HARVARD UNIVERSITY
DEPARTMENT OF STATISTICS

Science Center 4th Fl.
1 Oxford Street
Cambridge, MA 02138

July 7, 2021

To Whom It May Concern:

I have been asked by Austin Lockwood, the ICCF Services Director, to comment on a proposal by an ICCF member to modify the existing rating system. I am Senior Lecturer on Statistics in the Department of Statistics at Harvard University. My position involves teaching, research, advising undergraduate and graduate students, and performing administrative duties within the university. I am also Senior Statistician at the Center for Healthcare Organization and Implementation Research, a Veterans Administration Center of Innovation. I received my B.A. in Statistics in 1986 from Princeton University (Summa Cum Laude), and my Ph.D. in Statistics from Harvard University in 1993. I have substantial experience in authorship, refereeing peer-reviewed papers, editorship, and leadership within the sports analytics community at both the local and international level. I have been a member of the US Chess ratings committee continuously since 1985, having served as chair of the committee from 1992 to 2019. I have invented the Glicko and Glicko-2 rating systems, both of which are used in rating players in organized chess (e.g., chess.com and lichess.org) and for rating players in various online gaming systems involving head-to-head competition. I am also co-inventor of the Universal Rating System which has been adopted for rating players in the Grand Chess Tour.

I have been provided a copy of the proposal in a document I received on July 3, 2021, which corrected an error in the proposed system in which the winning expectancies between two players in a game did not sum to 1. I also have a copy of the ICCF rating system as part of the ICCF Rules document (effective Jan 1, 2021). Austin Lockwood, on my request, has performed some computations based on implementations of both the current ICCF system and the proposed system to game results from 2011 onward.

Based on the proposal, the rationale for the changes is that the current ICCF system does not correctly estimate winning expectancies under the existing formulas, and that the discrepancy is particularly egregious for higher-rated players whose strength tends to be underestimated with the existing formulas. The proposal asserts, as a consequence, that the underestimation of strong players' ratings causes rating deflation.

The most crucial change to the rating formulas is that the proposed winning expectancy "shrinks" closer to 0.5 when players' ratings are higher. As I understand, the frequency of drawn games has enormously increased over recent years particularly among higher-rated

players, so that this feature of the proposed changes seems sensible to me. A second proposed change is that the development coefficient, *k*, is increased relative to the current ICCF system. This change would make sense if there was evidence that players' abilities at the upper end of the rating scale are more variable than the current rating system reflected. I do not see evidence in the proposal for this particular concern, nor have I been informed otherwise that players' abilities at the upper end of the rating spectrum change more quickly than the rating system is currently tracking. I appreciate that an alternative use of increasing *k* is to reflect the large frequency of drawn games, but my feeling is that the variation in scores by player needs to be better understood before relegating this issue to changing *k*.

More generally, I am struck by the patchwork approach the proposal takes to addressing concerns with the current ICCF system, and the choice of particular constants without optimizing them. For example, the proposed winning expectancy formula involves two layers of computations, neither of which is justified. The proposed formula equates the winning expectancy of a 2500-rated player against a 2400-rated player to be equal to the winning expectancy of a 2000-rated player against a 1950-rated player. This is just one example of a strong and unjustified assumption, and I do not see the evidence supporting this particular feature. There are various constants in the system that seem to be chosen by hoping that their behavior will lead to improvements, but this approach is not consistent with basic data science principles. It is difficult to know, for example, the long-term impact of inflating ratings at the upper end of the rating distribution.

My own experience with rating systems is that their main use is for predicting game outcomes (on an average basis). To evaluate whether the proposed rating system improved in predictability relative to the current ICCF system, I asked Austin Lockwood to carry out the following computations: (1) Run the ICCF and proposed systems in parallel using tournament results from 2011 onward, treating the 2011-2015 period as a "burn-in" period for the proposed system. (2) From 2016 through now, record for each rated game between players with fixed ratings the game result and the players' ratings at the start of the rating period. I was sent a spreadsheet with these results. For each game, I computed a Log Loss and a Brier Score. Letting W be the outcome (1, 0.5, 0) for white, and We the winning expectancy (relative to each system's formula) for white, the Log Loss is computed as $(-W*\log_{10}(We) - (1-W)*\log_{10}(1-We))$, and the Brier score is computed as $(W - We)^2$. Both are measures of lack of predictability, and are commonly used in data science applications. The larger the value of each measure, the worse the predictability.

The following table summarizes the average Log Loss between the ICCF and proposed systems starting in year 2016.

	ICCF	Proposed
All Games	0.2765578	0.2770238
Games with at least one player with ICCF rating > 2450	0.2950945	0.2942219
Games with at least one player with ICCF rating > 2200 and <= 2450	0.2821632	0.2824887
Games with at least one player with ICCF rating > 2000 and <= 2200	0.2690178	0.2688965
Games with at least one player with ICCF rating > 1800 and <= 2000	0.2460499	0.2471808
Games with at least one player with ICCF rating > 1600 and <= 1800	0.2363376	0.2394879

The overall predictability according to the average Log Loss (averaged over all games) is slightly better for the ICCF system compared to the proposed system, but barely so. By

focusing the on games played with players in a particular rating range, it can be seen that the proposed system performs insignificantly better for games involving at least one very strong player. However, for lower rating ranges, particularly for games involving a player with a rating between 1600 and 1800, the proposed system has a more pronounced Log Loss. It would appear that an unintended consequence of the proposed approach is slightly worse predictability of lower rating ranges. The magnitude of the difference, however, is small; a difference in Log Losses of 0.003 corresponds to a 0.7% improvement in the probability of correctly predicting won games.

A similar conclusion is reached using average Brier scores over the same rating ranges as the summary, as shown in the table below.

	ICCF	Proposed
All Games	0.07417185	0.07425763
Games with at least one player with ICCF rating > 2450	0.03376504	0.03265299
Games with at least one player with ICCF rating > 2200 and <= 2450	0.05688327	0.05677355
Games with at least one player with ICCF rating > 2000 and <= 2200	0.08939423	0.08899681
Games with at least one player with ICCF rating > 1800 and <= 2000	0.11694943	0.11780104
Games with at least one player with ICCF rating > 1600 and <= 1800	0.13893654	0.14127949

Again, the predictability is slightly better under the proposed changes for games involving high-rated players, but is worse than the current system for games with lower-rated players.

My impression is that the proposed system could be revised by optimizing some of the system parameters to improve the out-of-sample predictability relative to the current ICCF system. One might hope that such an exercise would result in improved predictability in all rating ranges with proper system constants rather than the ones asserted in the proposal.

I also examined the comparison of the distribution of resulting ratings from applying the proposed modifications to the current ICCF system which appears at https://iccfratingslab.z35.web.core.windows.net/ . Under the current ICCF system, the mode of the rating distribution is between 2300 and 2399, but with the proposed changes the mode moves to the 2400 to 2499 range. The distribution of ratings under 1900 remains roughly the same, though the changes have slightly worsened the predictability of ratings for this group of players, on average. It is an open question whether the rating distribution will continue to inflate, and whether this level of inflation is appropriate. This issue is not addressed in the proposal.

I hope my comments are helpful.

Sincerely,

Mark E. Glickman, Senior Lecturer on Statistics
Senior Statistician, Center for Healthcare Organization and Implementation Research