The Use of World Rankings for the Association Croquet World Championships

A report by the WCF Ranking Review Committee

July 11, 2008

Contents

1 Introduction

In February 2008, the WCF Ranking Review Committee was asked by the WCF Management Committee to consider some issues relevant to the use of the World Rankings for the Association Croquet World Championships. The committee originally consisted of Chris Williams (the WCF Ranking Officer), Jonathan Kirby, David Maugham, and Louis Nel. For this report Chris Dent and Ian Burridge were also co-opted to the committee. Jonathan Kirby was appointed to chair the committee.

We were asked to consider:

- Do the rankings treat some countries unfairly?

- Using the rankings to select players for the Worlds, or how they might be adapted for that purpose.

- Using the rankings for selecting the top four seeds.

- The volatility of the rankings generally.

- What should be done about players who don't play many other strong players?

Section 2 of this report addresses the first point, and the appendix contains more details of the statistical analysis done. Section 3 is a general discussion about the selection process for the World Championships. There then follows a discussion in section 4 of different methods of using the world rankings to pick the top 50 or 60 players. A discussion of volatility in this context is included in this section. In a wider context, volatility cannot be separated from broader questions about the algorithm used to compute the rankings, and that is beyond the scope of this report (but see below). The question of seeding is considered in section 5. Some comments on the last point are included in the discussion of selection processes in sections 3 and 4.

It was not within the remit of this committee to consider the details of any allocation process. We have considered different approaches and have highlighted what we see as the important issues, but leave it to others to make the political decision of what method to adopt.

We have also been considering other possible ranking systems as replacements for the Croquet Grading System (CGS). This report does not discuss these possibilities, which we intend to report on at a future date. However some data from

the Bayesian Grade (BG), which is of particular interest to the committee in this regard, is included for comparison purposes.

1.1 Conclusions

1 We found no evidence that players of the same standard from different countries have different grades. The test using the drift of index points showed no significant drifts. The same results are reached if one only considers top players likely to play at international level (players above a given index).

2.1 Selection of the Top N players by the rankings: if "best players" is taken to mean with the best expected performance in the World Championship, the best way to use the rankings is to take the ordinary ranking list on the allocation date.

2.2 If "best players" is taken to mean those players with the best performance over the preceding 12-month period, the best method is by a weighted win/loss ranking.

2.3 A possible means of maintaining the general standard of play is to use a qualification standard. A player reaches the standard by attaining a CGS grade of 2100 at some point in the twelve months before the allocation date. Each country can nominate at most one player who does not meet the qualification standard.

3.1 The present system of seeding based on CGS grade after the block stage has no major flaws.

3.2 If the ranking system were changed to one with no lag, it would improve the seeding. In this case, it may be wise to use the rankings before the last day of block play (or even, for simplicity, before the event) to ensure there is no incentive for any player to lose a game. (However, changing the ranking system would have other implications which have not yet been considered, so we are unable to recommend changing the system at this time.)

3.3 Seeding by block results is inferior to the current method of seeding by rankings.

3.4 The alternative of using a seeding committee just for the top 4 or 8 seeds, with the rest done by rankings as above, could be considered if there is strong political support for it, but not otherwise.

3.5 One possible way of rewarding good play in the blocks would be to boost the seeding of block winners into the next group of eight seeds, although not into the top eight.

4 If a player has played few games against strong opponents, the rankings cannot be relied upon to give a good guide to how that player will play against strong opponents. A national selection committee can consider more evidence and make a better judgement. (In principle the same conclusion will hold for any player, but it is more pronounced in this case.)

5 In any match or tournament, the players should be told in advance whether or not the results will go into the rankings.

2 Comparison of countries

The world ranking system compares players from round the world in the same ranking system. The question has been raised of whether the rankings produced are fair to each country. For example, it could be the case that players from New Zealand are ranked higher than players of the same standard from England. How could this occur? The CGS works by transferring index points between players depending on the results of games. It is a *zero sum* system, which means that no points are created or lost – the winner gains the same number of points that the loser loses. At some point in time, a player may be playing to a standard above or below his CGS index. If he is playing above his index then his index will tend to increase to match his level, and he will take points from his opponents. The same occurs for groups of players. For example, if New Zealanders were ranked higher than English players of the same standard, then when they played each other, there should be a net transfer of points from the New Zealanders to the English. If they played enough games, this would tend to balance out any mis-ranking. However, the problem is that players from different countries may not play each other very much, and not enough to balance out the rankings. In this case, we would expect to observe a net drift of index points from the over-ranked country to the under-ranked country over the few games which are actually played.

In fact, countries gain and lose index points not just to each other, but also when new players enter the CGS and when other players leave. When a player first plays CGS games, the ranking officer, Chris Williams, estimates a starting index. This may be revised if, after a few games, it appears to have been wrong. After 30 games or so the starting index does not matter much for the player as the results of his games outweigh it. Chris made an investigation of the starting grades in 2006. He discovered that he had set a large number of new players' indices at 1700, particularly in Australia and New Zealand, and these had turned out to be too high. However, he had not subsequently adjusted them. The result was a large number of extra index point in those countries, which quickly became spread through the whole playing population. At the end of 2006, Chris adjusted a large number of these starting indices, and the rankings starting from January 2007 were changed to reflect the corrected starting indices. The result was that several Australian and New Zealand players lost a few index points, although not enough to make a substantial change to the ranking lists.

We have now addressed this issue more carefully and systematically to consider whether it is a significant problem. We considered all ranking games played since 2000. This is already a large number of games, and we felt that going back right to the start of the system in 1985 would not be measuring the system as it has settled down now that all countries have a large number of games in the system. We had

to group the countries into pools where most games are played inside the pools and fewer games are played between pools. It turns out that most countries do not form pools on their own. For example, Scottish players play as many games against English players as they do against Scottish players, so they must be considered part of the same pool. The table below is an example illustrating the number of games played between different groups of countries from 2000 until February 2008.

	Australia	GB&I	Minor	NA	NZ	SA
Australia	34002	1239	95	592	3029	50
GB&I	1239	43879	2275	2039	2773	1140
Minor	95	2275	1890	99	140	32
NA	592	2039	99	9542	410	82
NZ	3029	2773	140	410	13174	73
SA	50	1140	32	82	73	345

"GB&I" is Ireland, Guernsey, Isle Of Man, Jersey, England, Scotland, and Wales. "NA" is Canada and the USA. "Minor" is all other countries.

It turned out that we could only distinguish four pools: Australia, New Zealand, North America (consisting of the USA and Canada), and the Rest of the World, (including Great Britain, the rest of Europe, and South Africa). For example, from this table it is clear that the "minor countries" and South Africa play most of their games against GB&I, so they were incorporated in the same pool.

Between these pools the drift of index points changed direction from one year to the next, so there was no obvious evidence of mis-ranking. For example, although the USA did very well in the MacRobertson Shields in 2003 and 2006 and gained a lot of index points, they did relatively badly in other years, for example at World Championships, and lost index points. Chris Dent did a statistical test to see whether the overall drifts observed were suggestive of any mis-ranking. A full account of this test appears as an appendix to this report. Briefly, the idea is that the drift of points could happen either as a result of random fluctuations or as a result of systematic mis-ranking. We calculated the probability that the observed drift (or something more) would occur purely as a result of random fluctuations. In line with the standard practice for hypothesis testing, we would conclude there was systematic mis-ranking between the pools if this probability (known as a "*p*-value") were less than 5%.

The calculated value is 72%, which means there is no evidence of mis-ranking. However, in principle there could be mis-ranking of the top players which was balanced out by the other games. To check this, the test was repeated using only the games where both players had an index above a certain cut-off level. The results are tabulated below.

Minimum index	Number of games	p-value (%)
0	10621	72
2000	5611	72
2100	4362	23
2200	3263	43
2300	2189	21
2400	1160	14
2500	403	12

Although the p-values do get smaller as the minimum index increases, they are all well above 5%, so we conclude that the drift of index points is too small to suggest that there is mis-ranking. There is no evidence that the rankings treat any country unfairly, or even that it treats the top players of any country unfairly.

3   Selection for the World Championships

The WCF Association Croquet World Championship has, at present, 80 places which are supposed to be allocated mainly to the best players in the world. There are some other constraints: development places, host places, qualification tournament places, minimum numbers of places guaranteed for various countries, and wildcard places. After these constraints are taken into account, around 60 places will be available purely on merit.

The current allocation method is for each WCF member country to be awarded a quota of places, and for the countries to select their own players to make up that quota. The quota for each country is determined mainly by the world rankings, in particular the number of players the country has in the top 50.

The ranking review committee was asked to consider how else the world rankings could be used to select players for the world championship.

There are three possible basic methods of selection:

1. The rankings are used directly: the highest ranked available players are invited, irrespective of their countries.

2. A WCF selection committee or similar panel makes the selections (not purely based on world rankings).

3. The selection is delegated to member countries, with the number of places allocated to each country determined somehow from the rankings (or otherwise).

All three methods are likely to give places to most or all of the top 50 or so players in the rankings if they are available. The differences manifest themselves in the borderline cases, the last ten or so places (discounting the development places). So it only really matters which method is used for these borderline cases. In principle, the current method is number 3. However, any places which are not used by the countries are returned to the WCF as extra wildcards. At the 2008 world championships, the WCF allocated 16 wildcards, mostly on grounds of playing strength rather than for developmental reasons, so it could be argued that for all practical purposes the allocation was done by method 2.

Advantages of using the rankings directly:

1. The selection process is simple and completely transparent.

2. The selection involves very little work for anyone.

3. There is no need for a separate allocation of places to countries.

Disadvantages of using the rankings directly:

1. Any ranking system reduces each player to a single number, which is only a crude estimate of their playing strength, necessarily calculated by a "one size fits all" formula. A selection committee can look at the whole playing record and take into account any other factors they think are relevant (performance in matches, under pressure, etc).

2. Countries would lose control over the selection of their players. This could be a political issue.

3. Countries would not be able to select players using their own criteria. For example, a country may wish to allocate one of its places to the winner of its national tournament, or by a qualifying tournament. Alternatively they may wish to select an improving player or a volatile player who is not stronger overall than some other candidate, but whom they nonetheless believe has a better chance of performing well at the championship.

4. The ranking system does not cope well with players who play very few games against strong players. Such a player may or may not know how to cope against strong players. Their ranking will not be a good guide to this, because each player is ranked based on the profile of games they have actually played. A selection committee can judge each case on its merits, and consider how well that player is likely to play against the strong opposition in a world championship.

The fourth point here is related to one of the specific points the committee was asked to address, so we expand on it. Any ranking system reduces each player to a single number which is based on the games they have actually played. (This is not specific to the CGS.) It necessarily assumes that a very high percentage of wins against weak opposition is the equal of a 50–50 record against players of the same rating. For instance, take one player who plays most games against players of grade 2300, and wins 50% of them, and a second player who plays most games against players of grade 1800, and wins 90% of them. The CGS would grade both players at 2300. The second player may be tactically good against strong players (for example, has played them regularly in the past but has had limited opportunity recently) or may have no idea how to play against strong players. A selection committee with knowledge of the player and his/her history of results

can make a judgement here which the rankings do not. In this case a national selection committee is more likely to have knowledge of such players than a WCF committee and so is more likely to be able to make an informed decision.

We return now to the general discussion. If a committee is used, delegating to member associations is generally better since a local committee will have better knowledge of the players than a single WCF committee, and it is locally accountable to the players it is selecting. The WCF management committee is not elected to perform as a selection committee, and it is not directly accountable to the players in the same way. The only advantage of a WCF selection committee is that there is no need for a separate process to allocate the numbers of places to each country.

It has always been recommended that national selection committees should use the rankings only as a rough guide, not exclusively, to make selections. This is not a particular feature of the current ranking system – no other system can take into account all the potentially relevant factors in each individual case. To go exclusively to the rankings, a specific policy decision would have to be taken that this would be the sole criterion for entry, rather than any other measurement of who the "best" players were.

One issue which occurs at present is that, since countries are allowed to fill their allocation of players how they like, if their strong players are unavailable then they can send weaker players instead, rather than these places being given to stronger players from other countries. Given that they were allocated their places on the basis of the number of their strong players, this would seem undesirable. Using the rankings directly, or having places determined by the WCF initially are two ways to solve this problem, but they have the disadvantages mentioned earlier. A different approach would be to say that each country must nominate only players who have reached a certain "qualification standard". For example, they might have attained a grade of 2100 in the last twelve or twenty four months. (In practice, it may be best to say that each country can nominate up to one player who does not meet this standard, for developmental reasons. Any further players who have not met the standard could still be eligible for wildcards.) If a country cannot find enough players who have reached the qualification standard, their places should be redistributed to the countries who have a surplus of players who have met the standard, for allocation by their selection committees.

We have not considered the actual formula for the initial allocation of places to countries, as it is secondary to all the other issues here. The current formula seems overly complicated and is not well understood. It could easily be simplified.

4 Volatility and using the rankings to find the best N players

This committee was asked (email from Keith Aiton, 21 Feb 2008) *to consider, among other things, whether the ranking system, either as it is or adapted in some way, could be used simply to select the top N players for a world championship.* Our original task list included a request to report on *the volatility of the System and the algorithm utilized.* Since volatility usually enters discussions about ranking, we start with some observations about it. We then prepare a menu of available methods for top N selection.

What is here called a ranking system is in fact a rating system i.e. the players are not merely ranked in order of performance level, but also rated, e.g. by Grades, so that the Grade difference corresponds to the probability that the higher ranked player will beat the lower ranked one. We denote by CG the currently used CGS Grade.

4.1 Volatility

Volatility means different things to different people. Here we look upon it as the amount of change in rank positions from one month to the next. While rank positions change after every game, that is normally not visible. Ranking lists are typically published on a monthly basis.

It is not hard to design a ranking system with low volatility. One could modify Idx50 (single Index with step-size 50) so as to become Idx1 (single Index with step-size 1) or one could use the lifetime average of all Indexes of a player as ranking statistic. Either of these two systems will have very low volatility, but will reflect current performance levels very poorly. Of course, they are extreme examples, mentioned just to illustrate that low volatility may be achieved be at the expense of accuracy. For a ranking system to track rapid improvers and rapid decliners properly, a certain degree of volatility is inevitable.

One cannot be sure that the top, say, 60 players listed by any ranking system are actually the best 60 players. However, the top 60 players listed by a relatively accurate but possibly fairly volatile system will be reasonably certain to include the best 30 players.

To compare volatility of CG with that of other systems, we look at twelve ordinary ranking lists at monthly intervals, end September 2006 through end August 2007 (31 August 2007 is of interest as "Allocation Day" for the 2008 WCC). These

lists define a certain player population, namely those that appear on all twelve lists (10 games played in the preceding 12 month period applicable to each list). Every player in this population has a Rank Variation over the test period Sep06 through Aug07 with respect to the ranking system. It is the sum of the absolute differences of consecutive rank positions. For example, a player with CG rank positions:

$$9, \quad 10, \quad 10, \quad 8, \quad 8, \quad 8, \quad 8, \quad 8, \quad 6, \quad 3, \quad 4, \quad 5$$

will have the CG rank variation

$$|10 - 9| + 0 + |8 - 10| + 0 + 0 + 0 + 0 + |6 - 8| + |3 - 6| + |4 - 3| + |5 - 4|$$
$$= 1 + 0 + 2 + 0 + 0 + 0 + 0 + 2 + 3 + 1 + 1 = 10$$

The Average Rank Variation over all players of a population becomes a statistic for CG which quantifies its relative volatility (relative to that population over that sequence of ranking lists). Similarly, the Bayesian Grade BG has an Average Rank Variation. The following table lists these statistics with respect to the mentioned twelve ranking lists. In the bottom row, for further comparison, appears the entries for a CGS-like ranking system consisting only of an Index with step-size 50 (step-size is the maximum possible post-game Index adjustment). This table gives a quantified idea of how the mentioned systems compare as regards volatility.

Average Rank Variations

System	Top 100 players	All players
BG	40.3	102.5
CG	41.5	102.5
Idx50	80.5	163.8

Volatility as such should not be frowned upon. It is a natural phenomenon. A ranking system has the task of deducing the "true" ratings of a player population from game results. However, game results in any sport are subject to randomness. When players play at a constant performance level, game results could be likened to the tossing of a loaded coin. A coin loaded to give 70 heads and 30 tails per 100 tosses does not necessarily give a 70–30 split in any particular trial. It may give 74–26 on one occasion and 61–39 on other, along with numerous other outcomes. This randomness causes, at any moment, some players to be overrated while others are underrated. If an overrated player wins his next game, he becomes even more overrated. Thus, in the short term, rating accuracy is not necessarily improved by every game result. In the long run things average out and, over a period of time, the ratings produced by the system are a reasonable approximation of the true ratings. But they do not converge to the true ratings, no matter how many games

are played. The preceding remarks were based on the assumption of constant player performance. In the real world, fluctuating player form further complicates the situation and makes the task of the ranking system still more difficult. In view of all this, volatility is not surprising.

Nevertheless, the question can be posed whether a particular system is excessively volatile. This question is difficult to answer, because there is not a known norm against which volatility could be measured. However, we can say that CG and BG are similarly volatile, while Idx50 is much more volatile than either of them.

While conceptually quite different, BG and CG produce remarkably similar ranking lists in practice. This is reassuring. It suggests that while a ranking list (by either system) is never absolutely correct, it provides a credible comparison of player performance levels.

4.2   Criteria for finding the top N players

For allocation purposes, possible selection criteria include the following:

Expected performance (EP) i.e. how well the player is expected to play in the eventual World Championship.

Achieved performance (AP) i.e. how well the player has performed over a designated recent period e.g. the 12 month period that precedes Allocation Day.

No statistical analysis or logical argument can prove that one of these criteria is better than the other. The choice is ultimately political and thus beyond the task of this committee. We focus on methods for comparing players with regard to either criterion.

Ranking according to Expected Performance.   Any ranking system whatever can be used as a predictor of future success: the implicit prediction is that the better ranked player will do better. However, not all ranking systems are equally good at such prediction. A down-to-earth method for comparing predictive efficiency of systems of any kind is to determine the percentage of correct predictions (PCP) of each system over the same sample of test-games. The prediction could be based on

Pre-game grades: grades at the start of the game,

Pre-event grades: grades at start of the event used for each game in the event,

Well-before-event (WBE) grades: grades on a prescribed day several months before the event.

All three procedures are of interest. Pre-game grades are most relevant when updated rankings are used to seed a Knock-Out ladder after block play. Pre-event grades are of interest to tournament organizers generally because that is the form of rankings they mostly have available. WBE grades are relevant for Expected Performance ranking because the Allocation Day is always several months before the event. We report on comparisons involving all three kinds of prediction. All test-game samples are taken from World Ranking games.

COMPARISON A
Test-game sample (120371 games): the games played since Jan 1996 in which both players had at least 30 games recorded in the system database.
Results:

System	Pre-Game PCP	Pre-Event PCP
BG	69.21	68.93
CG	68.78	68.54

This gives a general comparison. Games since 1996 are used because the ranking system by then was in full swing. The PCP entries for Pre-Game and Pre-Event are comparable (because they were obtained via the same sample) and they illustrate the typical decline in PCP when older grades are used for prediction.

COMPARISON B
Test-game sample (1487 games): the games played in the first 52 events (first two months) of 2008 in which both players appeared on the 31 Aug 2007 ranking list.
Results:

System	Pre-Game PCP	WBE PCP
BG	70.28	68.80
CG	70.28	68.59

This comparison is of interest because it includes the games of the 2008 World Championships and it involves the players who appeared on the ranking lists of the presumed Allocation Day for that event. Thus it gives an idea of relative predictive efficiencies of a BG and CG ranking list on Allocation Day for Expected

Performance in the WCC. Again, the similarity is reassuring. Note that PCP values depend on the game disparities of the sample (higher disparity would give higher PCP). The Pre-Game and WBE entries for sample (B) further illustrate how older grades give weaker prediction than more recent grades. The results suggest that BG and CG have comparable predictive ability of Expected Performance. It is unlikely that any system will be found that will do significantly better than these two. Therefore, if the selection of players for the 2008 WCC had been based on an ordinary ranking list on Allocation Day (either BG or CG), it would have been a reasonable selection based on the criterion of Expected Performance.

However, there are qualitative considerations that deserve attention. There is a possible perception that the ranking list on Allocation Day will favor a player who happens to be in good form at that particular time over a player who happens to be in a slump. To the extent that this is true, it is equally true for the World Championship itself. Since the date is known in advance in both cases, it is up to all players to exercise their ability to reach peak form at the appropriate time. There also looms the possibility that some players may try to boost their grade by choosing the games they play as Allocation Day approaches. There is divided opinion on the extent to which such methods can be used to boost one's grade, or if they would be widely practised. Another consideration is the long known lag effect of CG.

Ranking according to Achieved Performance. We describe two systems for AP ranking, both of which can be derived from either BG or CG ranking. For simplicity and relevance we consider the 12 month period until end August 2007 (the presumed Allocation Day for the 2008 WCC).

CONSOLIDATED RANKING works like this. The statistic Cna of a player is defined to be the average CG-rank of that player on the 12 monthly ordinary ranking lists at the end of each month up to Allocation Day, where the average is taken over the first month and subsequent active months of the player. Let us illustrate by listing the following CG-based data for one of the players.

Month	Ranking	Games played
1	63	0
2	84	18
3	100	17
4	98	6
5	83	17
6	72	9
7	60	8
12	54	0

The table shows the players rank positions in months 1–7, and also in month 12. He was inactive in months 8–12 but his rank positions did change slightly also in those inactive months. For this player the statistic Cna = average of ranks 1 through 7 (average of rank over first and subsequent active months) is 80. His Consolidated Rank position (when players are ranked according to Cna) turned out to be 70. The average is not taken over all twelve months to avoid an unintended weighting of the rank position on which the player became seasonally inactive.

While this ranking method has appealing simplicity, it is not pleasing in all respects. If a player has 5 games in one month and 25 in another, the ranking associated with those months are equally weighted. This implies an unintended weighting of rank positions is at work. The order in which games are played in an active month does influence the next rank position: those closer to the ranking generally have more influence. The lag effect of CG makes the relative influence of the games more mysterious.

It has been speculated that a period-based ranking based on CG would give better Well-Before-Event prediction than an ordinary CG ranking list. Observed results do not support this speculation. Here is an example. The sample games are chosen to consist of all World Ranking games in the first 52 events of 2008 in which both players qualified for Consolidated Ranking i.e. they appeared on each of the 12 monthly ranking lists in question. This gives a sample of 1152 games. The PCP for Consolidated CG-Ranking was 68.84, compared to 69.18 for the ordinary CG ranking list on Allocation Day. The corresponding numbers for Consolidated BG-ranking and the ordinary BG ranking list on Allocation Day both came to 69.18.

Let us now describe WEIGHTED WINS/LOSSES RANKING. To grade a player P we use only the following known facts about him: his wins and losses over the period in question and the Grades of his opponents at the time of each game. Note that we don't use any previous ratings of P himself. We then guess a Trial Grade for P and use it to compute the Losers Win Probability in each game played by P in the test period. Each such game is weighted by its Losers Win Probability: positive in case of a win and negative in case of a loss. If the net sum of these weights is positive, it means that the Trial Grade assigned to P is too low; if it is negative the Trial Grade is too high. The computer program can easily determine (within six or so trials) the Trial Grade which gives a zero net sum (within reasonable tolerance). That gives an appropriate Achieved Performance Grade for the player in terms of the assumed known Grades of his opponents. WWL ranking has a potential problem as regards a player who has a large number of wins and few if any losses. An effective way to overcome this problem is to require a certain minimum number of moderate disparity wins as well as moderate disparity losses to be eligible for ranking. For the vast majority of

players of interest in Top N ranking, this kind of requirement presents no obstacle.

WWL ranking provides an assessment of Achieved Performance in which there is no unintended weighting. In particular, the order in which the games are played has no influence on the eventual grade. That is considered desirable for Achieved Performance ranking (not desirable for EP ranking). In this respect it is superior to Consolidated Ranking.

4.3 Discussion

If two Top N lists are created, one via EP and one via AP, one could expect to find the usual suspects near the top on both lists but considerable dissimilarity near the bottom. In the case of AP listing, a slump during the last three months of a 12 month period will be equivalent to a slump during the first three months. For EP listing the timing of such a slump has considerable influence. A steadily improving player will do better with EP ranking than with AP ranking while the opposite holds for a player with declining form. These observations point out differences which ought to be taken into account in an eventual choice between EP and AP. Based on the attributes of the relevant systems discussed above, we summarize our conclusions by making the following recommendations.

- If Allocation is to be made on the basis of Expected Performance, then the ordinary ranking list on Allocation Day should be used.

- If Allocation is to be made on the basis of Achieved Performance, then a period based system such as Weighted Win/Loss ranking for the 12 month period ending on Allocation Day could be used, although this is not the current official world ranking system. The allocation process is more transparent if the official rankings are used, so this would raise additional issues.

5 Using the rankings for seeding

The committee was asked to consider the seeding process for the knock-out stage of the world championships. The current system is to specifically order seeds 1 to 8, then have seeds 9—12 equal, 13—16 equal, 17—24 equal, and 25—32 equal. All these seeds are allocated according to the world ranking list as computed at the end of the block stage.

5.1 Context

The system of having groups of equal seeds seems to work well. In practice it is hard to distinguish all the players, and grouping the seeds like this prevents manipulation because one cannot know in advance exactly which opponents one will face. In practice the top eight players can be better distinguished, so seeding them individually will be fairer. The real issue is the method of allocating the seeds. The alternatives are by a seeding committee, seeding using the block results, or using the rankings (perhaps in a different way).

There were specific issues in 2005 and 2008. In 2005, Fulford and Bamford were by a margin the best players in the difficult conditions at Cheltenham. Any seeding committee would have made them the top two seeds, and it surprised no one that they both reached the final. However, they were actually seeded 3 and 4, and Clarke and Maugham were seeded 1 and 2. They were ranked above Fulford and Bamford on account of having slightly better results in easier conditions in the earlier part of the season, in particular at the British Opens.

In 2008, the top five seeds were all very close. Indeed, the seeding of the top 4 would have been different had the warm-up match (England v Other internationals) not been removed from the rankings. Bamford's additional two games would have been enough for him to go down from first to second in the rankings. That would, for example, have put Clarke and Fulford in opposite halves of the draw. The decision that the games would not count for the rankings was not made with the seeding in mind, and there is no suspicion of manipulation, but it is unfortunate that the seeding should have been affected by this decision which was made after the match, hence after the results of the games were known. There was no clear position on whether these games should have been included. Although many of the players were concerned more with getting practice than necessarily with winning, that is little different to any start of season event or plate event. There have to be some restrictions on what can go in the rankings, or players can invent closed events for themselves at short notice specifically to manipulate their rankings. In this case the main issue is not whether or not the match should have

been included in the rankings, but that the position was not made clear to all the players in advance.

5.2 Discussion

The system of seeding based on the block results was used in the early WCF World Championships. It was abandoned after 1992, when it produced a draw with 7 of the best 8 players in the same half, making the knock-out stage very unfair. Half the players who won their blocks were rewarded by being in the easy half of the draw, but the other half were penalised by being in the difficult half. That year may have been extreme, but the same effect is commonly seen in tournaments which seed purely on block results, hence seeding based on block results is not appropriate for croquet events.

A seeding committee would have avoided the problem in 2005 of the best two players not being the top two seeds. As in the Wimbledon tennis tournament (where the seeding committee takes into account performance on grass) the committee could take the lawn conditions into account, which no ranking system will do. Using a seeding committee also has the benefit that the players have no incentive to manipulate the seedings by deliberately losing games. However, it is a difficult job for a committee to decide seeding, even if they just decide the top four or eight players and use the rankings for the rest. Although this particular issue of the top two seeds in 2005 seems straightforward, it is not always so clear cut. Nor perhaps would it have been such a travesty if the best two players had met in the semi-final rather than the final. It is also difficult to find a committee of people who know all the international players well enough to make a good judgement, and which is seen to be impartial. Thus, unless there is strong support from the players, and a belief that a committee seen to be impartial can be found, we cannot recommend that a seeding committee be used. If it is used, we suggest it only decides the top four or eight seeds, with the others being decided by the rankings.

The current system has its flaws, which are relatively minor compared with the two systems considered already. The main complaint is that players with lower rankings coming into the event cannot be rewarded for good play in the block stages. However, the purpose of seeding is to give everyone as fair as possible a chance of winning the event. The rankings are a better predictor of expected performance than the block games alone, so seedings based on rankings are fairer than seedings based on block results. The one significant flaw with the current system is due to the game lag in the CGS grade, which means that the grade is always somewhat out of date. For example, the top four seeds at the 2008 World

Championship were decided according to the CGS grade as given in the table below.

Position	Name	CGS Grade	Index
1	Reg Bamford	2755	2634
2	Chris Clarke	2752	2769
3	Robert Fulford	2727	2719
4	Rutger Beijderwellen	2711	2693

Reg Bamford's low index shows that his recent form was not as good as his grade suggests, and his grade was dropping. Had the effect of the twenty or so games before the KO stage been taken fully into account (they count for less than the previous twenty, and less even than the twenty before that in the CGS Grade), Reg would not have been the first seed. The Bayes Grade does not have the lag effect, and the corresponding table is given below. (David Maugham and Rutger Beijderwellen were close together on both systems.)

Position	Name	Bayes Grade	SD
1	Chris Clarke	2610	55
2	Robert Fulford	2603	59
3	Reg Bamford	2592	61
4	David Maugham	2567	65

Using the rankings after the blocks for seedings, when the players are ranked close together (as in 2008), it is possible for just one or two games to change the seedings. There are cases where a player may prefer to be 3rd seed rather than 2nd, or 4th rather than 3rd etc. When the seedings can depend only on one or two games, this can give an incentive for a player to lose games. Often a player will have "dead" games at the end of the block, having already qualified. It is a poor format which puts any player in the position where their chance of winning the event may be higher if they lose a game than if they win it, and it is unfair for other players for whom those games may still be live. The game lag of the CGS grade actually makes this largely irrelevant at present. However, if the ranking system were changed to a system with no lag, then this may become an issue. In this case, it may be preferable to use the ranking list after only two days of block play for seeding rather than that after the block stages. This ensures that no "dead" games are counted, so players can have no incentive to lose these games. The benefit in ensuring there is no incentive to lose games would outweigh the seedings being based on rankings which were one day out of date.

5.3   Conclusions

The present system of seeding based on CGS grade after the block stage does not seem to have major flaws. If the ranking system were changed to one with no lag, it would improve the seeding. In this case, it may be wise to use the rankings before the last day of block play (or even, for simplicity, before the event) to ensure there is no incentive for any player to lose a game.

The alternative of using a seeding committee just for the top 4 or 8 seeds, with the rest done by rankings as above, could be considered if there is strong political support for it, but not otherwise.

It may be considered desirable to reward good play in the blocks in the seeding. Seeding purely on block results does not achieve this. However, a possible way of rewarding good play in the blocks would be to boost the seeding of block winners into the next group of eight seeds, although not into the top eight as this would cause too much distortion.

A Statistical analysis of inter-pool drift in the Croquet Grading System

Chris Dent

A.1 Introduction

Concern has been expressed that the ranking indices in the croquet grading system have found different levels within the four main continental pools of players: North America (NA), Australia (Aus), New Zealand (NZ) and the rest of the world (ROW, essentially Europe). In particular, a net drift of points from the other pools to NA has been observed over the last decade or so. This might seem to indicate that players of the same ability have lower ranking indices in NA compared to the other pools – this report sets out to answer the question of whether this drift of points is statistically significant or just a random exchange of points between pools with equivalent rankings.

The methodology used is outlined here. I've tried to tread the fine line of showing how my analysis works, without including the detailed mathematics. Some background on hypothesis testing and confidence intervals is included – a knowledge of basic probability theory is assumed, but there are no gruesome equations to follow. The full gory details may be obtained from either the Ranking Review committee or myself.

Throughout this appendix (as in the rest of the report), the indices are given on the "new" scale where A class players have indices between 2000 and 3000, as in the published rankings. (In some documents the "old" scale is used, where the corresponding range is 100 to 200.)

A.1.1 Hypothesis testing in classical statistics

Classical hypothesis testing takes a hypothesis about the distribution of a random variable, called the null hypothesis or H_0, and tests it against an alternative hypothesis H_1. This is achieved by putting H_0 on trial in a suitable way. (Here H_0 will in essence be that all is fine with the CGS, H_1 that it isn't.) H_0 is then "innocent until proven guilty".

Guilt is deemed proven if when assuming H_0 the probabilty of observing the given data *or something more extreme*, called the p-value, is less than some critical

value p_c. The data is then said to be significant at the $100p_c\%$ level, and the null hypothesis is rejected.

If $p > p_c$, then H_0 is found not guilty, as it explains the observed data reasonably well. Note that as in a criminal trial this does not prove that H_0 is true, merely that the evidence does not suggest sufficiently strongly that it isn't.

Statements such as "H_0 is accepted/rejected with $100(1 - p_c)\%$ confidence" are sometimes seen. This is at best vague, at worst meaningless or wrong, with respect to what the significance test actually demonstrates.

Here, the data will be deemed statistically significant if there is less than a 5% chance under H_0 of obtaining it or something more extreme. (5% is the critical probability most commonly used in hypothesis testing. One might call it the default which is used unless there is a good reason to go for a different, usually more stringent, value.)

A.2 Significance test for inter-pool drifts

A.2.1 One pool versus all pools

The suggestion that the rankings might have found different levels on different continents originated in the observation that there had been a net drift of points to the USA over the last n years. However, it is important to remember that even if there is no systematic difference in index between players of the same ability in different pools, there will still be some net drifts of points between pools as players randomly win and lose games. It is extremely improbable that all of these drifts will be very close to zero. To put this another way, if one looks at enough inter-pool relationships, it becomes very likely that something improbable will be seen in one of them even if there is no systematic difference in the ranking levels of players between pools (an example of this effect from everyday life is that while it's very unlikely indeed that any one ticket will win the lottery, it's quite likely that there will be a winner. It would be wrong to infer from the low single ticket probability that the lottery is somehow unfair, and that no one will win.)

As a result, even if the most extreme inter-pool drift of points is deemed to be statistically significant when tested in isolation it does not follow that the combination of drifts is significant. It is therefore necessary to find a way of testing the significance of the combination of inter-pool drifts of points observed.

A.2.2 Null and alternative hypotheses for the hypothesis test

- H_0: all players' ranking indices represent their true abilities at all times, and the true win probabilities are correctly given by the CGS formulae.

- H_1: H_0 not satisfied

This form of the hypotheses might seem very strong (no one would claim that any ranking system can actually deliver correct win probabilities at all times.) However, the question posed is whether *the data observed alone*, i.e. the actual inter-pool drifts, is consistent with the assumption that all win probabilities obtained from the CGS are correct. This hypothesis permits the analysis below, as it provides a way of obtaining the correct win probabilities for each game from the players' ratings.

The assumption of H_0 also implies that the ranking indices of players of the same ability are the same in all pools (although as game results are random there is still some drift of points.)

A.2.3 Probability distribution for the drifts

As the win probabilities for a game follow directly from the CGS formulae, assuming H_0 it is easy to work out the mean (equal to zero) and the standard deviation of the probability distribution for the index points exchanged by the participants after each game based on the two players' pre-game indices (which are assumed to be correct).

The probability distribution for the points drift per game between any pair of pools is derived using the Martingale Central Limit Theorem (MCLT). This rather fearsome-sounding beast simply implies that the sum of the single-game points exchanges between pairs of pools is Normally distributed.

The number of games played, standard deviation of the probability distribution for the average drift of points per game, and the actual observed average drift of points per game[1], are show in Table 1. Because of the very different numbers of games, the average drifts cannot be compared directly (this is expected to decrease in size as the number of games played increases, just as the proportion of heads approaches 1/2 as the number of coin tosses increases.) Comparison can me made, however, if the observed drifts are scaled by the standard deviation of the drift

---

[1]The points transferred between pools must necessarily sum to zero. At first sight this might appear not to be the case in the table, however for each of the pairs of pools the drift shown is added from one and subtracted from the other.

Pools	Games	S.D.	Drift	Z	χ^2
NA v Aus	592	0.094	-0.039	-0.415	0.172
NA v NZ	410	0.119	0.070	0.588	0.346
NA v ROW	2220	0.047	0.069	1.462	2.138
Aus v NZ	3029	0.040	0.009	0.234	0.055
Aus v ROW	1384	-0.060	-0.058	-0.969	0.939
NZ v ROW	2986	0.040	0.004	0.095	0.009

Table 1: Calculation results for the inter-pool drifts between pairs of pools. The columns are: standard deviation of the drift per game probability distribution assuming all players have 'correct' indices; actual drift per game observed; drift per game scaled by the standard deviation; the derived chi-squared statistic.

distribution to obtain the Z-statistic in the table (under the assumptions in H_0, the Z-statistic is Normally distributed with mean 0 and standard deviation 1.)

A.2.4   Requirement for a test statistic

In order to perform a significance test, an univariate test statistic (i.e. a single representative number derived from the data) is required. As described above, the Z-statistics permit direct comparison between the pools, but there are six of them.

The simplest test statistic is the largest of these scaled drifts, however this suffers from two major disadvantages. It clearly doesn't use all of the available information, and in particular it can be conservative in rejecting the null hypothesis (if the second largest drift is only slightly less than the largest, then this seems intuitively to be more extreme than if there is one large drift and the rest are very small.)

A.2.5   Chi-squared statistic

It is possible to measure the total amount of drift between all of the pools by squaring and adding the six scaled drifts (Z-statistics) between each pair of pools. The squared drifts for each pair of pools are shown in Table 1, column χ^2.)

Under H_0, this test statistic follows a standard probability distribution, called the χ^2 (chi-squared) distribution with six degrees of freedom. The value of χ^2 observed is 3.659, and the probability of observing this amount of drift or larger is 0.723. There is therefore not sufficient evidence to suggest that the ranking levels

in the four pools are different.

In order for the data to be statistically significant, the χ^2 statistic would need to be 12.592, which is 3.4 times the value observed. As χ^2 is calculated from the squares of the scaled drifts, this would require the typical drifts between pools to be about twice as large as those seen.

A.2.6　What if the data had been statistically significant

If the chi-squared statistic were statistically significant, this would be taken as an indication that the ranking levels between pools are different, i.e. that players of the same ability do not have the same index in all pools. It would then be possible to estimate by how many points the index levels differ.

To give the range of index level differences which are plausible with respect to the data, the statistical concept of a confidence interval can be used. This is essentially the range of values of a parameter which pass a significance test of the kind described above.

Assuming that Aus, NZ and ROW have the same ranking levels, and working to the same significance level as before, a confidence interval for the ranking level of North American players is from 28.5 points under-ranked to 6.4 points over-ranked. Looking at the ends of this interval, the observed data implies that any systematic misranking of NA players must be small. As the interval includes zero, this analysis once more provides no evidence of any misranking at all.

A.3　Conclusion

Even if the ranking system gave correct win probabilities for all games at all times, there would still be some exchange of index points between pools of players, as games are randomly lost and won. The quantity of points exchanged between North America, Australia, New Zealand, and the Rest of the World, is sufficiently small that it can be explained very plausibly by these random events alone, without any requirement to assume different ranking levels in different continental pools of players. There is thus no evidence that players of the same ability have different ranking levels in the different pools.