The Use of the K-factor in Estimating Individual Ability

Advanced Study in Individual Differences

Paul Lodder

University of Amsterdam, Amsterdam, The Netherlands

In ability estimation, the k-factor weighs the importance of the current performance data, relative to past performance data. In this paper I review different techniques that are used to determine the value of the k-factor. I also come up with a new way of determining the k-factor, based on a method called the Kalman Filter.

Imagine Dr. Garry, a scientist who wants to know the chess ability of individual chess players. To do this, dr. Garry invents a scale on which all chess players will be rated. He gives each player a position on this scale based on the number of matches they won in the past. Subsequently, he decides that a change in rating should depend on whether a chess player performs above or below expected during a match. Tomorrow there will be a match between a grandmaster and an amateur chess player and dr. Garry expects based on the ratings of both players that the grandmaster will win. If the amateur wins, then this player performs better than expected and the grandmaster performs worse than expected. In such a scenario, dr. Garry wants to increase the rating of the amateur and decrease the rating of the grandmaster.

However, dr. Garry is thinking about the degree to which he should adjust the rating of both players. If he weighs the information about current and past performance equally, then the Grandmaster's rating would decrease severely, but if dr. Garry gives past performance more weight than current performance, then the Grandmaster's rating will only decrease a little. The weight that dr. Garry should give to current performance, relative to earlier performance, is called: the k-factor (Elo, 1978). The goal of this paper is to find out whether there is an ideal k-factor in estimating individual ability, or whether the value of the k-factor should depend on other factors.

First, I will give an overview of the different ways in which the k-factor is used to estimate individual ability and indicate the weaknesses of these different approaches. After that I will argue that there is not one ideal k-factor and I will propose a method to determine the k-factor for each individual.

I would like to thank Han van der Maas and the students of the course *Advanced Study in Individual Differences* for giving me useful feedback on the ideas presented in this paper.

► E-mail: **p.lodder@uva.nl**

K-factor

The k-factor weighs the current performance relative to the earlier performance in estimating individual ability. In updating an individual rating, the k-factor is multiplied by the difference between the expected and the actual score. A low k-factor implies a conservative updating process because a deviation from the expected score does not influence the rating that much. On the other hand, a high k-factor implies a swift and dynamic updating process because each deviation from the expected score can cause a completely different individual rating. Figure 1 illustrates the difference between a high and low k-factor in estimating the individual ability over time. A drawback of using a high k-factor can be that ability estimate will be too sensitive to recent outcomes. A disadvantage of using a low k-factor can be that the system will respond slowly to actual changes in a player's ability. Note that the k-factor does not determine the ability level of a chess player; it only determines the speed with which the estimated ability moves towards a player's ability level.

Elo's Rating System

Physics professor Arpad Elo (1978) invented a new method for estimating the individual chess ability. The novelty in his Elo Rating System (ERS) consisted of a statistical technique to relate game results to a latent factor that represents the individual chess ability. Below are two important ERS equations.

$$E_A = \frac{1}{1 + 10^(R_B - R_A)/400} \tag{1}$$

$$R'_A = R_A + K(S_A - E_A) \tag{2}$$

With equation 1, Elo calculated the expected score of player A (E_A) by using a model that is closely related to Rasch's IRT model. The expected score in a match between player A and B is a function of the rating of both players (R_A&R_B). After a match, Elo used equation 2 to update the rating of a player. The updated rating (R'_A) is based on the old rating (R_A) plus a multiplication of the k-factor (K) with the difference between the expected

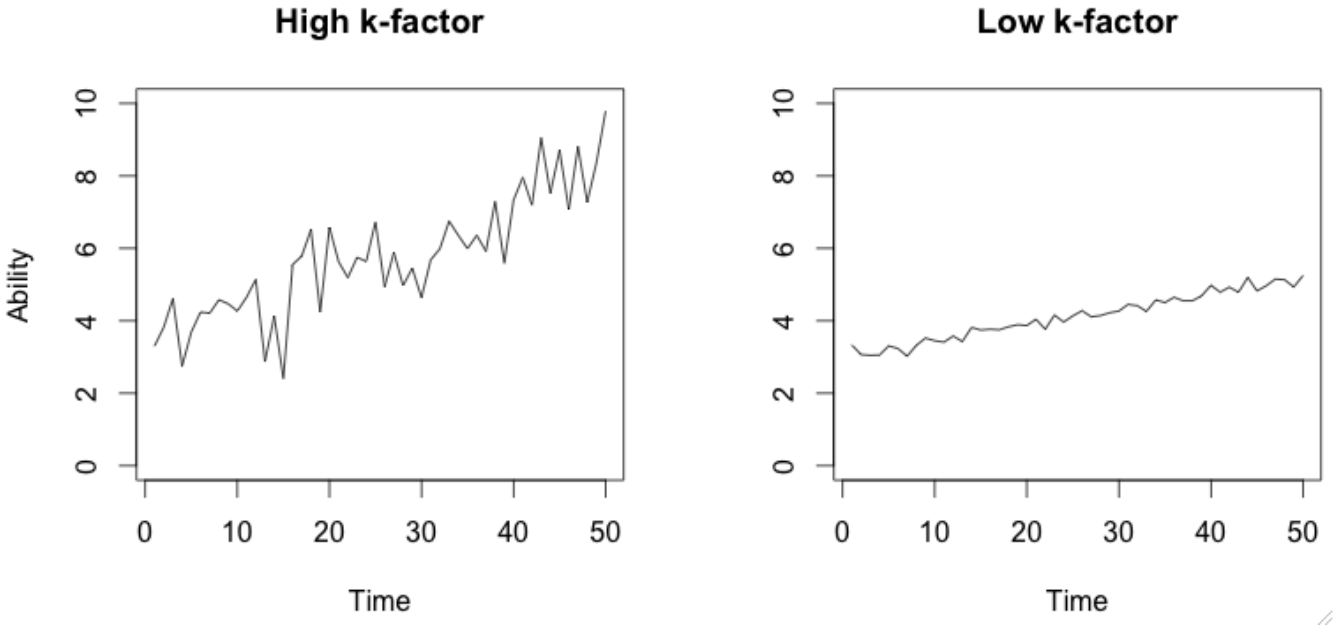

Figure 1. Ability estimate over time when using a high or low k-factor.

and the actual score ($S_A - E_A$). Elo decided to let the k-factor vary over three different rating groups. Players with a low rating get a larger k-factor than players with a medium or a high rating. The rationale behind this choice is that new players who start with a low rating will likely show more changes in their rating than experienced players who already have a high rating. As we will see in the next paragraph, there is one major disadvantage to this approach.

Sonas' Analysis

Jeff Sonas is a statistical analyst who studies the ratings of chess players. He identified a major drawback of Elo's choice to give highly rated players a lower k-factor. Sonas (2012) argued that it could lead to an overestimation of highly rated players with decreasing ability levels. For example, Karpov was a grandmaster who, after losing a championship, decreased in strength during the last few years of his career. However, due to the low k-factor he remained among the top ten players long after his playing strength decreased to that of a top 50 player (Thompson, 2012). According to Sonas (2002), a higher k-factor for highly rated players will prevent overestimation because the estimating system will respond more quickly to decreasing chess ability.

To find the ideal k-factor, Sonas (2002) performed a simulation study in which he retroactively calculated how accurate chess ratings were in predicting future results. He used rating calculations of 266.000 game results between 1994 and 2001 and applied different k-factors to the rating formulas. Figure 2 shows the results of his analysis. It turns out that the rating system most accurately predicts future performance if the k-factor is set to 24. The k-factor originally used in the ERS for highly rated players was equal to 10, which is suboptimal according to Sonas. He concludes that k-factors smaller than 24 lead to inaccurate ability estimations because the system is too slow to respond to actual changes in ability. There are some problems, however, with Sonas' decision to use one k-factor of 24. The next paragraph will describe why the use of a single k-factor can be problematic.

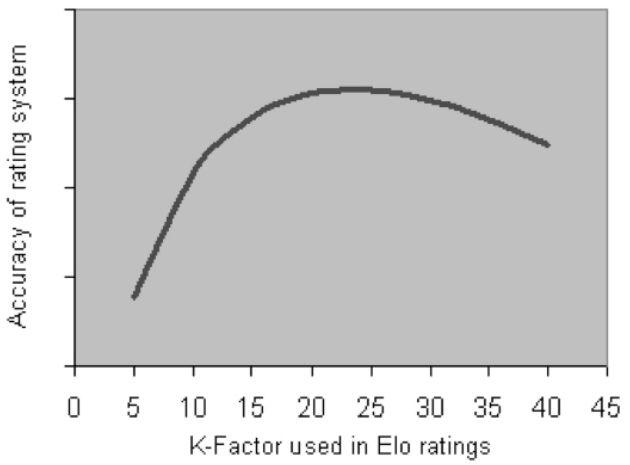

Figure 2. Accuracy of the ERS for different values of the k-factor.

Glickman's K-factor

According to Glickmann (1995), the k-factor should not be one value for all chess players, but should differ between players. We should give players a low k-factor if we are confident in their rating and a high k-factor if we are unsure about the accuracy of their rating. Glickman identifies two scenarios in which we should be unsure of a player's rating accuracy. First, if a new player did only compete in a few tournament games, then we can have no confidence in the accuracy of the rating because it is based on little data. Secondly, if a player did not compete for a very long time then we can also have no confidence in the accuracy of the rating because the current rating might differ from the rating based on his past performance.

In order to overcome these problems, Glickman (1995) suggests that each player should have a unique k-factor that indicates the confidence we have in that player's rating. This confidence level will depend on playing frequency and recency. Thus, the k-factor will be a function of the time since the last game played, and of the total number of tournament games played.

Klinkenberg, Straatemeier, and Van der Maas (2011) used Glickman's k-factor in estimating the individual math ability of children. Their data consisted of children's answers on math questions in an online learning environment called the Math Garden. In the math garden – as opposed to chess matches – players do not play against other players, but they play against items. The math garden is an adaptive environment, which means that each child receives items adjusted to the child's ability level.

Equations 3 and 4 show how Klinkenberg et al. calculated the k-factor of each player and of each item. Following Glickman (1995), they let the k-factor be a function of the rating uncertainty (U) of a player (j) or item (i). Equation 5 shows that the uncertainty depends on the recency and frequency of play: the updated uncertainty ($\hat{U}$) decreases after every item administered and increases after each day of not playing (D).

$$K_j = K(1 + K_+U_j - K_-U_i) \tag{3}$$

$$K_i = K(1 + K_+U_i - K_-U_j) \tag{4}$$

$$\hat{U} = U - \frac{1}{40} + \frac{1}{30}D \tag{5}$$

By using those three equations, Klinkenberg and his colleagues managed to give each player and item a unique k-factor. It turns out their approach leads to a reliable estimate of children's math ability (Klinkenberg et al., 2011). However, one drawback of the approach is that they assume that rating uncertainty will decrease if more items are played. This assumptions sounds reasonable, but is not verified and therefore the next paragraph will describe how this assumption can be improved upon.

Updating Uncertainty

Zult and his colleagues (2012) argue that the assumption of decreasing rating uncertainty after every administered item can be improved upon. It does sound reasonable that uncertainty decreases after gathering more data on the performance of a player, but it would be even better if the uncertainty could be based on an actual response pattern. After all, response patterns can indicate whether players are performing around their true ability levels. For instance, successive overperformance might indicate that the real ability of a player is at a higher level than the level at which the player is currently playing. Successive underperformance might indicate that the real ability is at the lower level than the level at which the player is currently playing. In both scenarios it would be beneficial to have a higher uncertainty and a higher k-factor to let the ability estimation converge faster to the true ability level. Conversely, alternating over- and underperformance might indicate that the person is playing approximately at the true ability level. Therefore, both the uncertainty as well as the k-factor should decrease to let the ability estimation stabilize around the present estimate.

Zult et al. succeeded in creating a model in which the rating uncertainty depended on a player's response pattern. It turned out the correlation between their method and the Elo ratings did not differ from the correlation between the old method and the Elo rating, which implies that both methods describe the data equally well. Unfortunately I cannot evaluate the bias and measurement precision of this method, because the Zult et al. (unpublished) article is still under construction.

Interim Summary

Up to this point, I explained different ways of using the k-factor in the estimation of individual ability. Elo (1978) first introduced the k-factor and Sonas (2002) discovered its optimal value after analyzing chess data. Subsequently, Glickman (1995) argued that there should not be a single k-factor, but that each player should have its own k-factor, depending on the confidence we have in the rating of a player. Klinkenberg et al. (2011) used Glickman's method to estimate math ability of children, basing the k-factor on the frequency and recency of play. Finally, Zult et al. (2012) based the value of a player's k-factor on the response pattern.

The summary above indicates that in the beginning of its existence, the k-factor was equal for almost everyone. Subsequently it started to depend on individual data such as the frequency and recency of play and at the moment it becomes even more tailored to the individual by letting it depend on a player's response pattern. This move from one universal k-factor to an individual k-factor seems to be in the right direction if we think about what a k-factor ideally should do. The size of the k-factor determines the speed with which the

estimated ability converges to the true ability. The k-factor should be high if the estimate is far from the true ability and the k-factor should be low if the estimate is near the true ability. This indicates that a single value is not appropriate and that it should depend on individual data.

Another problem with determining the value of the k-factor is that the data is always developing. People do not stay at the same skill level, but they develop and get better (or worse) at a particular skill. This makes it difficult to determine the value of an individual k-factor, because what does high current performance mean? Is a person really getting more skillful? Or is the higher performance due to random factors or systematic factors external to the person? Developing data implies that the optimal k-factor is also continuously changing. It is not like fitting a regression model and identifying the optimal parameters. The optimal parameters depend on the data and the data changes with every item or match played. If the data would not show any variation, then it was not necessary to keep adjusting the optimal parameters. Therefore, a possible direction we could take in determining the value of the k-factor is to let it depend on the variation an individual player shows in its performance. In the next paragraph I will describe a way to do this.

A New Direction

Recently, the world chess federation (FIDE) organized a contest on finding the most accurate system to predict chess outcomes (Kaggle, 2012). Tim Salimans, a Dutch PhD student at the Econometric institute, won the contest. In an article he wrote about the system he developed, he talks about a method to estimate an unknown variable based on measurements observed over time (Salimans, 2012). This method is called the Kalman Filter (Welch & Bishop, 1995) and it is already heavily used to analyze time-series data in economics (Ritter, 2012).

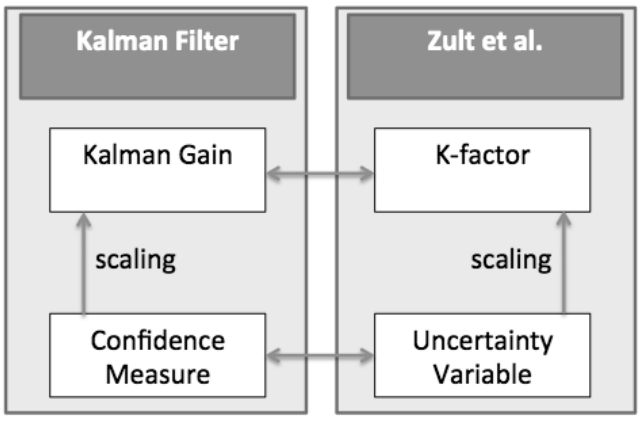

Figure 3. Similarities between the Kalman Filter and the k-factor calculation of Zult et al.

The Kalman filter calculates model parameters, including a measure of the confidence the model has in its parameters, which can be seen as an a priori uncertainty. This makes the confidence measure rather similar to the uncertainty variable that determines the k-factor in the Klinkenberg et al. (2011) and the Zult et al. (unpublished) articles. The Kalman filter also uses a factor called the Kalman gain, which is similar to the k-factor because it determines the weight of the current data relative to the past data. The Kalman gain is scaled by the Kalman confidence measure, just like the k-factor is scaled by the uncertainty variable. Figure 3 illustrates the similarities between the Kalman filter and the method used by Zult and his colleagues.

The Kalman confidence measure is simply a covariance matrix with measurement residuals of the parameters. Therefore, with the Kalman filter it is the measurement residual that determines the optimal Kalman gain. We could translate this method to the estimation of individual ability in the math garden by making the k-factor a function of the measurement residuals of the predicted individual ability. If the predicted ability has high measurement residuals then this implies that there is uncertainty concerning the estimation of the ability level. This uncertainty should lead to a higher k-factor, whereas a low measurement error and thus low uncertainty should lead to a lower k-factor.

It could be argued that the uncertainty variable in the Zult et al. (2012) article is based on the response pattern, a pattern that indicates whether a player successively over- or underperforms, or alternates between the two. It could be argued that this uncertainty variable is actually already determined by the measurement residuals. If a player continuously overperforms, then such a response pattern might also be visualized by higher measurement residuals of the ability estimate, because the ability is not estimated accurately. This suggests that the method used by Zult et al. to calculate the uncertainty variable is similar to the calculation of the confidence measure in the Kalman filter. However, there are mathematical differences between the two methods. The Kalman filter uses the exact values of the measurement residuals, while equations 6 and 7 show that Zult et al. do not take into account the size of the residual. They take the sign of a residual, which implies that all negative residuals are transformed to -1 and all positive residuals to +1.

$$\hat{U}_j = (1-\alpha)U_j + sign(S_{ij} - E(S_{ij}))\alpha \quad (6)$$

$$\hat{U}_i = (1-\alpha)U_i + sign(E(S_{ij}) - S_{ij})\alpha \quad (7)$$

It remains unclear which of the two methods most adequately determines the value of the k-factor. The Kalman filter is already successfully used in disciplines such as economics, finance and engineering (Ritter, 2012), but not yet in Psychology. Future research will

have to indicate whether we can use the Kalman filter to estimate individual ability and whether we should base the k-factor calculation on the exact value or the sign of the measurement residuals. Because without any empirical evidence, the match between those two different methods will end in a draw.

References

Elo, A. (1978). *The rating of chessplayers, past and present.* New York: Arco Publishers.

Glickmann, M. (1995). A comprehensive guide to chess ratings. *American Chess Journal*, *3*, 59–102.

Kaggle. (2012, October). *Deloitte/fide chess rating challenge.* (`http://www.kaggle.com/c/ChessRatings2`)

Klinkenberg, S., Straatemeier, M., & Van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education*, *57*, 1813–1824.

Ritter, S. (2012, October). *Using a kalman filter to predict ticket prices.* (`http://seatgeek.com/blog/dev/using-a-kalman-filter-to-predict-ticket-prices`)

Salimans, T. (2012, October). *How i won the deloitte/fide chess rating challenge.* (`http://people.few.eur.nl/salimans/chess.html`)

Sonas, J. (2012, October). *The sonas rating formula: Better than elo? Chessbase news.* (`www.chessbase.com/newsdetail.asp?newsid=562`)

Thompson, K. (2012, October). *Leave the k-factor alone!* (`www.chessbase.com/newsdetail.asp?newsid=5410`)

Welch, G., & Bishop, G. (1995). An introduction to the kalman filter. *Department of Computer Science, University of North Carolina at Chapel Hill*, 1–16.

Zult, D., Van Harreveld, F., Klinkenberg, S., Wagenmakers, E.-J., & Van der Maas, H. L. J. (2012, October). *A dynamic paired comparison based computer adaptive testing method.* (In preparation)