Minute 7/12/19: *Project of Team Pairing System for Olympiad* DRAFT

Investigation about Batumi 2018 Chess Olympiad Pairings

ABSTRACT

In this paper, the pairings of Batumi Olympiad are scrutinized and compared to some previous Olympiads, with the aim to verify their fairness. The pairings were examined mainly by analysing the frequency of very unbalanced matches and of average opposition met by teams. Also, some consideration is given to technical aspects of the pairing systems such as the sorting method inside scoregroups and its effects.

PREMISE

SPP Commission was asked to investigate upon some facts related to the pairings made for the 2018 Batumi Chess Olympiad. Namely, the Commission was asked to discuss three proposals from GS Commission – an extract follows:

1. ***Proposal for the individual Swiss pairings system***

   The pairing system currently used in individual Swiss Tournament does not ensure equal chances for all the participants: statistically, players with lower ratings encounter much stronger opponents in order to reach the top of standings compared with higher rated competitors. GSC proposes to find the fairer pairing system. Dubov's Pairing System is likely to be tested.

2. ***Proposal for the team Swiss pairings system***

   Taking into consideration numerous complains related to the current pairing system, GSC proposes to revise the current pairing and tie-break system for the World Chess Olympiad.

3. Taking into consideration "extra Black game" for the individual Swiss tie-break system

GSC proposes to introduce the "extra Black game" adjustment – a number between 10 and 15 (to be specified) added to the Rating Performance (when the latest is used as a tie-break criterium) for players having played more games with Black in a Swiss tournament.

Two more proposals were sent by Mr. Holowczak, Chairman of TAP – an extract follows:

While there is a general criticism of the pairing sorting criteria being different from the ranking sorting criteria, there are also specific issues with the current system of resorting by game points early in the tournament, specifically in Round 2. (…)

15	24	ESP	Spain	4	2	**3½**	:	**½**	2	4	Dominican Republic	DOM	82
16	83	KGZ	Kyrgyzstan	4	2	**½**	:	**3½**	2	4	Greece	GRE	25
17	26	SLO	Slovenia	4	2	**4**	:	**0**	2	4	South Africa	RSA	84
18	86	IRQ	Iraq	4	2	**1½**	:	**2½**	2	4	Vietnam	VIE	27
19	87	AND	Andorra	4	2	**0**	:	**4**	2	4	Brazil	BRA	29
20	32	SWE	Sweden	4	2	**3½**	:	**½**	2	4	Tunisia	TUN	88
21	34	ITA	Italy	4	2	**1**	:	**3**	2	3½	Azerbaijan	AZE	4
22	5	IND	India	3½	2	**3½**	:	**½**	2	4	Austria	AUT	35
23	36	UZB	Uzbekistan	4	2	**1½**	:	**2½**	2	3½	Ukraine	UKR	6
24	8	ARM	Armenia	3½	2	**2½**	:	**1½**	2	4	Kazakhstan	KAZ	37
25	38	NOR	Norway	4	2	**2**	:	**2**	2	3½	Georgia 1	GEO1	14

In Round 1, there were enough 4-0 wins such that Sweden, seeded 32, is playing Tunisia, seeded 88. Both teams won 4-0. However, Italy are seeded 34, but they won 4-0 in Round 1 and their reward was to be paired against the highest-seeded team that won 3½-½ in Round 1, Azerbaijan, seeded 4. Notwithstanding the result of the match in Round 1, this doesn't seem to have been very fair on Italy, who played a much higher-rated team than Sweden did, despite them both winning 4-0 in Round 1 and being very similar strength teams on rating. This doesn't appear to be fair. TAP investigated two potential solutions to the problem.

***Solution 1**: Rather than sort by gamepoints, sort by the Olympiad tie-break. Due to the way this is calculated, this has the effect of simply pairing by seed in Round 2, because the lowest result is dropped in the calculation of the Olympiad tie-break, and thus everyone's score is 0 in Round 2 because they have only played 1 match, which must be dropped.*

***Solution 2**: Sort the scoregroup by matchpoints and then seed, ignoring gamepoints won. This is logical if a comparison is drawn to an individual Swiss tournament; there is never* *any suggestion that each scoregroup should be sorted by the tie-breaks before doing the pairing, so why should that apply in a team competition?*

The European Chess Union has opted for Solution 1, but as this paper hinted earlier and will go on to explain, TAP is not minded to retain the existing Olympiad tie-break on the grounds that it is too difficult to be calculated. For that reason, TAP is minded to propose solution 2 to solve the pairing problem.

DISCUSSION

For the sake of simplicity, we will subdivide the discussion of the above issues into several points, even if every aspect of pairings actually interacts with every other one.

Stronger opposition for lower-ranked players

Let's then begin with noting that any Swiss pairing system can only work on a statistical basis – this means that, in looking for fairness, we can only analyse the overall, statistical behaviour of the system, while sparse cases of "bad luck" remain always possible and are in fact unavoidable.

The first objection to the pairings is that "*players with lower ratings encounter much stronger opponents in order to reach the top of standings*". Actually, this behaviour is deeply rooted inside the theoretical foundation of all rating controlled Swiss systems. Its rationale is that the convergence of the selection process (and subsequent formation of standings) is faster - and way more reliable - if weaker players have early games with stronger players. In doing so, stronger players will soon get a higher score than weaker players, as should (statistically) happen. When ratings are meaningful, the opposition to higher rated players is unavoidably formed by lower rated players, at least on the average. This happens just because they are higher rated, i.e. stronger – for example, the lowermost rated player can only meet higher rated opponents, and will therefore have the hardest path to the top standings. On the contrary, the topmost rated player will have the easiest one. Even in a round robin tournament, the lower-rated players will (of course) get higher AROs.

Should weak players initially meet weak players, and strong players meet strong players, after a few rounds we would have high-scored weak players, and low-scored strong players. This is just what happens with an accelerated system - hence the need to have enough "normal" (un-accelerated) rounds after the accelerated ones, to allow the "abnormally-high" scores to subside and the "abnormally-low" scores to rise, until equilibrium is reached. Very peculiar pairings may appear during this settling phase, with large differences in ratings that upset players.

(Because of this, acceleration is usually used only when the differences in ratings are so large that the results of the games in the first rounds are so much predictable as to be pointless; or when the presence of too many low-rated players would seriously impinge on the probability of title norms.)

The attenuation of the overall strength of the opposition (in practice, of ARO) for lower rated players can be obtained only by the use of some kind of accelerated pairing - however, in view of the well-known Dresden Olympiad experience, SPP does *not* recommend the use of acceleration for Olympiad pairings before enough experience is acquired.

The question of unfair opposition is also put forward in the foreword to TAP proposals, by means of the example of a scoregroup in the second round pairing (see table in TAP proposals), which yielded some "easy" pairings together with some "tough" ones. It is however worth noting that, in that scoregroup, the pairings would have been the very same even if the pairings had been made by means of different sorting criteria, namely either pairing number ("seed") or tie-break (OSB) order.

Use of Dubov pairing system

The goal of Dubov pairing system is to equalise opposition in the sense of obtaining as equal as possible AROs for player having the same score. This result is sought for by using ratings as a measure of the real players' (and thus teams') playing strength – it can however be pursued for each team only (approximately) on every other round, because Black players' ratings are used to level out their (White) opponent's ARO.

Because of its nature, Dubov system can only be effective if all the ratings are reliable. Actually, however, this only happens for professional players, while ratings for amateur or very young/very old players are often unreliable. In Olympiad, of course, we have many amateur-level teams that, especially in the first rounds, unavoidably mingle with highly professional ones.

Moreover, Dubov system puts much store on colour balancing, which is however far less important in team competitions than in individual ones.

Because of all this, Dubov doesn't seem to be a first choice system for Olympiad pairings.

Very unbalanced pairings

A very unbalanced pairing (VUP for short) is a pairing that yields a 4:0 or 3½:½ outcome. Usually, such a result shows a decisive difference in strength between the paired teams. This is normal – and sought for – in the first rounds of a Swiss tournament, but should not happen too often in late rounds. To analyse the behaviour of the pairing system in this respect, all the chess Olympiads since year 2000 were examined.

The results are collected in the following table, where only actually playing teams are counted (see Table 1). For each round, the average ***m*** and standard deviation ***σ*** of the number of very unbalanced pairings are calculated, and a confidence interval ***m±σ*** is determined (this interval should contain approximately 67% of all the items). Rounds falling below this range are marked in green, meaning a very well balanced pairing, while round exceeding this range are marked in red, meaning disequilibrium.

#	Olympiad Site	Year	Teams	Rnd 1	Rnd 2	Rnd 3	Rnd 4	Rnd 5	Rnd 6	Rnd 7	Rnd 8	Rnd 9	Rnd 10	Rnd 11	Rnd 12	Rnd 13	Rnd 14	AVG	Total	total/tms
34	Istanbul	2000	126	55	23	18	18	16	17	11	14	13	19	10	12	21	11	18,4	258	2,05
35	Bled	2002	135	56	22	13	10	13	19	10	13	16	17	13	11	14	11	17,0	238	1,76
36	Calvià	2004	129	48	25	18	14	10	17	13	13	13	11	8	15	10	14	16,4	229	1,78
37	Torino	2006	148	56	31	24	24	13	18	17	9	15	16	19	17	18		21,3	277	1,87
38	Dresden	2008	146	38	24	40	30	18	22	19		15	23	20				24,8	273	1,87
39	Khanty-Mansiysk	2010	148	63	30	28	19	21	24	18	20	17	16	13				24,5	269	1,82
40	Istanbul	2012	157	63	27	28	24	25	26	20	22	25	16	15				26,5	291	1,85
41	Tromso	2014	172	77	44	29	38	21	36	31	23	15	24	22				32,7	360	2,09
42	Baku	2016	170	81	45	30	32	24	20	28	23		16	21				31,1	342	2,01
43	Batumi	2018	185	81	51	30	30	29	22	30	25	26	20	24				33,5	368	1,99

		Rnd 1	Rnd 2	Rnd 3	Rnd 4	Rnd 5	Rnd 6	Rnd 7	Rnd 8	Rnd 9	Rnd 10	Rnd 11	Rnd 12	Rnd 13	Rnd 14	AVG	Total	total/tms
Average		61,8	32,2	25,8	23,9	19,0	22,1	19,7	18,6	17,7	17,8	16,5	13,8	15,8	12,0	24,6		1,91
Sigma		14,3	10,5	7,8	8,7	6,1	5,7	7,7	5,8	4,8	3,8	5,4	2,8	4,8	1,7	6,4		0,12
Confidence range	min	47,5	21,7	18,0	15,2	12,9	16,4	12,0	12,8	12,9	14,0	11,1	11,0	11,0	10,3	18,2		1,79
	max	76,1	42,7	33,6	32,6	25,1	27,8	27,4	24,4	22,5	21,6	21,9	16,5	20,5	13,7	31,0		2,03

Table 1: Number of very unbalanced pairings (see text) per round

In a normal (i.e., not accelerated) pairing, the number of VUPs should decrease (statistically) exponentially from the first round on. The effect of acceleration, quite apparent in Dresden Olympiad, is to push down the initial peak, at the price of an increment in the 1-2 rounds immediately following the removal of fictitious points. The number of VUPs is likely determined by many interacting causes, one of which is the number of surprise results in previous rounds. Such unexpected results, although always present, tend to happen more often when there are many “unpredictable” teams – that is, teams whose ratings are not so good a measure of strength as those of highly professional teams.

Inspecting the table we find that the first three editions, where only a limited number of very good teams took part, show a fairly good balance in pairings. The Dresden Olympiad, first one in which accelerated pairings were used, shows some unbalanced rounds, namely the third (acceleration removal) and the eighth and tenth. The overall count for all rounds, compared with the number of participating teams, was under average. However, unbalanced pairings in late rounds are not well liked by players, as they give a sense of "unfairness".

The last editions (since Tromso), which had a far larger attendance than the previous ones, show not only many VUPs in the first two rounds (which are, as we saw above, intrinsic in Swiss systems) but even in much later rounds.

The eye-catching difference between these and the previous Olympiads suggests that we should try to analyse the data with an eye to number of teams too, as the latter is the most apparent difference. The above table was therefore recalculated considering for each item the ratio between number of VUPs and number of participating teams (see Table 2 below).

#	Olympiad Site	Year	Teams	Rnd 1	Rnd 2	Rnd 3	Rnd 4	Rnd 5	Rnd 6	Rnd 7	Rnd 8	Rnd 9	Rnd 10	Rnd 11	Rnd 12	Rnd 13	Rnd 14	AVG	Total	st.dev.
34	Istanbul	2000	126	0,44	0,18	0,14	0,14	0,13	0,13	0,09	0,11	0,10	0,15	0,08	0,10	0,17	0,09	0,15	2,05	0,09
35	Bled	2002	135	0,41	0,16	0,10	0,07	0,10	0,14	0,07	0,10	0,12	0,13	0,10	0,08	0,10	0,08	0,13	1,76	0,09
36	Calvià	2004	129	0,37	0,19	0,14	0,11	0,08	0,13	0,10	0,10	0,10	0,09	0,06	0,12	0,08	0,11	0,13	1,78	0,08
37	Torino	2006	148	0,38	0,21	0,16	0,16	0,09	0,12	0,11	0,06	0,10	0,11	0,13	0,11	0,12		0,14	1,87	0,08
38	Dresden	2008	146	0,26	0,16	0,27	0,21	0,12	0,15	0,13	0,16	0,10	0,16	0,14				0,17	1,87	0,05
39	Khanty-Mansiysk	2010	148	0,43	0,20	0,19	0,13	0,14	0,16	0,12	0,14	0,11	0,11	0,09				0,17	1,82	0,09
40	Istanbul	2012	157	0,40	0,17	0,18	0,15	0,16	0,17	0,13	0,14	0,16	0,10	0,10				0,17	1,85	0,08
41	Tromso	2014	172	0,45	0,26	0,17	0,22	0,12	0,21	0,18	0,13	0,09	0,14	0,13				0,19	2,09	0,10
42	Baku	2016	170	0,48	0,26	0,18	0,19	0,14	0,12	0,16	0,14	0,13	0,09	0,12				0,18	2,01	0,11
43	Batumi	2018	185	0,44	0,28	0,16	0,16	0,16	0,12	0,16	0,14	0,14	0,11	0,13				0,18	1,99	0,10

Average		0,41	0,21	0,17	0,15	0,12	0,15	0,13	0,12	0,12	0,12	0,11	0,10	0,12	0,09	0,16
Standard deviation σ		0,06	0,04	0,05	0,04	0,03	0,03	0,03	0,03	0,02	0,02	0,03	0,02	0,04	0,01	0,02
Confidence range	min	0,35	0,17	0,12	0,11	0,10	0,12	0,09	0,09	0,09	0,09	0,08	0,09	0,08	0,08	0,14
	max	0,46	0,25	0,21	0,20	0,15	0,17	0,16	0,15	0,14	0,14	0,13	0,12	0,15	0,11	0,18
min		0,26	0,16	0,10	0,07	0,08	0,12	0,07	0,06	0,09	0,09	0,06	0,08	0,08	0,08	
max		0,48	0,28	0,27	0,22	0,16	0,21	0,18	0,16	0,16	0,16	0,14	0,12	0,17	0,11	

Table 2: Average number of very unbalanced pairings per team, per round (see text)

Now, looking at the VUPs per team, the situation appears different. The unbalanced rounds are far less than it seemed – and we can also see that Dresden 2008 Olympiad accelerated pairings seem to be just a little worse than expected.

In pre-Dresden editions, the Burstein system was used, with Buchholz as main sorting criterion inside scoregroups. After Dresden, 2010-2012 editions used plain pairing numbers, while 2014-

2018 used game points for sorting. Now we can appreciate that the total VUPs per team is essentially the same for editions since 2006 through 2012, which had similar attendance, although three different pairing system were used for those four Olympiads. The last three editions, which had a significantly larger attendance (+15÷20%) show a total more or less +20% larger, while the first three editions are unstable in this regard (Bled and Calvià are a little better, but Istanbul 2000 is on a par with recent editions). Thus, the total number of VUPs per team seems to be only loosely correlated to the number of players – however it seems rather difficult to discern between the effects of attendance and of the pairing system itself.

We can also see that the variability from round to round, expressed as standard deviation (Table 2, last column to the right), is minimum for accelerated pairing. This is to be expected, as acceleration "spreads" the VUPs widely in unexpected rounds – this is a consequence of "queer" pairings in moderately late rounds, and is also one strong reason why players object to acceleration.

This variability is essentially the same with Burstein system and with pairing-number-driven system, while it is moderately higher for the game-points-driven pairing system used in the last Olympiads. Since the latter also had a fairly larger attendance, it is hard to say whether the reason for larger variability resides in larger attendance or in the pairing system – however, the fact that previous Olympiads had similar behaviour, independent on the pairing system, seem to hint that the cause might sooner be found in attendance.

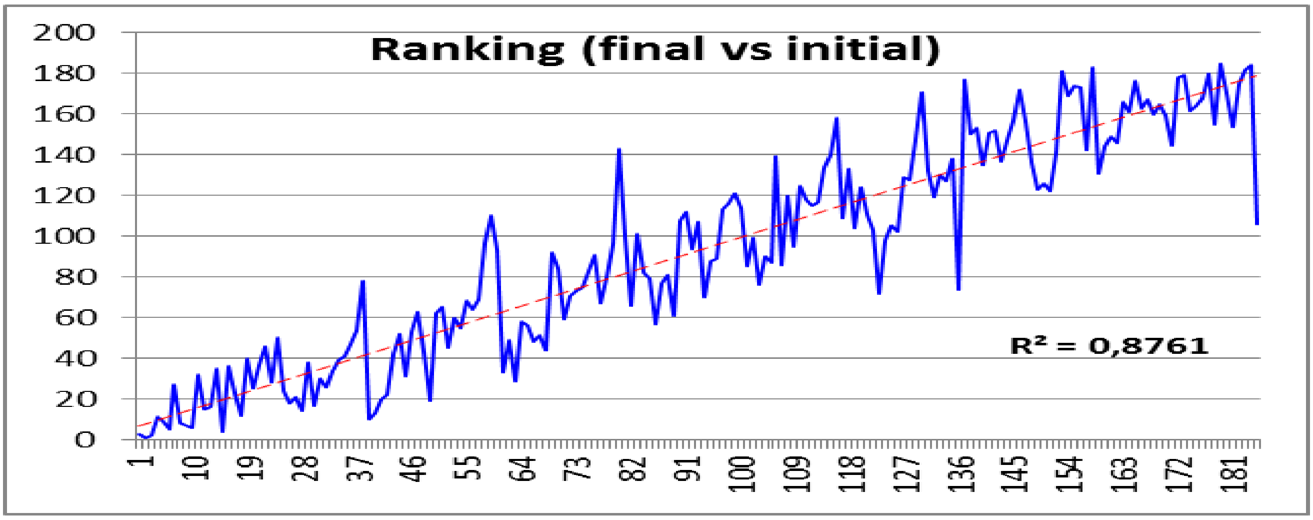
Fairness of pairing system

The question of the pairing system fairness is of course a central issue to teams and organizers. SPP therefore tried to investigate that matter, analysing the Batumi pairings in some depth. First of all, however, we should set some criteria by which to decide whether the pairings are fair.

The goal of any Swiss pairing system is of course to yield a final standing that sorts the participants (individual or teams, as the case may be) in order of playing strength. If the ratings of all teams were well correlated to their strength, the final standing should reflect the initial order list, which is represented by the pairing numbers ("seed"). In fact, the current strength of a team is a stochastic variable, whose average value is probabilistically correlated to the average strength, but which of course varies with players' conditions, opponents' behaviour and other environment parameters.

The final standings can therefore only statistically be correlated to the initial order, while we must accept some random differences as normal statistical variability.

All this is apparent in the graph below (Graph 1), which shows the correlation between initial and final ranking for all teams. The correlation coefficient is high enough to show a good correlation between the two variables, meaning that, on a statistical basis, better teams did actually obtain better places in standings. Moreover, we observe that variations are significantly smaller for higher ranked teams, as should be expected.



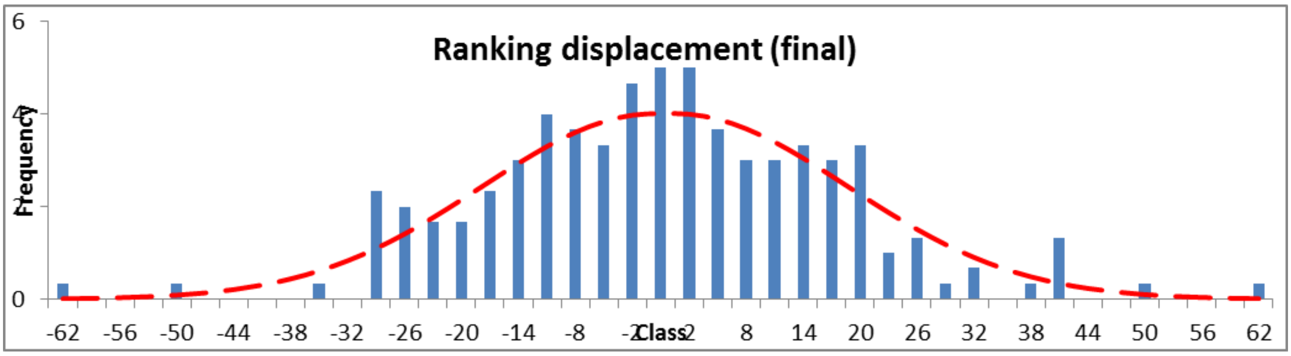
Graph 1: Correlation between final vs. initial ranking

We want now focus on the top ten teams' path through the tournament (see Table 3). Criticism has been raised against the "easy ride" of China, which however had a harder path than the runners up USA and Russia. A very hard path was indeed that of Poland, caused by the really impressive row of six won games, then two draws and then again a won game in the first nine rounds – Poland met very strong opponents because at that time it was in fact the strongest team in the competition – and, when finally it lost a match, in the tenth round, it was only to the Olympiad winner.

Final ranking	Initial ranking	Team	Team	opposition (final ranking)											average opposition
1	3	China	CHN	69	64	49	26	12	17	10	40	15	4	2	28,0
2	1	United States of America	USA	96	51	40	6	39	65	26	15	4	8	1	31,9
3	2	Russia	RUS	104	82	43	4	49	6	52	30	33	5	9	37,9
4	11	Poland	POL	81	57	47	3	9	10	15	8	2	1	6	21,7
5	9	England	ENG	124	71	63	15	33	9	41	39	29	3	21	40,7
6	5	India	IND	92	14	23	2	67	3	19	12	8	40	4	25,8
7	27	Vietnam	VIE	113	107	56	9	62	43	23	21	34	13	37	47,1
8	8	Armenia	ARM	58	21	42	20	15	48	30	4	6	2	13	23,5
9	7	France	FRA	79	77	86	7	4	5	18	10	13	26	3	29,8
10	6	Ukraine	UKR	109	16	27	41	25	4	1	9	36	15	12	26,8

Table 3: Average opposition for top ten teams in Batumi Olympiad 2018

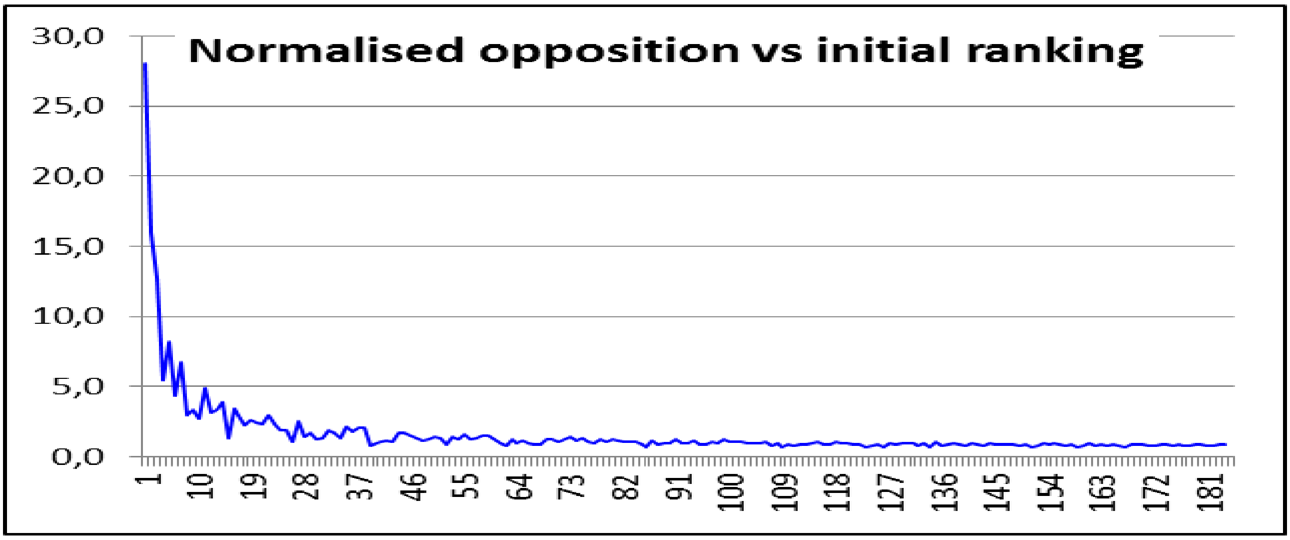
All this is further confirmed by the statistical distribution of ranking displacements (differences between initial and final ranking), shown in the graph below (Graph 2).



Graph 2: Probability density of ranking displacement (final ranking - initial ranking)

Here we can appreciate that the probability density of the displacement fits rather well to a Gaussian bell curve, meaning that the distribution is actually stochastic, and its mean is nearly zero (actually, 0.43). In other words, there is no apparent bias of the system.

From this data we can also analyse the average opposition for each team, obtaining the graph below (Graph 3). Here, the "normalised opposition" for a given team is defined as the ratio between the average final ranking of opponent teams and the final ranking of the team itself. A unity value therefore means that, on the average, the team was matched with its equals, while higher values show weaker opposition. From the graph it is readily apparent that the normalised opposition is fairly near unity for a very large majority of teams.



Graph 3: Normalised opposition (see text)

Of course, it gets rapidly larger and larger as we near the top ranked teams. As we already observed, this is *a priori* unavoidable, because there are not enough strong opponents to balance the "easier" matches of top teams (we may call it a "border effect").

The ranking displacement was also inspected by means of fast Fourier Transform for cyclic regularities (for example, differences repeating every n places in the standings) but no such anomalies were observed.

Scoregroup sort strategy in pairings

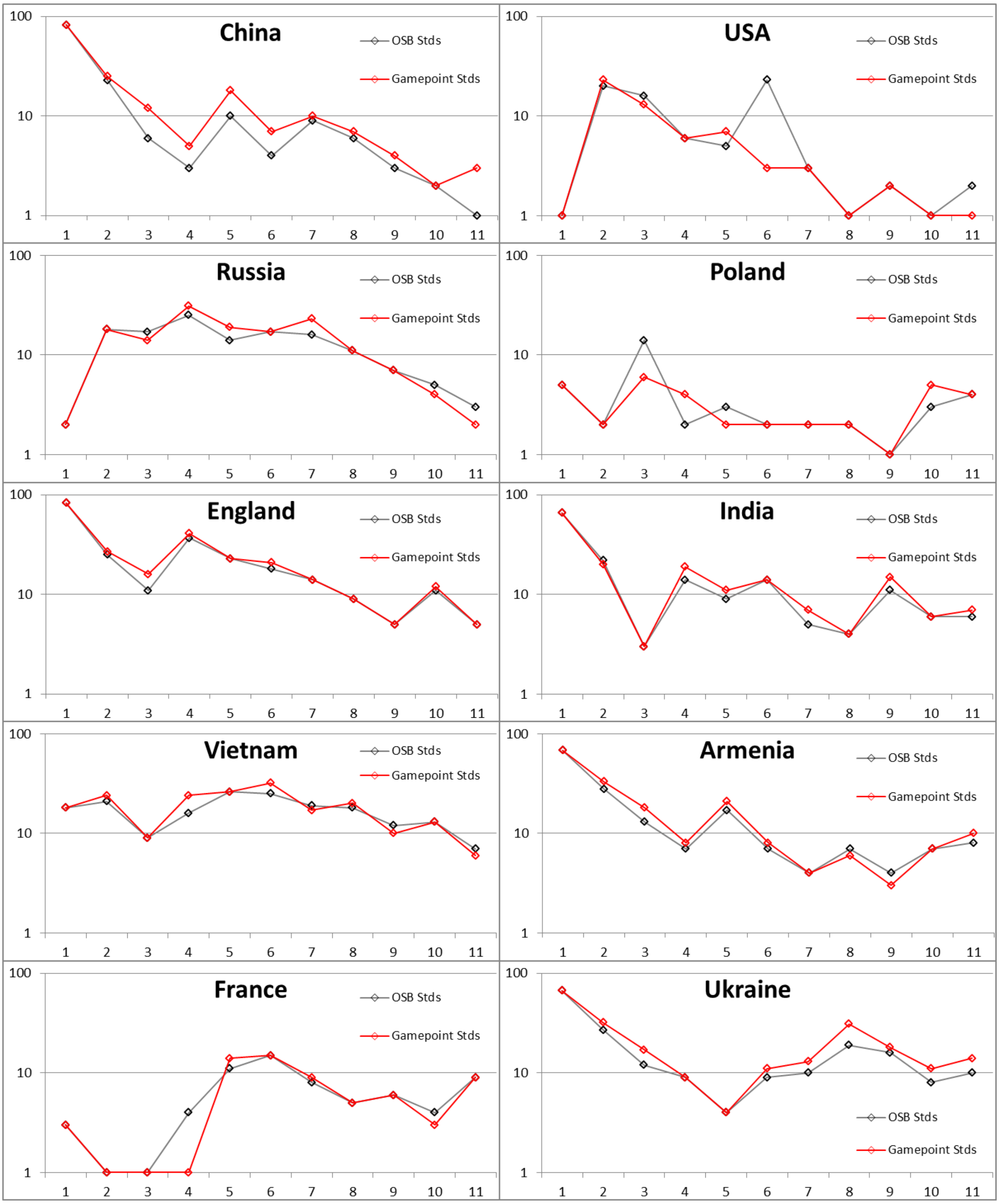
The current method for sorting teams inside scoregroups uses game-points as a driver. It is readily apparent, however, that in the last three Olympiads, which used this sorting strategy, the number of very unbalanced pairings was sometimes high even in unusual rounds, and that aroused some unfavourable reactions. As we mentioned above, it is really hard to say whether the pairing system can be blamed for it – however, some proposals were advanced to change this scoregroup sorting to some other one, namely to pairing numbers or to a tie-break, possibly the same used for standings.

Pairing numbers were used as a sorting criterion inside scoregroups for the 2010 and 2012 Olympiads. They provide a fairly simple sorting method, which is strictly related to ratings and shares therefore their pros and cons. In particular, ratings can safely be considered reliable for professional teams, so we can rely on pairing numbers to give sound and fair pairings. For weaker teams, ratings are not just as much reliable, so we could have some peculiar results, giving birth to unusual pairings – however, this behaviour should affect mainly the lower half of the ranking.

In favour of pairing numbers we ought to mention that, since they are vastly used as sorting criterion in FIDE Swiss (Dutch) system, they are very well known to most players.

As mentioned in the TAP letter, the use of a “cut” type tie-breaker like the Olympiad Sonneborn-Berger as a sorting criterion for scoregroups is inherently meaningless in the second round. Its discriminating capability is only moderate also in the immediately following rounds. By using an uncut tie-breaker we can remedy this limitation to some extent, but we can never overcome it.

The use of a tie-breaker, namely Buchholz, is part of the Burstein pairing system and was experimented during Olympiads in the years 2000 through 2006, so it is not really new. In Burstein system, however, the pairing strategy is completely different than the current Olympiad system, so that the results cannot be readily extended to our case.



Graph 4a-h: Round by round comparison between gamepoint and OSB standings for top ten teams

To try and shed some light on the matter, an analysis was made on the top ten ranking teams, to visualize the differences in standings – and hence in ranking positions, were the tie-breaker used for scoregroup sorting. Of course these results are only meaningful for the top teams. The graphs (Graph 4a-h, above) show that in general gamepoints and Olympiad Sonneborn-Berger yield similar results, but in some cases there are significant differences. This happened for example in the third round for Poland; in the fourth for France; in the sixth for USA. In all three the order obtained by gamepoints gave a stronger estimate of the team. For China, India and Ukraine, the OSB gave on the contrary a weaker estimate that was far smaller but lasted many rounds. Changing the scoregroup sorting to OSB would have immediately produced different pairings – for example, Poland would have got an easier pairing in the third round, and thus an increase in its winning probability (however, the team *won* that round). Thus it would have got a tougher opponent in the fourth round, decreasing its winning probability. There's of course no way to know what the outcome of the match would have been – however, the average opposition would likely remain more or less the same.

"Extra Black Game" criterion in tie-break

It is well known that having Black rather than White statistically entails a lower actual rating. However, at the moment there is no way to know exactly how large the difference is, although some research on the subject was done in the past. (Mr. Roberto Ricca, former Secretary of SPP and now member of the TEC Commission, can probably supply more information on the matter.)

It would seem reasonable that, for tie-break purposes, a correction be applied to average ratings based on colour, possibly on a game-by-game basis. However the matter requires much analysis and SPP Commission is not in charge of the subject of tie-breakers, except insofar it may affect pairing systems (e.g., Burstein system).

CONCLUSIONS

The analysis of the above data shows that there is a good correlation between playing strength (as represented by ratings) and final ranking position of high level teams, and that there is no apparent bias in the pairings. We can therefore conclude that the pairing system was fair, even if better systems can exist.

The discussion yields no certain conclusion about the use of tie-break criteria for use in scoregroup sorting. The adoption of pairing numbers as a sort driver seems to be a possible choice, all the more in view of the fact that it is an easy and fairly well-known scoregroup sorting strategy.

SPP Commission cannot recommend Dubov system at present, because data regarding its use in team competitions is almost inexistent. Moreover, the Dubov system, by its nature, requires very reliable ratings, which many Olympiad teams have not.

SPP Commission also cannot recommend the use of an accelerated system, particularly in view of the negative reactions caused by Dresden Olympiad pairings and of the still insufficient experience with such systems in team competitions.