Essential elements of a rating system for correspondence chess: A Delphi survey of players

ICCF Ratings Implementation Workgroup (Uwe Staroske, Dennis Doren, Phillip Beckett, Manfred Frerichs, Eric Ruch, Gerhard Binder, Michael Millstone, Kristo Miettinen and Austin Lockwood)

June 2022

Abstract

Background: ICCF Congress voted in 2021 to develop a new rating system for correspondence chess. The proposed system was to be informed by a survey of correspondence chess players on what they consider to be the most important elements of the ICCF ratings system.

Methods: A 'Delphi' exercise was carried out to identify how far there was consensus on the essential elements of a correspondence chess ratings system in a group of 882 correspondence chess players. An initial free text survey of players produced 1,154 statements, 121 of these were considered by the authors to be out of scope and the remainder were categorised by similarity and reduced to 95 descriptive elements.

Results: Overall there was good player agreement on the importance of 52 (54.74%) of the 95 elements. Of the elements on which there was consensus, 9 (9.47%) were rated essential, meaning that *"without this element, the rating system would be unable to function"* or desirable, meaning that *"without this element, the rating system would be less effective, but still functional"*.

Conclusions: The degree of player consensus over the essential elements of a correspondence chess ratings system suggests that any new system commissioned by ICCF should take these elements into account.

Background

The chess rating system originally used by ICCF was based on the probabilistic Elo system used by FIDE and first introduced in 1991; this system has been successfully used continually by ICCF with minor adjustments for more than 30 years.

The emergence in recent years of powerful chess engines has meant that the assumptions of winning probabilities of the original Elo system are no longer valid, so in 2021 delegates to the ICCF Congress voted to consult a leading statistician for guidance in developing a new rating system suitable for the modern game.

One of the criteria of the new system was that its features should reflect what correspondence chess players find to be important about a rating system.

The aim of this study was to determine which elements of a chess rating system are the most important to players, so that an effort could be made to incorporate these into the new system.

The Delphi method is an incremental and structured technique for eliciting and converging the opinions of experts (in this case, correspondence chess players), where the results of each stage in the process are fed back anonymously to the panel to facilitate convergence.

Methods

Participants in the Delphi exercise were self-selecting correspondence chess players who responded to an announcement on the front page of iccf.com and an email circular to all players who opted into the ICCF mailing system.

The Delphi exercise took place in three stages:

For Stage One, participants were asked to identify three essential elements of a rating system for correspondence chess; they were shown three free text boxes, it was possible to leave one or more of the boxes empty, so not all participants left three entries. Participants were forced to log onto the system using their ICCF ID and password, so it was possible to prevent multiple entries from the same participant.

The study team then compiled the responses by (i) eliminating elements that appeared to be out of scope, and (ii) combining elements that appeared to be duplicated.

For Stage Two, an online questionnaire was constructed from the compiled list of elements from Stage One. Participants were invited to rate each element on a five-point Likert scale; the anchor points of the scale were:

1. Without this element, the rating system would not be able to function
2. Without this element, the rating system would be less effective, but still functional
3. This element is desirable, but the rating system could function without it
4. The absence of this element would have little effect on the functioning of the rating system
5. The presence of this element would have a detrimental effect on the rating system

For Stage Three, elements where there was already good agreement among the panel (determined by a semi-interquartile range of 0.5 or less) were excluded and a new questionnaire constructed from the remaining elements. The second questionnaire was similar to the first, but with the addition of two extra pieces of information for each element. The first piece of information was that the participant's previous rating was indicated by underlining the relevant anchor point. The second piece of information was the level of agreement within the panel as a whole was indicated by shading each score within one point of the median rating for each item (for example, if the median rating for a particular element was 2, then the ratings 1, 2 and 3 would be shaded). Each participant was asked to reconsider their original ratings from stage 2 in light of this new information. If their new rating was outside the shaded area (indicating that they disagreed with the rest of the panel), they were asked to enter a comment on their reasons for making their rating.

Results

Stage One

448 players submitted 1,154 elements.

Of the 1,154 elements, 121 were eliminated as being out of scope: for example, unclear responses like *"I don't know"*, general comments, elements relating to entry fees, or other rules not directly related to the rating system.

The remaining 1,033 elements were categorised into 95 elements by the working group by grouping similar or duplicate elements.

The average rating of the 429 players submitting the included elements was 2,208 (standard deviation 281.38, minimum 1,104, maximum 2,626), there were 19 grand masters (GMs), 42 senior international masters (SIMs), 52 international masters (IMs), 90 correspondence chess masters/lady GMs (CCM/LGM) and 46 correspondence chess experts/lady IMs (CCE/LIM).

Stage Two

662 players completed the questionnaire at Stage Two. The average rating of these players was 2,218 (standard deviation 270.30, minimum 992, maximum 2,609), there were 22 GMs, 66 SIMs, 83 IMs, 144 CCM/LGMs, and 69 CCE/LIMs.

Of the 95 elements identified in Stage One, 30 were found to have a good agreement (semi-interquartile range of 0.5 or less), the median ratings for each of these elements were:

Without this element, the rating system would not be able to function

- The ICCF rating system should be open and transparent
- The ICCF rating system should be fair and be applied consistently

Without this element, the rating system would be less effective, but still functional

No elements

This element is desirable, but the rating system could function without it

- Rating inflation and deflation should be controlled so that there is some continuity between historical and current titles (CCE, CCM, IM, SIM, & GM)
- The rating system should be stable and not allow excessive fluctuations
- ICCF rating should be a qualification criterion for tournaments
- ICCF should continue to release a rating list four times a year
- Provisional ratings should be adjusted quickly to allow a player to reach his or her true strength
- There should be a balance between stability and responsiveness to changes in strength
- Provisional ratings should be based on FIDE or national ratings
- It should be easier for lower-rated players to enter title norm tournaments
- The system should be resilient to the effects of provisionally rated players with an unknown playing strength

The absence of this element would have little effect on the functioning of the rating system

- There should be some direct comparability between ICCF and FIDE ratings
- Provisional ratings should be based on national CC ratings
- Players should not be marked as inactive until at least ten years after their last game
- Games played against provisionally rated players should not be rated for the experienced player
- Ratings should reflect a player's willingness to value social and cultural exchange
- Preliminary rounds should not be rated because of the risk of mismatches; only later rounds should be rated
- There should be more variation, for example a rating of 3000 should be possible.

The presence of this element would have a detrimental effect on the rating system

- ICCF should remove the rating system and not replace it at all

- Shorter games should be scored higher than longer games
- Not playing over a certain period of time should lead to a lower rating
- The rating system should reflect good moves on the chessboard, not just results
- In the case of a draw, the higher rated player's rating should increase
- The rules of the game (laws of chess) should be changed so that players cannot rely on opening knowledge or chess engines
- A player's initial rating should be determined by games played against a computer engine
- In the case of a draw, both player's ratings should increase
- All ratings should be rounded down to the nearest 100 points
- Longer games should be rated higher than shorter games
- Defeats in longer games should not lose as many points as defeats in faster games
- A player's response time to moves should be taken into account for ratings

Stage Three

606 players completed the questionnaire at Stage Three. The average rating of these players was 2,208 (standard deviation 278.95, minimum 992, maximum 2,609), there were 19 GMs, 57 SIMs, 73 IMs, 120 CCM/LGMs, and 66 CCE/LIMs.

22 of the remaining 65 elements converged on good agreement after Stage Three, these were:

Without this element, the rating system would not be able to function

- Ratings should never decline following a win

Without this element, the rating system would be less effective, but still functional

- Ratings should accurately reflect a player's strength relative to other players
- ICCF ratings should reflect a player's progress over time
- The ICCF rating system should be built on sound statistical principals
- If a new system is introduced, there must be some continuity and comparison with the old system
- New rating lists should be produced on time
- Provisional ratings should be fair and realistic

This element is desirable, but the rating system could function without it

- The ICCF rating system should be tuned to recognise the high number of draws in CC
- ICCF should continue to use the Elo system, but modified in some way
- The rating system should be sensitive to changes in playing strength and allow rapid changes
- Lower rated players should be allowed to enter higher rated tournaments to give an opportunity to increase their rating
- If ICCF introduces a new system, players' historical ratings should be corrected retrospectively
- ICCF should not recognise national ratings; all member federations should use ICCF ratings for national events

The absence of this element would have little effect on the functioning of the rating system

- ICCF should adopt the Glicko rating system
- There should be a different rating list for each time control
- More points should be given for draws
- ICCF Should not use the Elo system
- The number of games before a player can have a published rating should be increased

- Provisionally rated players should not lose a lot of points
- Players who have lost a lot of games through resignations should receive a special rating category

The presence of this element would have a detrimental effect on the rating system

- Very short games (for example shorter than 16 moves) should not be rated
- A draw against a much stronger player (for example +400) should be rated as a win

Elements with only moderate agreement

After Stage Three, the remaining 43 elements all achieved moderate agreement (defined by a semi-interquartile range of between 0.5 and 1), these were:

Without this element, the rating system would not be able to function

No elements

Without this element, the rating system would be less effective, but still functional

- Wins should be rewarded higher than draws
- The ICCF rating system should be easy for players to understand
- There should be some protection against rating manipulation
- Players should be able to see ratings as a personal motivation for improvement
- Defaults and ETLs (flag falls) should be rated the same as lost games
- All ICCF and national federation events (except Chess960) should be rated
- Games completed by substitutes in team events should not be rated for the substitute if he or she didn't make any moves
- ICCF should recognise that correspondence chess is not over the board chess, and there should be no expectation of a link between ICCF and over the board or FIDE ratings

This element is desirable, but the rating system could function without it

- Ratings should be updated after every game
- Rating should be updated more frequently than currently
- The difference in players' ratings should predict the outcome of a game played between those two players
- A player's rating when a game is ended should be used to calculate the player's new rating
- The rules for chess titles (CCE, CCM, IM, SIM, & GM) should be based on the rating formula
- It should not be possible to reach a high rating in a small number of games, there should be limits on early ratings
- Additional rating points should be awarded for winning a tournament
- ICCF should discontinue the rule where more than 80 games in a period are rated using a different formula
- A player's rating at the start of the tournament should be used to calculate the new rating
- Any new rating system should be responsive to changing patterns of game outcomes (for example an increased draw rate)
- All new players should start with the same provisional rating
- Defaults and ETLs (flag falls) should be penalised with a rating deduction
- Ratings should be updated at the end of each tournament
- There should be a different rating for regular games and thematic events (events which start from a specific position)
- The rating system should be sensitive to small differences in strength because of the influence of powerful computers

- Title norm requirements should be lowered to reflect the difficulty in achieving a higher score
- If ICCF use a new system for ratings, players should retain their old rating with no retrospective correction

The absence of this element would have little effect on the functioning of the rating system

- ICCF should keep the current rating system with no changes
- ICCF should adopt the FIDE rating system with no changes
- ICCF should not recognise FIDE ratings for new players
- A win with black should be scored higher than a win with white
- ICCF ratings should be approximately equal to over the board and online ratings
- Instead of being based on win expectancy, the ICCF rating system should be based on tournament points (as in tennis, snooker, etc.)
- The number of games played should contribute to a higher rating
- Draws should not be rated, or have minimal impact on rating
- There should be a separate rating for players who do not use a computer
- Defaults and ETLs (flag falls) should be rated differently to wins
- ICCF should adopt a "rolling" rating system, where only the results of recent games should be considered
- Games between players where there is a large difference of ratings should not be rated
- Players should receive a ratings boost when they earn a title (CCE, CCM, IM, SIM, & GM)
- A player's rating should change more slowly the longer they have been playing
- ICCF ratings should somehow reflect a player's over the board strength
- Titles (CCE, CCM, IM, SIM, & GM) should be taken into account when calculating rating
- Higher rated players should not be penalised for drawing against lower rated players
- Players with higher ratings should be protected when playing lower rated players

The presence of this element would have a detrimental effect on the rating system

No elements

There were no elements with poor agreement after Stage Three (defined by a semi-interquartile range of greater than 1)

Conclusions

The Delphi approach appears to have been successful in converging expert opinion in an area where there is naturally a wide range of views and opinions. Of the 95 elements identified in Stage One, good agreement was achieved for 30 items at Stage Two, and for a further 22 elements at Stage Three. Fewer than half the elements (43) eventually achieved less than good agreement, and none eventually had poor agreement. These results appear to vindicate the methodology used.

The next stage of the ICCF Ratings Review project is to consult with a leading statistician; this survey will ensure that players views are fully taken into account when the new system is submitted to delegates for approval at the 2022 ICCF Congress in Glasgow.