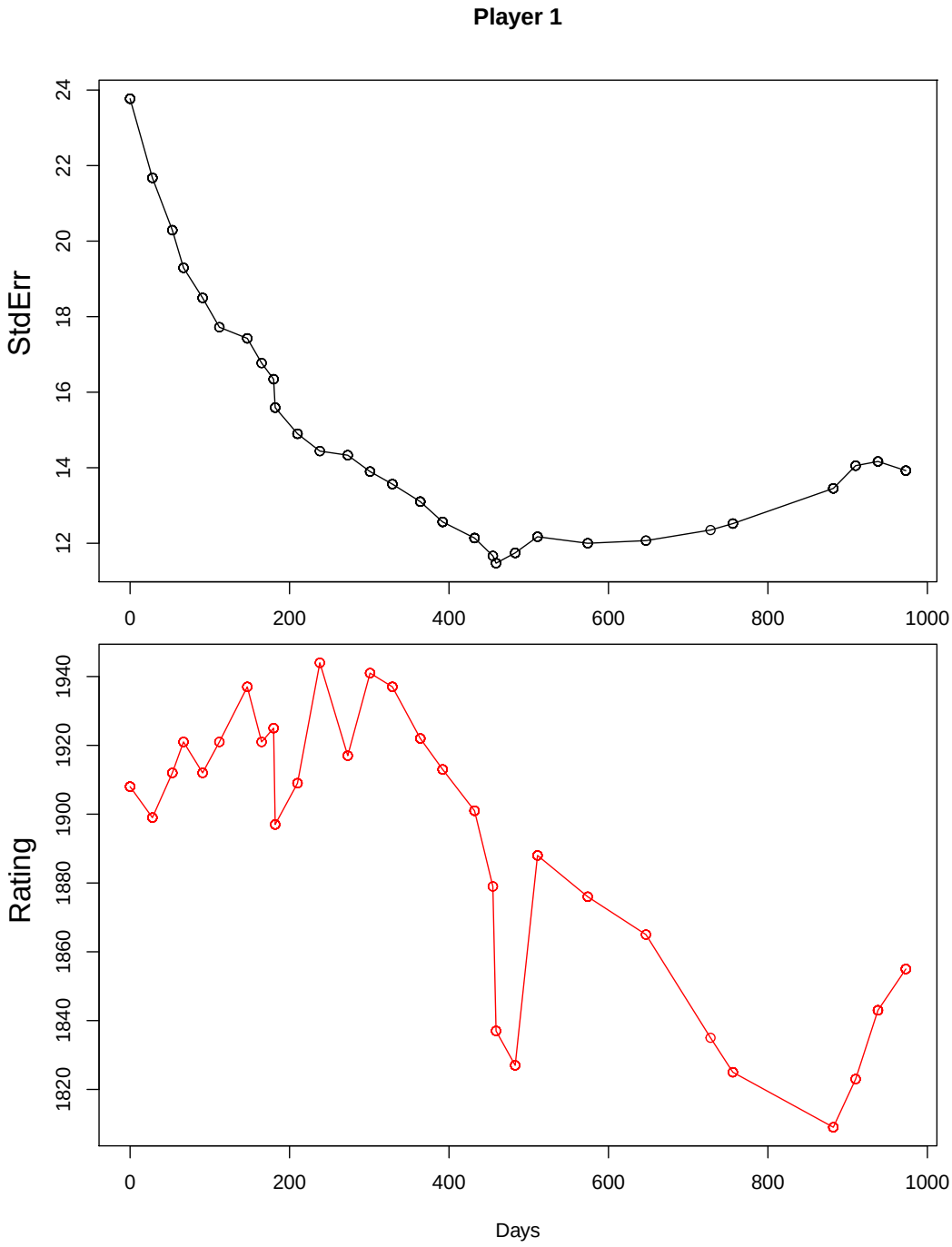

Figure 3: Standard error calculations for a randomly selected player for events rated between May 2014 and May 2017. Top panel is the standard error computation over time, and the bottom panel is the player's computed rating over time.

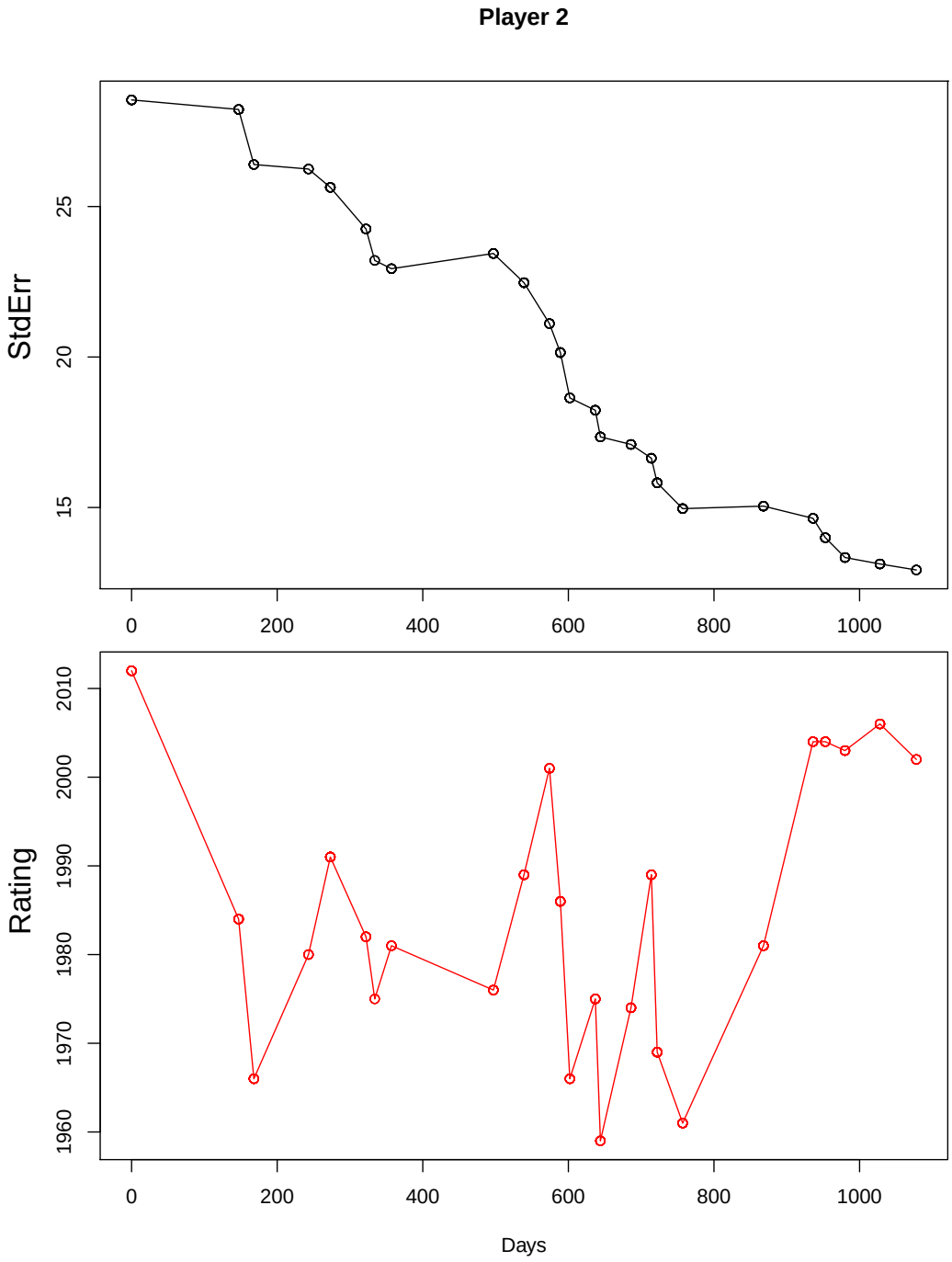

Figure 4: Standard error calculations for a randomly selected player for events rated between May 2014 and May 2017. Top panel is the standard error computation over time, and the bottom panel is the player's computed rating over time.

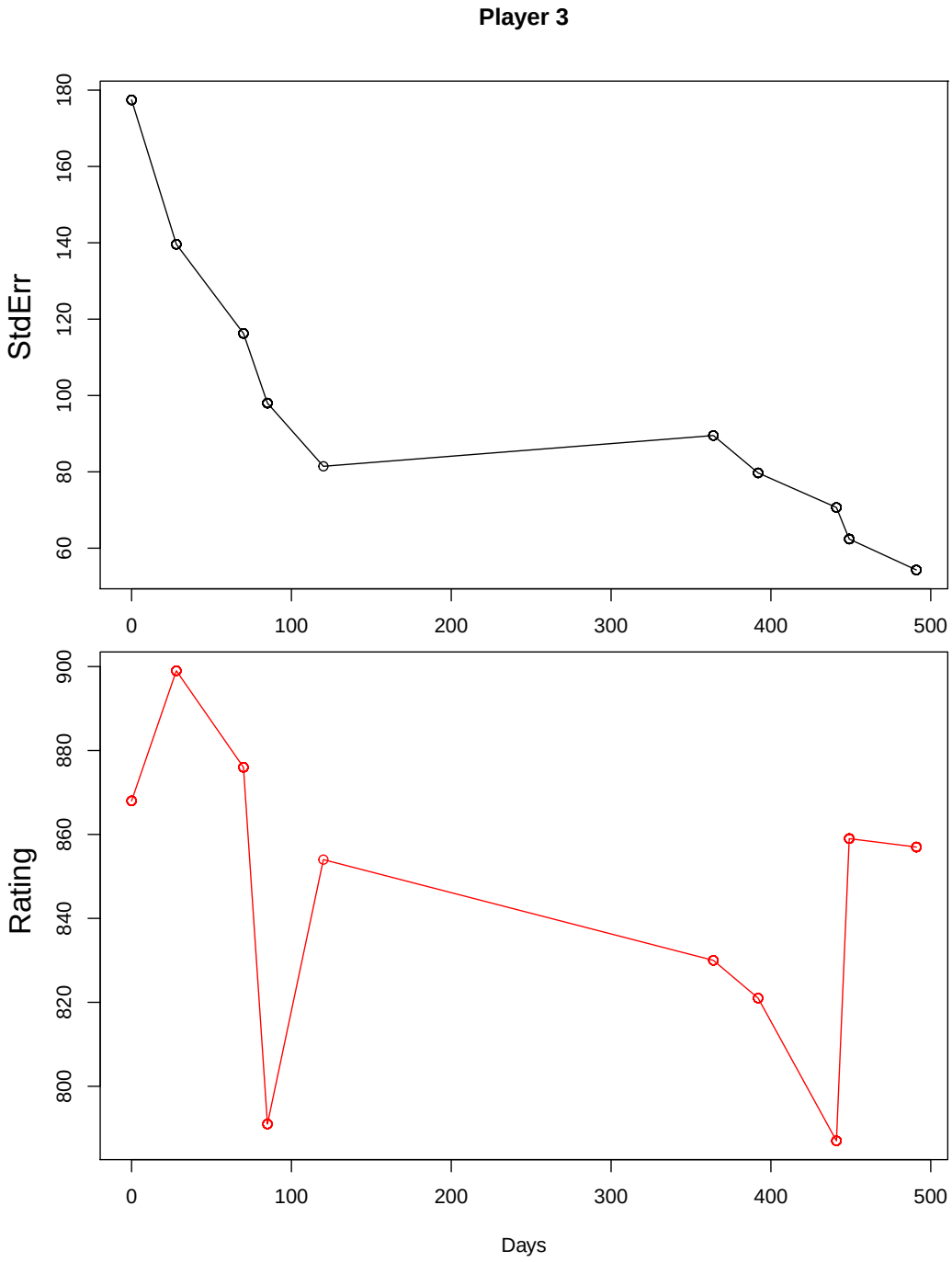

Figure 5: Standard error calculations for a randomly selected player for events rated between May 2014 and May 2017. Top panel is the standard error computation over time, and the bottom panel is the player's computed rating over time.