# REPORT OF THE GOLF CROQUET RANKING REVIEW COMMITTEE

This committee was formed in November 2012. It currently consists of the following members:
Paul Billings, Chris Clarke, James Hopgood, Stephen Mulliner, Louis Nel.

It had the following terms of reference:
"To look at the technical aspects of the formulae and factors used, given the differences in playing characteristics between AC and GC:

- to consider the current CGS algorithm, particularly whether characteristics such as lag are appropriate, and summarise the strengths and weaknesses.
- to consider how the system should deal with rapid improvers and sporadic players
- to consider whether the DG algorithm would be appropriate for GC, and if so, whether an improvement over CGS.
- to recommend the most appropriate algorithm for GC World Rankings
- to decide whether class factors should continue to be included
- to consider how new players entering the system are introduced (starting grade etc.)
- to clarify which games should be included in the rankings
- to look at how to make it easier for countries to enter their own data directly."

We are herewith submitting our report. It comes after extensive fact-finding, which is reported in the attached companion document

[GCGP] GOLF CROQUET GRADING POSSIBILITIES.

Much of this fact-finding is concerned with long term trends, which was not within reach during the first few years because we did not have enough game results available.

## 1. MEASURING GRADING ACCURACY

Our task calls for comparison of various grading systems – in particular comparison as regards grading accuracy. For purposes of this comparison we organized the available set of game results on 14 March 2017 into 53 successive batches of 3000 games. On each batch one can apply the long known Chi Squared statistic. It gives a measure of how accurately the grading system estimates win probabilities. We translated the Chi Squared statistic into an essentially equivalent (numerically more convenient) one: the **Grade Deviation** statistic (GD). Since the GD over any given batch often deviates considerably from the GD over other batches, we introduced the concept

**rGD = recent Grade Deviation = average GD over the most recent 30 batches**

Detailed definitions appear in [GCGP]. We noticed a deterioration of rGD as the number of batches increases and this led us to introduce also the statistic **GAT = Grading Accuracy Trend,** given by

**GAT = (average GD over first 30 batches) – (average GD over last 30 batches).**

These two performance statistics enable us to make useful objective comparisons.

## 2. START GRADE REVISION

After long contemplation of the mentioned deterioration we diagnosed it as being mainly caused by start grades, aggravated by the fact that the majority of players in the GC database have fewer than 30 recorded game results. Having made this diagnosis, we put a lot of effort into the development of a method for detecting poor start grades and for revising the detected ones. It turned out that about one in every five start grades needed to be revised, some of them by several hundred grade points. We developed a method for automatic revision. We were eventually able to transform every grading system under consideration into one that automatically performs this start grade revision.

## 3. A QUICK GLIMPSE OF OUR FACT-FINDING

Every post-game updating algorithm considered is an elaboration of the Fixed Modulator algorithm that we now recall for convenient reference. It expresses the New Grade (= grade after the game) in terms of the Old Grade (= grade at start of the game), in terms of a single parameter (Modulator) as follows

Winner's New Grade = Winner's Old Grade + Modulator * (Loser's Win Probability)
Loser's New Grade = Loser's Old Grade – Modulator * (Loser's Win Probability),

where the Loser's Win Probability is estimated by the system. This algorithm becomes a grading system, when a specific positive value is assigned to the parameter. For example we may assign Modulator = 20 or Modulator = 0.2 or Modulator = 2000. Each such assignment gives a different grading system. When the system estimates the Loser's Win Probability to be 0.4 (= 40%), we note that

Modulator = 20 will cause the winner to gain 20 * 0.4 = 8 grade points
Modulator = 0.2 will cause the winner to gain 0.2 * 0.4 = 0.08 grade points
Modulator = 2000 will cause the winner to gain 2000 * 0.4 = 800 grade points

Clearly, 0.2 is far too small a modulator value: it will practically retain the start grade of every player and not allow the grade to keep up with the form of a typical improving player. On the other hand, 2000 is far too large a modulator value: a grade that happens to be correct at the start of the game will be woefully incorrect after the game. This sheds light on the role of the Modulator.

We are now going to present performance statistics for a certain four grading systems and indicate what can be learned from them. Let **FM** denote (temporarily) the Fixed Modulator system with Modulator = 18. Let **FMR** denote the closely related system with Modulator = 18.2 and equipped to revise start grades. Let **CSQ** (Continuous grading of the Status Quo) denote the system currently in use to do world rankings and let **CRE** denote the modernized version of CSQ obtained by changing the basic modulator assignment from 50 to 19.2, the Class 1 factor from 1.2 to 1.16, the Class 3 factor from 0.8 to 0.76, the Primary Smoothing Parameter (PSP) from 0.9 to 0.5 and setting the Secondary Smoothing Parameter = PSP (which effectively means it no longer varies with the Index). (Complete details are presented in

[GCGP], where several more special systems are reported on). The table to follow provides performance statistics for the mentioned four systems.

| System | rGD | GAT |
|---:|---:|---:|
| CSQ | 1.488 | -0.198 |
| FM | 1.233 | -0.166 |
| FMR | 0.992 | 0.030 |
| CRE | 0.989 | 0.020 |

It can be seen that the latter two systems, both equipped to revise start grades, perform significantly better than the former two. The above table suggests that grading accuracy has the following three prerequisites:

(i) Good update algorithm
(ii) Good parameter assignments
(iii) Good start grades.

Indeed, the latter two systems have all three prerequisites, the system FM has two out of the three (it lacks only good start grades), CSQ has only the first prerequisite: that its algorithm is good, is known from the performance of CRE (which has the same algorithm). CRE also shows up the poor parameter assignments of CSQ.

It should already be clear from the above that CSQ calls for replacement, but it is not clear yet what the most appropriate replacement should be. The grade-smoothing present in CSQ and in CRE has the widely-disliked feature that it creates situations in which a grade can increase even when a game is lost or it can decrease even when a game is won. This could occasionally discourage participation.

## 4. DYNAMIC GRADING SYSTEMS

Neither CRE nor FMR makes special provision for rapid improvers. This calls for dynamic grading. The underlying idea of dynamic grading is to detect a rapid improver (or rapid regressor) and to use an increased modulator when the player is in that state. This idea was pioneered a few years ago by the AC Ranking review committee. We discovered a more efficient detection method and also a more efficient way to apply the temporary modulator enlargement (see [GCGP] for details). This led to the creation of three dynamic grading systems: DR, DRE and DREA. Let us outline them, while referring to [GCGP] for details.

DR is nothing but the system FMR with appropriate adjustment of parameter values and equipped with dynamic grading; DRE is DR further equipped with traditional event classification leading to Class Factors (as in CSQ) but with more judicious choices for the Class Factor values. DREA is much like DRE except that the Class 1 events are algorithmically selected rather than at the discretion of the ranking officer.

We provide performance statistics of these three systems in the section to follow.

## 5. FINAL PERFORMANCE STATISTICS

Final performance statistics were calculated over more recent data (up to 30 September 2017) which included 59 batches of 3000 games. The notation rGD59 and GAT59 is used for performance statistics calculated over these 59 batches while the previous ones are now written rGD53 and GAT53.
The new rGD59 and GAT59 are of interest not only because they involve more data, but particularly because they measure performace in at least 6 batches of games that had no influence over the choice of parameter assignments.

The three systems that employ dynamic grading outperformed all other systems considered and yielded the following performance statistics.

| System | rGD53 | GAT53 | rGD59 | GAT59 | rGDdif |
|---|---|---|---|---|---|
| DRE | 0.967 | 0.031 | 1.000 | 0.005 | -0.033 |
| DREA | 0.939 | 0.043 | 1.011 | -0.004 | -0.072 |
| DR | 0.968 | 0.050 | 1.015 | 0.008 | -0.047 |
| CSQ | 1.488 | -0.195 | 1.555 | -0.278 | -0.067 |

It appears from these numbers that DRE is marginally ahead of the other two, in particular as regards rGD59, which is arguably the statistic of greatest interest. It is not surprising that the rGD59 and GAT59 are generally slightly worse than rGD53 and GAT53 because parameters that yield a good fit over the data set for which they were calculated can be expected to yield a less good fit over the expanded data set. DRE also had the smallest deterioration over the six new batches (indicated in the rGDdif column) -- only half that of CSQ.

How is the conspicuous difference between rGD(DRE) = 1.000 and rGD(CSQ) = 1.555 reflected on ranking lists produced by DRE and CSQ? It gives an average difference of 12.2 rank positions for top 100 players and an average difference of 22.9 rank positions for Top 500 players (see [GCGP] for more details about this calculation). This contrasts with the average Top 100 difference of less than 1.0 rank position on rank lists produced by any two of DRE, DR, DREA and less than 2.4 rank positions for Top 500 players.

## 6. RECOMMENDATIONS and SUGGESTIONS

**Recommendation 1.**
**That the WCF assume responsibility for Golf Croquet ranking, by designating a specific system as official and by appointing a Ranking Officer and a Deputy Ranking Officer to implement that system.**

The GC rankings has been a voluntary unpaid service provided to the CA since 2001 by Bill Arliss, for the first seven years, and then by Stephen Mulliner. These rankings have been used by the WCF. It is clear that these two gentlemen have rendered a great service to the sport for a long time and their industry is greatly appreciated. However, the committee feels that the WCF should now take on full responsibility for the GC world rankings.

We suggest that the execution of Recommendation 1 should go accompanied with an effort to reduce the workload of the Ranking Officer, which has become a lot for one person to handle. The calculations are not the issue, because the computer should produce the ranking list within seconds. The main burden arises from getting the input data properly prepared. If every member country could appoint somebody to oversee the submission of their own game results in the proper prescribed from, without ambiguity of names, that would go a long way toward a solution of the workload problem.

**Recommendation 2.**
**That the system DRE be designated as the official grading system for Golf Croquet.**

The three systems DRE, DREA and DR are so close in performance that they produce practically the same ranking list. DRE produced the best rGD59, arguably the most important performance statistic. It also shows the greatest resistance to deterioration. It is well equipped to handle rapid improvers properly and it does not have features that are widely disliked. It could be regarded as the most appropriate replacement of the system currently in use.

**Recommendation 3.**
**That a computer generated cumulative audit of the official grading system be published on the WCF website at least once every year, during the month of October. It should list the GD entries of the last 30 batches along with the rGD and GAT statistics, accumulated up to the last 5 years.**

Being automatically generated by the computer this should require very little effort. When these data are shown side by side for the preceding 5 years it will enable early detection of a serious shortcoming if there is one. Early detection could facilitate early remedy.

Continued good performance cannot be guaranteed. It is clear from the above tabulation of rGD59 and GAT59 that in just 6 months and with 6 new batches some deterioration of performance became visible in all systems. What will the performance look like after two more years? Or after 5 more years?

There are a number of influences over which nobody has control. The quality of new start grades is one of them. The automatic start grade revision that we introduced does not address all potential start grade problems. For example, if a local pool of players should develop who mainly play each other, then their grades may be accurate relative to each other while all being 200 points too high relative to the general population. Such a problem can go undetected for a few years and then suddenly come to light. It cannot be addressed by changing parameters; it requires a change in start grades additional to what our method is offering.

**Suggestions as regards games to be included**.

We noted that in earlier years the game scores were optional. We feel strongly that the more recent practice of requiring scores for all games should be continued.

We suggest that the WCF consult with Members with regard to the desirability of having minimum criteria for game length (points and time) and lawn size before games are included for ranking purposes

**Suggestion as regards published ranking lists.**

1. Requirement of either 20 games in the preceding 12 months for appearance on the list or 30 games in the preceding 24 months.
2. The following should be shown in addition to the current items:
   The career total of games played.
   On a DRE rank list, the Mobility Index and the Mobility Streak.

The career total is of interest as indicator of how much the grade may still be influenced by the start grade. It is also of interest in case of new players aspiring to play in a World Championship. Hopefully, there will be a regulation about that in the near future and a displayed career total could facilitate its implementation.

7. SOFTWARE RELATED ISSUES

The software used for all calculations presented by the committee is written in Delphi. This programming language, along with C++, is promoted by Embarcadero. Among programming languages, Delphi is among the most human readable. So there should be no problem for the software used by the committee to be translated into C++ or any other language that may be preferred by the WCF. The situation is quite similar to that of the AC Ranking Review committee a few years ago, when the source code for DG was translated from Delphi into C++.