MARIO HELD, IA

SWISS DUBOV AND FIDE SWISS (DUTCH)

A COMPARISON BETWEEN SWISS PAIRING SYSTEMS

ABSTRACT

FIDE Swiss (Dutch) and Swiss Dubov pairing systems are compared by means of extensive simulations (15000 Dubov tournaments. 15000 FIDE (Dutch) and 15000 double round robin), with the aim to verify the reliability of the obtained standings and whether Dubov system can actually achieve its goal of equalising ARO in scoregroups. To highlight the differences between the two systems, we analysed ARO in scoregroups, and players' positions in final standings, comparing the Swiss systems between themselves and with round robin.

Players' distribution in scoregroups is similar, but for a slight tendency of FIDE Swiss towards a sharper selection of top and bottom ranks, and a slightly more difficult path for top players. Dubov system partially equalises ARO in middle standings, but the effect is smaller for top and bottom ranks, where FIDE Swiss gives similar or better results. This seems to imply that the tournament winner meets a stronger and more uniform opposition with FIDE Swiss than with Dubov system. Standings linearity seems to be comparable – however, standings created by FIDE Swiss show a smaller deviation from the ideal and a slightly smaller uncertainty. The podia composed by the two systems are fairly similar, with only small and fluctuating differences.

In conclusion, the goal of ARO equalisation is at least partially achieved. However, this does not seem to yield "higher fidelity" standings and therefore fairer results for the tournament.

GENERALITIES

The strategic goal of the Dubov Swiss system is that of equalising players' ARO in each scoregroup, with the aim to make the encountered opposition as uniform as possible for players with the same score. In the philosophy of the system, this ensures fair pairings.

In the present paper we compared the behaviours of FIDE Swiss (Dutch) and Dubov pairing systems; and tried to verify whether the latter achieves its own strategic goal better than the far more well-known and widespread FIDE Swiss (Dutch) system.

The tournament as a measurement process

The primary goal of a tournament is to decide who, among a set of players, is the strongest; and, subordinately, to classify all players, or at least a part of them, based on their respective playing strength. The tournament can thus be construed as a measurement process for the playing strength of players.

It is however a peculiar process, in that we have no "example player" to be used as a unit of measure. In fact, our examples are the measured players themselves, in an iteration process working per successive comparisons. At each iteration (i.e., round), each player is compared to another one, whose playing strength is only approximately known – by means of the rating and the results achieved up to that moment. The outcome of each comparison is quantised in only three values (win, loss, draw) and brings a consequently small quantity of information, no more than approximately 1,6 bit[1].

This process, like any measurement, is characterized by some key parameters:

- Resolution (or the ability to reveal small differences), which depends on the difference in playing strength between paired players; and on the number of measurement cycles (rounds). For example, if player B lost to A and won against C, we can say that B's playing strength is intermediate between A's and C's. If A is very strong, and C is very weak, the uncertainty interval is wide, while if A and C have similar strengths, the uncertainty if definitely narrower

- Systematic error (depending on instruments and methods), due in part to playing conditions (which, for how hard we can try, we cannot make perfect, and may disturb players in different fashion); and in part to phenomena that are intrinsic to the game (for example, playing with black rather than white pieces is a well-known and partially quantifiable disadvantage[2]). For a round robin tournament, every player meets every other one once or twice. In Swiss type tournaments, each player usually meets only a small part of all the possible opponents. In this case, the systematic error also contains a component that depends on how the system chooses opponents

- Stochastic error (i.e., random), due to the many unforeseeable fluctuations that can make the needle of the scales lean to one side or the other: a small incident, a slight discomfort, a moment's distraction…

In simple measurement processes, the measured quantity (weight, length …) is constant – it does not change during the measurement. We can therefore repeat the measurement over and over again and, by averaging results, attenuate the random error and, up to a point, simulate a better resolution (while the systematic error can be kept in check with calibration and adjustment procedures).

---

[1] *The quantity of information ("informational entropy") is maximum when all outcomes are equally probable, and decreases as one result becomes much more or much less probable than the other(s). This is actually a platitude: the more a game result is foreseeable, the smaller the quantity of information is. Hence, the first rounds bring little information (we can estimate an average well under 1 bit for first round(s), and sometimes as low as 0,75 bit/game), while the last rounds usually bring nearly the maximum possible information – although this is not always true of the very last round! (By the way, Accelerated Swiss systems work on the principle of substituting low-information rounds in order to enhance the total information obtained from the games.)*

[2] *See Milvang, Otto - Probability for the outcome of a chess game based on rating, SPP report 2016.*

Our case is a bit more complex, because a player's playing strength is not constant, not even on an average: it varies slowly with age and education; a little more rapidly with condition cycles; and, finally, in a fast way with physical and mental state – so fast that it can be even fairly different from game to game. Moreover, players' playing strengths are not uncorrelated, because several players (especially amateurs) are somewhat intimidated by the opponent.

No tournament, not even a round robin, can therefore really say who the strongest player is: to know that, we would need to play an infinite number (i.e., many…) of tournaments – and even doing so we could not get the "right" answer, simply because the playing strength is never the same.

The conclusion is that we cannot expect a tournament to tell us which player is the strongest; at best, it can tell us, with reasonable reliability, which player *played best in that particular event*.

From measure theory we know that, the larger is the number of the measures done, the better is the approximation of the result. In our case, each measure is a game – hence, we expect a result that is the more reliable, the larger the number of rounds is. For example, in a knock-out tournament, where the number of games per player is minimum, chance plays an important role – notwithstanding all the corrective method we can apply (e.g., top seed), though luck may sometimes knock out a definitely strong contestant.

On the contrary, a double round robin tournament, in which each player meets each opponent twice (once with white and once with black pieces), has the maximum possible number of rounds (if N is the number of players, the tournament comprises N*(N-1) games). We therefore expect it to minimise the random error (and, by the way, also that part of the systematic error due to colour assignment).

Somewhere between those extremes, we find Swiss tournaments, in which the number of games per contestant is fixed – and the same for all (except in case of forfeits…). However, different Swiss systems pair players with different methods – originating by different assumptions about the best way to obtain fairness:

- the FIDE Swiss (Dutch) system implicitly assumes that the best way to select players is to equalise, as far as possible, the differences in playing strength for each pair of the scoregroup

- the Dubov Swiss system explicitly assumes that the best way to select players is to equalise, as far as possible, the average rating of opponents (AROs) of the players in each scoregroup

Those are "philosophical choices", which we should not discuss; we shall rather investigate on the actual ability of the Dubov Swiss system to achieve its strategic goal of equalisation, and on the achievements of both the systems in terms of standings composition.

The simulations

First, I wish to thank Mr Roberto Ricca, former Secretary of the FIDE SPP Commission, currently a member of the FIDE Technical Commission – and one of best pairing experts in the world – who prepared all the simulated tournaments. In particular, the Simulations for Dubov 2019 were prepared by his new Dubov pairing engine included in JaVaFo and soon to be released. A similar analysis will also be run for the Burstein system – as soon as its rules (see FIDE Handbook, Section C.04. 4.2) are consistently defined as per the indications given by the FIDE SPP Endorsing Subcommittee report[3], and a pairing engine is available.

Our goal is to compare the behaviours of the two Swiss systems, in order to shed some light on the respective ability to create a "fair path" for the players, and "fair standings" that best denote the real playing strength of contestants.

[3] *See the Meeting Minutes of Systems of Pairings and Programs Commission for the 89th FIDE Congress, held in Batumi, Georgia, in 2018. The Minutes are available on the FIDE website: https://spp.fide.com/wp-content/uploads/2020/04/2018_minutes.pdf.*

In practice, we cannot simulate all the possible types of tournaments encountered in real life, hence we should limit the scope of our examination. We therefore chose to create three sets of samples, representative of some typical situations:

- Group "A": a typical "Masters" tournament with comparatively few players, from high-level amateurs, or Candidate Masters, up to Grand Masters

- Group "B": an integral "Open" with players of every level, from lowly amateurs up to Grand Masters

- Group "C": another "Masters" tournament but larger – such as we may have in a large event

We used random tournaments generators to create several samples of the different types of tournaments, as shown in the following table.

Group	Sample	System	Rating	Number of players	Number of rounds	Number of tournaments
A "Masters"	F68	FIDE Swiss (Dutch)	2150-2615	68	9	5000
	D68	Dubov	2150-2615	68	9	5000
	R68	Double RR	2150-2615	68	(2x67)	5000
B "Open"	F140	FIDE Swiss (Dutch)	1416-2599	140	9	5000
	D140	Dubov	1416-2599	140	9	5000
	R140	Double RR	1416-2599	140	(2x139)	5000
C "Masters" (Large)	F142	FIDE Swiss (Dutch)	2150-2615	142	9	5000
	D142	Dubov	2150-2615	142	9	5000
	R142	Double RR	2150-2615	142	(2x141)	5000

The ratings used for the simulations were partly coming from actually played tournaments and partly generated locally, in such a way as to give them a realistic distribution. All tournaments have nine rounds - a typical length of a tournament valid for international titles.

All the tournaments generators assigned win, draw, or loss results in a random fashion, based on the probabilities determined by players' ratings and colours[4], in an attempt to simulate realistic tournament outcomes.

In such simulations we know right from the beginning the "correct" standings. The outcomes of the tournaments statistically depend on the playing strengths assigned to players by means of their respective ratings, which are assumed to be constant. Sorting the players by descending ratings (as normal), we expect an ideal measurement to yield standings exactly related to the initial ranking list of players. Any deviation from this list is a (random) deviation of players' behaviour from the expected playing strength given by their ratings – analogous to those that happen in real tournaments. We will therefore be able to verify if, and how much, the different pairing systems yield results that, on an average, are different from the expected ones, thus obtaining an approximated but statistically reliable evaluation.

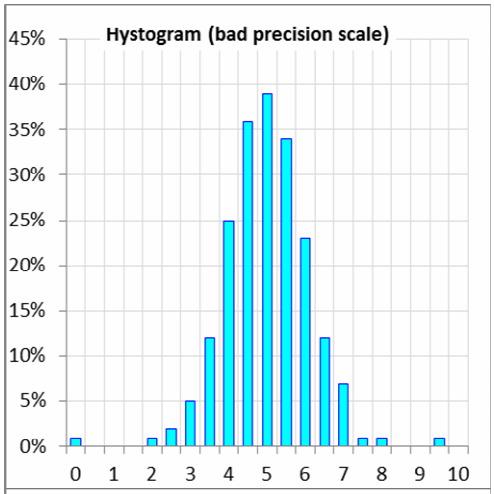
The standings coming from different pairing systems are the compared to each other and to the initial ranking list (ideally expected outcome), based on the average final position, the average final score, and the standard deviation of the position. Where needed, the final position in standings for Swiss systems was

[4] *See note 2, page 2.*

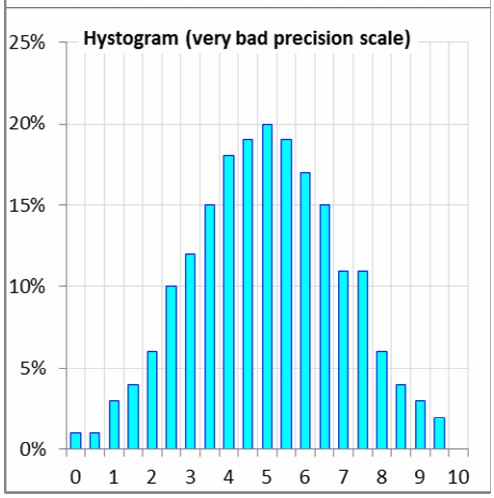
determined by Buchholz Cut-1 and APRO. In previous simulations, this combination proved to yield best results for the composition of the top standings, which are of prime concern.

The standard deviation

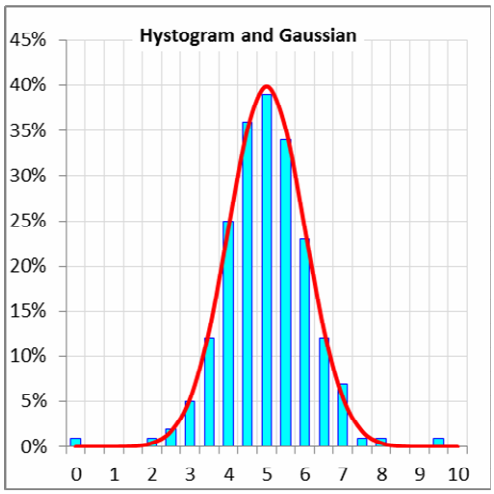
In this paper we extensively use the standard deviation, which is among most common statistical indicators. Before getting into samples analysis, we want to give some hints about the meaning of this quantity (the reader will find more information in a basic manual about statistics, or measure theory, or even physics). For simplicity, in this paragraph we always refer to the measurement of the weight of a generic object - however, what follows is equally true for all kinds of measurements.



If we repeatedly put an object on the weighing plate of scales that have enough resolution, (almost) each time we read a different weight. By averaging all such results (i.e., adding all them together and then dividing the sum by the number of results), we obtain their *mean*, a value that is the nearer to the real value, the better the scales are (of course!) and the larger the number of measures we took is.



The mean by itself does not tell us how precise our measurement is. For example, suppose we weighed a given object many times, using two different scales, in both cases obtaining a mean value of about 5 kg. However, on the first scales, all the readings were randomly distributed between 0 and 10 kg; while, using the second scales, the readings were randomly distributed between 4.5 ÷ 5.5 kg. Intuition tells us at once that the uncertainty of the measure is far larger in the first instance than in the second, but statistics gives us theoretical tools to *quantitatively evaluate* the precision obtained in those two cases.



Among the simpler of such tools there is the standard deviation, which is easily computed: first, we average all the measures, thus obtaining the mean; we then compute the differences between each measure reading and the mean and square them; last, we average those squares. We thus obtain a number that is called the *variance* of the given data[5]. The standard deviation is (by definition) the *square root of variance*, and can be seen as a measure of the "average distance" between the given data and their mean. The standard deviation is usually indicated by **σ** (lowercase Greek letter "sigma").

Let's go back one step, and take another look at the measures: first we group and sort them in "classes", based on their value (e.g., between 0-0.5, 0.5-1, 1-1.5 and so on). Then, we count the number of readings in each class. Lastly, we draw a diagram in which each class, in order, is represented by a rectangle whose height is proportional to the number of "events" (i.e., measurements readings) that belong to it.

We thus obtain a "histogram" – a most common and useful kind of diagram. In the illustration (see side, above), the histogram above shows better (more precise) measures than the one below (the measurement

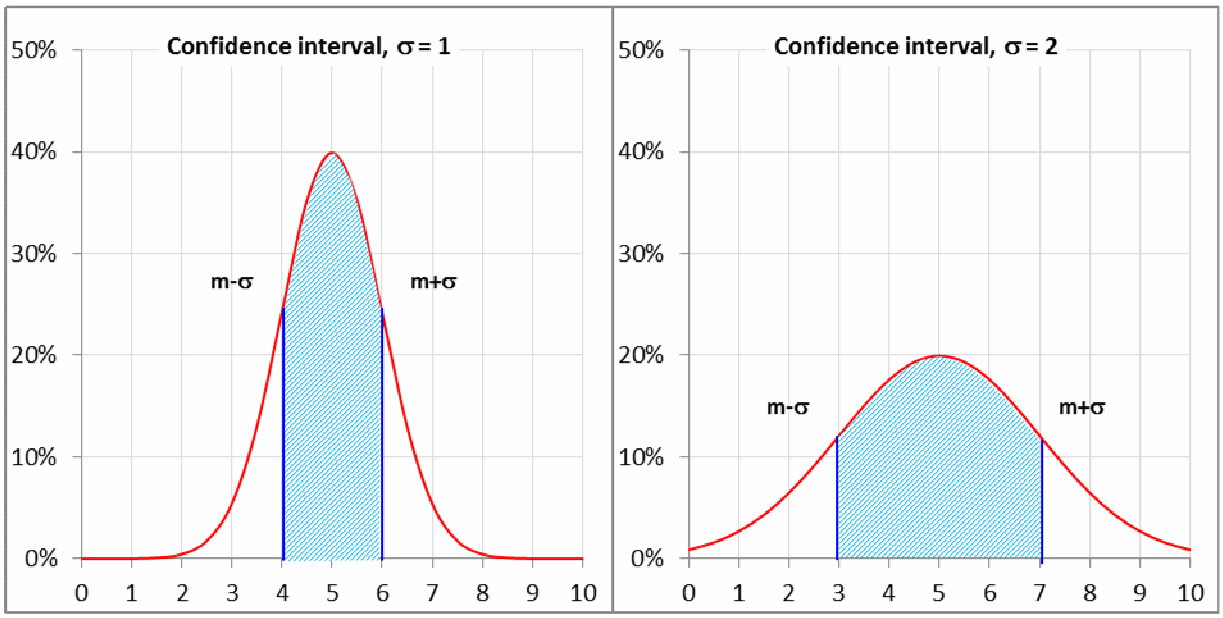
[5] *There is another quantity, the "sample variance", which is very similar to this, but not just the same. We will not use it.*

error in both histograms has been made very large, to make it evident). In those histograms we can observe that counts become larger and larger as we approach the mean value – on the contrary, if scales are good, as we move away from the mean they rapidly become fewer and fewer. If scales are not good, the measures are much sparser – however, the sum of counted events is a constant, because it is simply the total number of readings. Hence, the wider the distribution is, the squatter it becomes.

The most interesting observation is that the distribution has always (approximately) the same shape, corresponding to a special curve, called a "Gaussian" (see picture to the side). When we meddle with random phenomena, we find this curve so often, that it is called "normal distribution"[6]. This however does not mean that *all* phenomena are Gaussian, and we need always be very cautious in using properties valid only for Gaussian phenomena in unknown situations – that is, situations that may well obey to completely different laws!

Gaussian distributions have a very useful property: if ***m*** is their data mean and ***σ*** is their standard deviation, approximately[7] a 63% of the readings belong to the interval **m±σ** (that is, it is contained between ***m-σ*** and ***m+σ***). We call such a range a *confidence interval* – this is a very important practical parameter to evaluate the quality of the measure. Moreover, approximately 99% of the readings belong to the interval **m±3σ** (i.e., from ***m-3σ*** to ***m+3σ***) – in practice, all data fall in this range.

All this should be interpreted in a statistical way and not as a certainty – sometimes we may find many readings outside the interval ***m±3σ*** , while some other times we may find none… *statistically*, however, that is *on average*, we expect to find more or less 99% of the readings inside this interval, and more or less 1% outside. The picture below shows the confidence intervals ±σ in two Gaussian distributions, both having mean value of 5 and standard deviation respectively of 1 (left) and 2 (right).



In conclusion, we can use standard deviation as a reliable indicator of the quality of the measure, the latter being the more precise, the smaller the deviation is; and define *confidence intervals* for which we know the probability, i.e. the average percentage of measures belonging to the interval.

Now, the quantity we want to measure is the players' playing strength, and the variable that interests us most is therefore each player's position in the final standings. However, for a more complete comparison

[6] *In fact, Bernoulli's "Central Limit" theorem, a very important theorem related to the Law of Large Numbers, shows that every time a large number of stochastic events is involved, the larger that number is, the more the distribution resembles a Gaussian.*

[7] *Percentages are specific to each probability distribution. Those given here are only valid for Gaussian phenomena.*

between Dubov and FIDE Swiss systems, we should first examine some other aspects of the pairings, and in particular players' ARO distribution, which is the optimisation goal of the Dubov system.

All parameters have been evaluated after the last round – that is, at the end of the tournament.

POPULATION OF SCOREGROUPS

Scoregroups composition

In an ideal tournament (without draws), all the possible scoregroups are created, but their populations are very different, with many players in central scoregroups, and fewer and fewer as we move away from the average score. Draws reduce the number of players having either maximum or minimum score, because they decrease stronger players' scores and increase weaker players', thus inflating central scoregroups. This detracts from the population of extreme scoregroups, sometimes even completely emptying them, and therefore making them disappear. Because of this, we first examined the distribution of scoregroups, which shows the effectiveness of the tournament in selecting best (and worst) players. The histograms below show the *frequency* of the scoregroups – that is, how many times the scoregroup was created in the 5000 tournaments of each sample set (i.e., there were players in the tournament who ended up with that score).

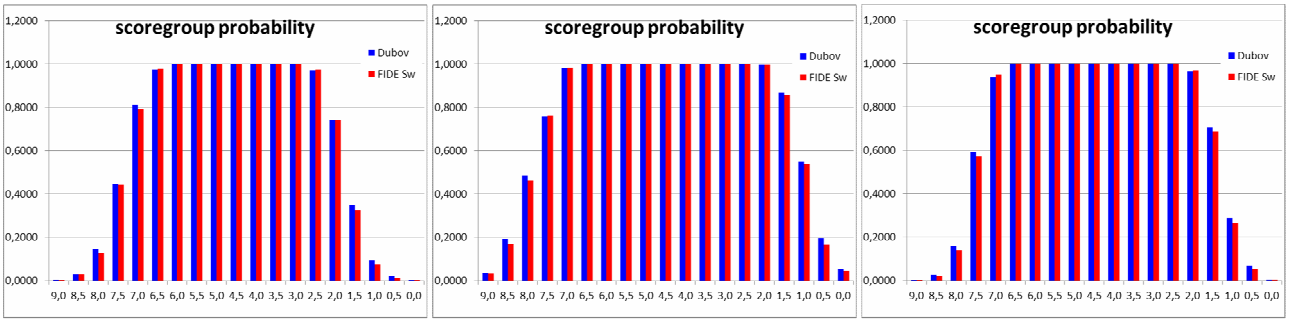


Figure 1: Probability of scoregroups formation (left: Group A; middle: Group B; right: Group C)

The central scoregroups were of course created in all tournaments. The probability of scoregroup formation decreases as we move away from the average score, and 'extreme' (outmost) scoregroups are definitely rare (see table and graphs below).

score group	Group A		Group B		Group C	
	Dubov	FIDE Swiss	Dubov	FIDE Swiss	Dubov	FIDE Swiss
9	0,5%	0,2%	3,6%	3,4%	0,1%	0,1%
8,5	2,9%	2,8%	19,2%	17,0%	2,7%	2,2%
8	14,5%	12,7%	48,6%	46,1%	16,2%	14,0%
7,5	44,6%	44,3%	75,8%	76,2%	59,3%	57,5%
7	81,1%	79,2%	98,1%	98,4%	93,7%	94,9%

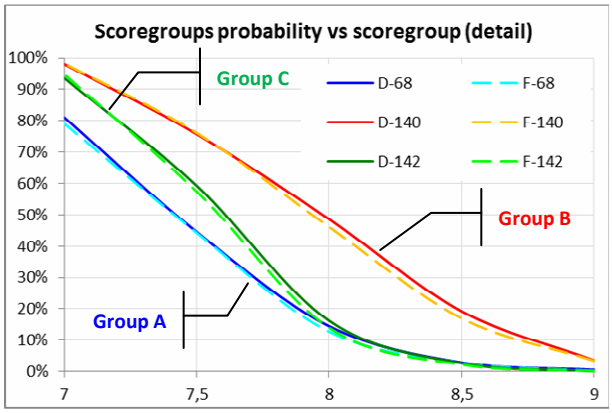


Figure 2: Probability of formation of top scoregroups

For example, the 7.5 points scoregroup in Group A was created only in approximately 45% of the tournaments, while probability gets below 15% for the 8 points one. In Group B the probability of extreme scoregroups decrease definitely slower (there are many players with very sparse ratings). Group C, which contains many players but less sparse ratings, shows a 'halfway' behaviour, but resembles Group A in its rapid decrease of extreme scoregroups probability. This happens because, when ratings are less sparse (Groups A and C) the number of draws increases, thus deflating extreme scoregroups.

The FIDE Swiss system always shows a slightly lower probability of extreme scoregroups creation than Dubov. FIDE Swiss therefore makes a (slightly) stricter selection of the players forming the top standings – however, the difference is definitely small.

Another very important parameter for the composition of standings is the average number of players contained in each scoregroup, which tells us whether tiebreak criteria are needed to decide players' positions in standings. We therefore examined the average distribution of the number of players per scoregroup, in the 5000 tournaments of each sample set. This too is very similar between the two systems – however, Dubov puts some more players in top and bottom scoregroups than FIDE Swiss.

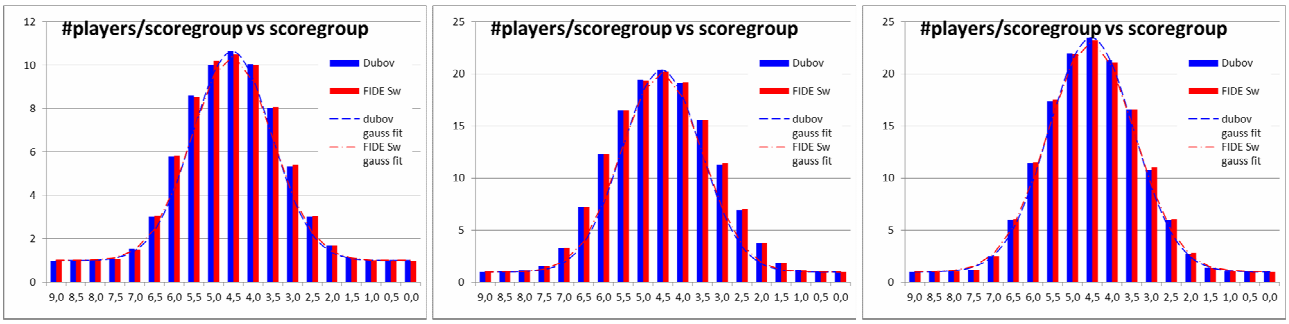


Figure 2: Average number of players by scoregroups. (from left to right: Group A; B; C. Please note that scales are different.)

The difference is once again very little, as indicated by the comparison of the distributions of the total number of players per scoregroup (as before, on 5000 tournaments).

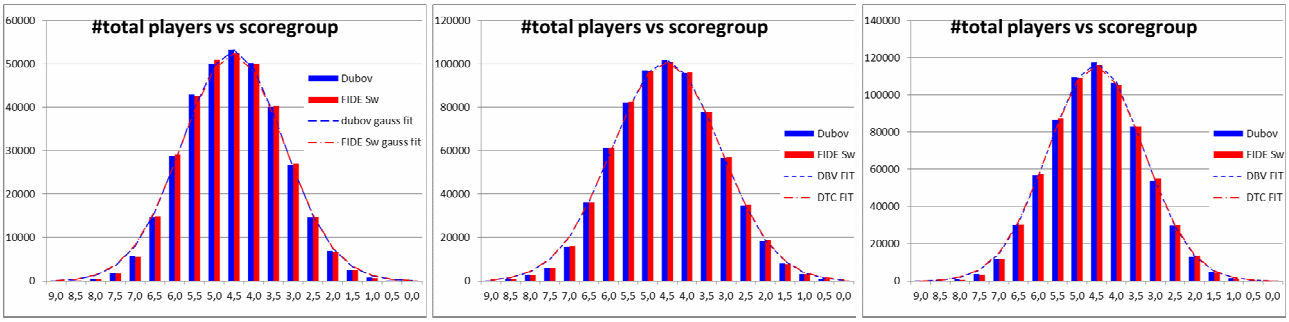


Figure 3: Total players vs. scoregroups. (from left to right: Group A; B; C. Please note that scales are different.)

To summarize, the formation of scoregroups shows only marginal differences between the two systems, with only a slight tendency of FIDE Swiss to select top and bottom standings stricter than Dubov does.

ARO Comparison

The Dubov system, in its rules preface, establishes as its goal the fair treatment of players, meaning that the path to victory should be equally difficult for all players who finish the tournament with the same score. The difficulty of this path is evaluated based on the supposed playing strength of the opponents, which is estimated by ratings. The reference value for each player is therefore the ARO (Average Rating of Opponents), which is the mean of ratings of all the *actually* encountered opponents –unplayed games are excluded from the computation. The strategic goal of the system is therefore to equalise the ARO for all players belonging to any given scoregroup.

To evaluate whether this goal is achieved, we compared the distribution of AROs and of respective variations in scoregroups. The first scrutinized parameter is the mean of AROs on 5000 tournaments, versus scoregroup (see graphs below).

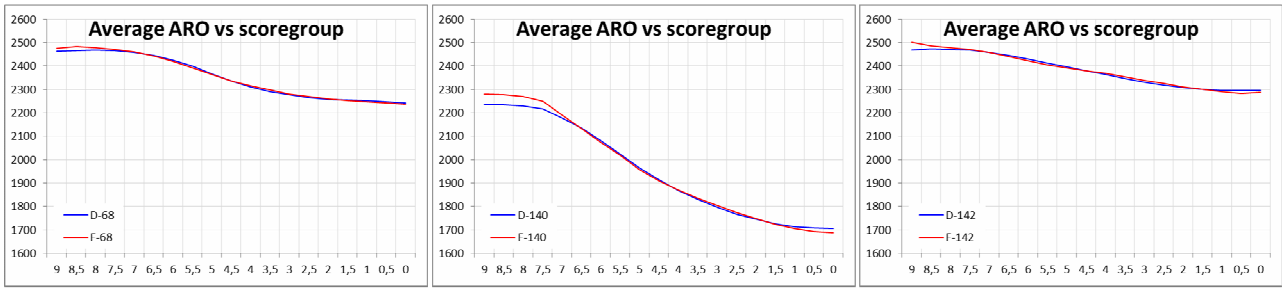


Figure 4: ARO vs scoregroup (from left to right: Group A; B; C.).

The evolution of ARO can be subdivided in three zones:

- a top standings zone, where the curve is almost constant versus scoregroups – players in those scoregroups encountered an approximately equal opposition
- a wide central zone, where the encountered opposition is essentially proportional to the final result. Players in those scoregroups encountered an opposition that was as tough, as good their final result was
- a bottom standings zone, where the encountered opposition is once again approximately the same for all players

We should observe that in central scoregroups the average encountered opposition per player is essentially the same between the two systems – hence, to a majority of players, the two systems seem fairly equivalent, as in practice there are (small) differences only in top and bottom standings. To give a sharper vision of this behaviour, we depicted the differences between AROs mean values in FIDE Swiss and Dubov systems (graphs below).

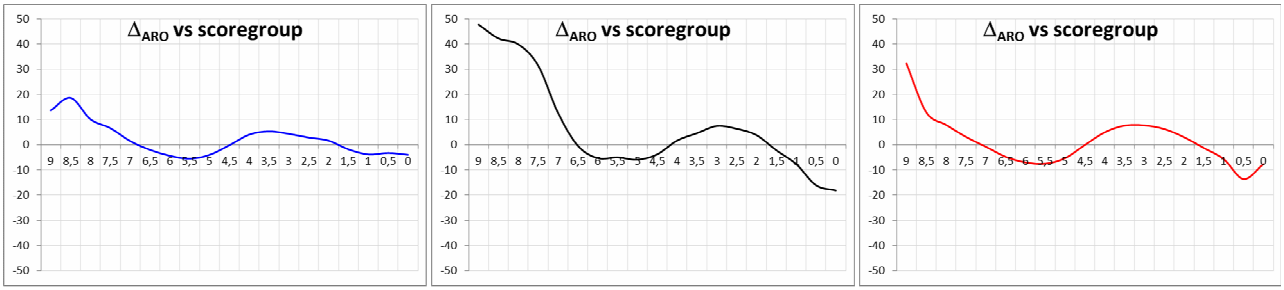


Figure 5: Difference between average AROs obtained by FIDE Swiss (Dutch) and Dubov systems. (from left to right: Group A; B; C.)

In top scoregroups, the average encountered opposition is smaller for the Dubov system (≈15 Elo points in Group A, ≈50 points in B, ≈30 points in C). In bottom scoregroups we find just the opposite, and the average encountered opposition is (slightly) higher for the Dubov system. (The difference is more evident in Group B, because of the very wide rating range.)

This pretty much means that the path to victory is just a little easier with Dubov than with FIDE Swiss. This is consistent with the previous observation that FIDE Swiss makes a stricter selection of the standings top and bottom.

All this however still tells us nothing about the success in equalising oppositions - even if means are equal, AROs could still be almost equal or very different. To evaluate this aspect of the question, we computed the standard deviation σ_{ARO} of AROs in each scoregroup (for the full sample of 5000 tournaments), and depicted its evolution (see graphs below).

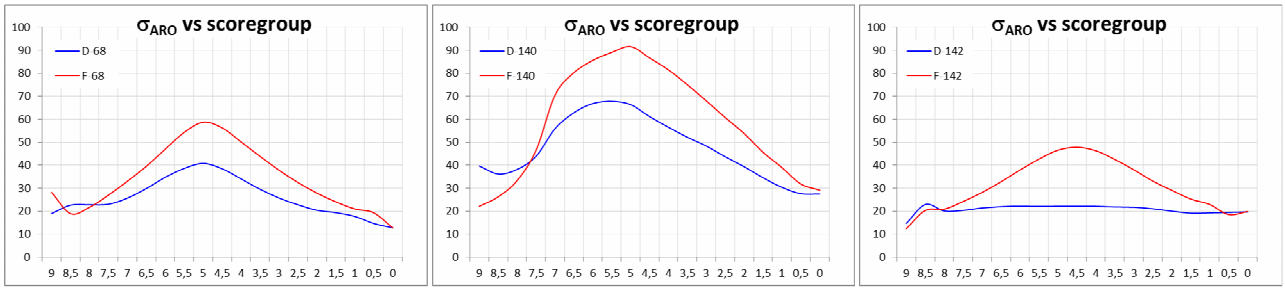


Figure 6: Spread of AROs versus scoregroups (standard deviation in Elo points) for FIDE Swiss (Dutch) and Dubov systems. (from left to right: Group A; B; C.)

The deviation is at its maximum for Group B, where ratings are distributed in a very wide range. In all samples, it is at its maximum for central scoregroups, and decreases moving to extreme scoregroups – where there are only few players, and of fairly similar playing strength.

The deviation in FIDE Swiss is usually larger than in Dubov for central scoregroups – while it is comparable or even smaller for extreme scoregroups, and especially for top ones (see graphs below). Data show that, in central scoregroup, ARO variability is smaller (approximately 20-30 Elo points) for Dubov than for FIDE Swiss – however, this advantage becomes smaller and smaller, and even vanishing, as we move towards top standings positions.

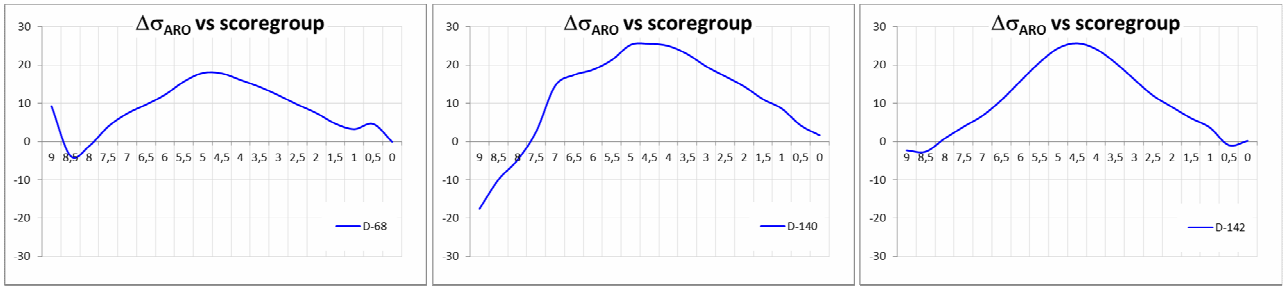


Figure 7: Difference in the spread of AROs between FIDE Swiss (Dutch) and Dubov systems (from left to right: Group A; B; C.).

For example, for the "large Open" format tournaments (Group B), the average standard deviation of ARO in central scoregroups is about 70 Elo points for the Dubov system and about 90 points for FIDE Swiss. Hence, in case of Dubov tournaments, the AROs of approximately 63% of the players belong to an interval around ±70 Elo points, while for FIDE Swiss this interval is around ±90 points. Group A shows a similar difference, while there is a clearly bigger advantage for the "large Master" formula (Group C) – where, by the way, with the Dubov system we also observe an almost uniform dispersion of AROs versus scoregroups.

To better understand this phenomenon, we also examined the standard deviation of ARO at single tournament level, depicting its distribution in some representative scoregroups (see graphs below).

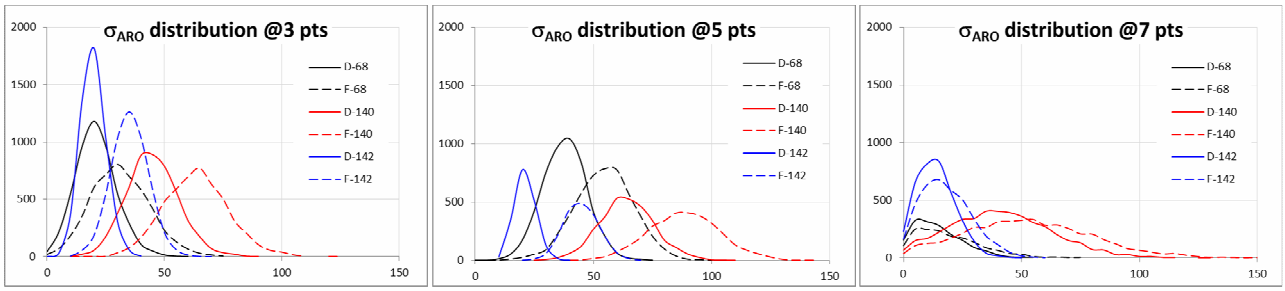


Figure 8: Spread of AROs in tournaments; x: deviation σ (Elo points); y: frequency (number of tournaments showing deviation σ)

For central scoregroups (5 pts.), the distributions seem to resemble Gaussians – especially for Groups A and C, where rating ranges (and hence ARO ranges) are definitely narrower.

On the contrary, in outer scoregroups, the small number of players and a limitation intrinsic to scores (which can neither become less than zero, nor more than 9) both concur to distort the symmetry and shape of the distribution.

In summary, in top scoregroups the simulations show an equalisation of AROs that is comparable or better for the FIDE Swiss system – which also exhibits a slightly higher ARO. This seems to indicate that the winner encounters a superior and more uniform opposition with respect to that given by the Dubov system. On the contrary, the equalisation of ARO given by the Dubov system in central scoregroups is clearly better – but not excessively so.

Rating distribution

The rating distribution in scoregroups is also an interesting element in the comparison between the two systems (see graphs below). It gives a simple indication of the global effectiveness of the pairing systems in measuring and classifying the playing strength of contestants.

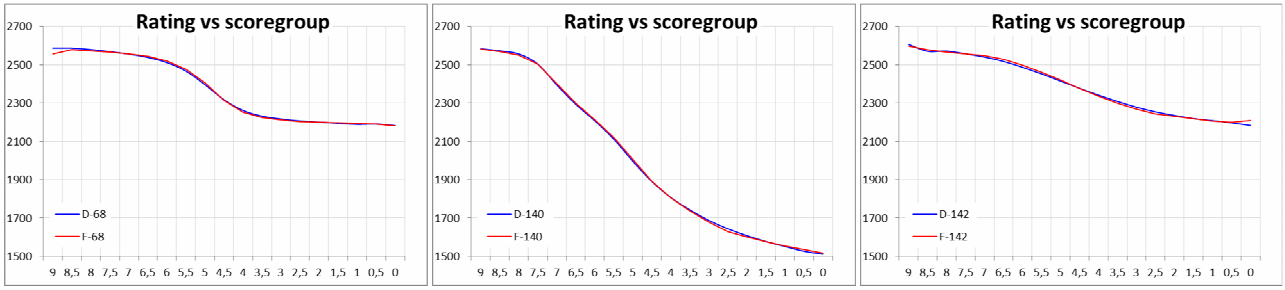


Figure 9: Average ratings in scoregroups obtained by FIDE Swiss (Dutch) and Dubov systems (from left to right: Group A; B; C.).

Once again, the outcomes of the two systems are almost equal, and evolutions are quite similar to those of ARO: in central scoregroups the distribution is fairly linear, while it tends to flatten for outer scoregroups. This similarity is not at all unexpected, because in the course of the tournament, stronger players tend to gradually move towards top standings, and to play more and more with their equals, thus strengthening each other's encountered opposition. Graphs below show the difference in average rating versus scoregroups.

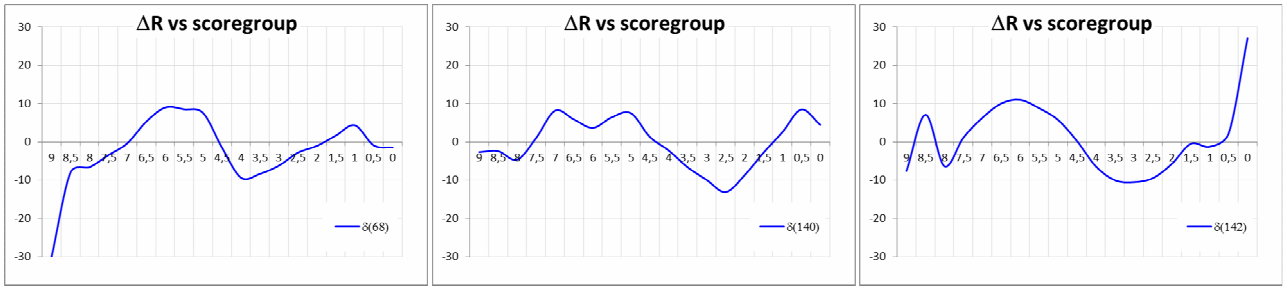


Figure 10: Difference between average ratings obtained by FIDE Swiss (Dutch) and Dubov systems (from left to right: Group A; B; C.).

In the top (9 points) scoregroup – which exists only in 0.1÷0.2% of the tournaments – the average rating for FIDE Swiss is approximately 30 Elo points less than for Dubov. In all other scoregroups, the difference between the two systems is inside ±10 Elo points, and in practice can be ignored.

STANDINGS

Linearity of standings

As mentioned above, the ideal tournament should classify players based on their playing strength. Since in our simulations (contrary to real tournaments!) the playing strength is by definition *exactly* given by the player's rating, an ideal tournament should yield standings that are identical to the initial ranking list. In this paragraph we compare the standings made by Dubov and FIDE Swiss systems to each other and to the ideal standings – that is, the initial ranking list. For a better understanding of these results, the comparison also included the standings obtained by double-round robin pairings with the same sample sets of players. In order to ensure the statistical equivalence of the sample sets, the results for the games were generated based on the expected outcome of the game, as mentioned before[8], both for Swiss and round robin tournaments.

In each sample set we examined the average position of players in the final standings as a function of the initial ranking (see graphs below, left), and the respective deviation (right) from the ideal evolution (the latter being a perfectly linear correspondence between pairing id and position in standings).

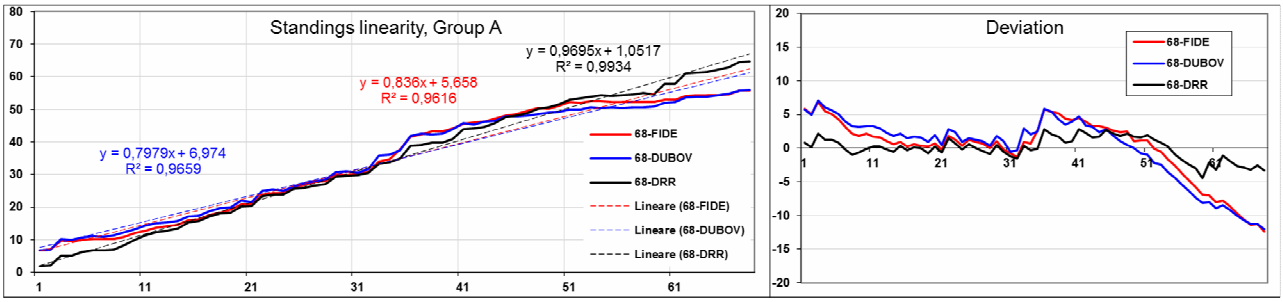


Figure 11: Standings linearity (Group A). Left: correlation between initial ranking (x) and final standing (y). Right: Deviation of final standing from intial ranking.

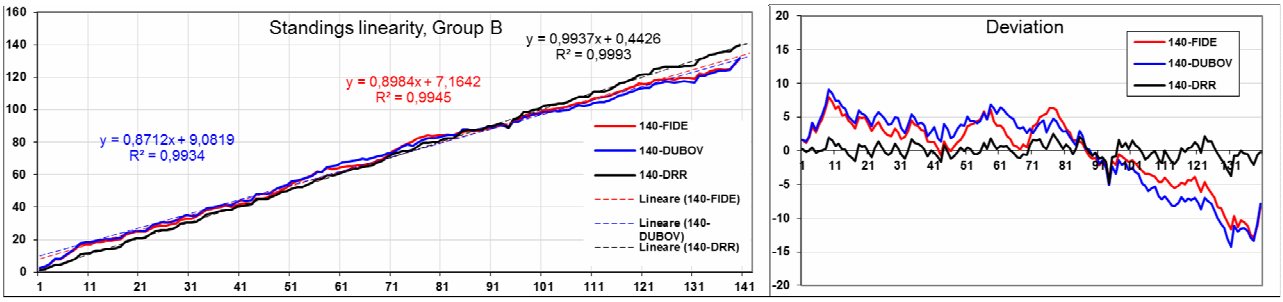


Figure 12: Standings linearity (Group B). Left: correlation between initial ranking (x) and final standing (y). Right: Deviation of final standing from intial ranking.

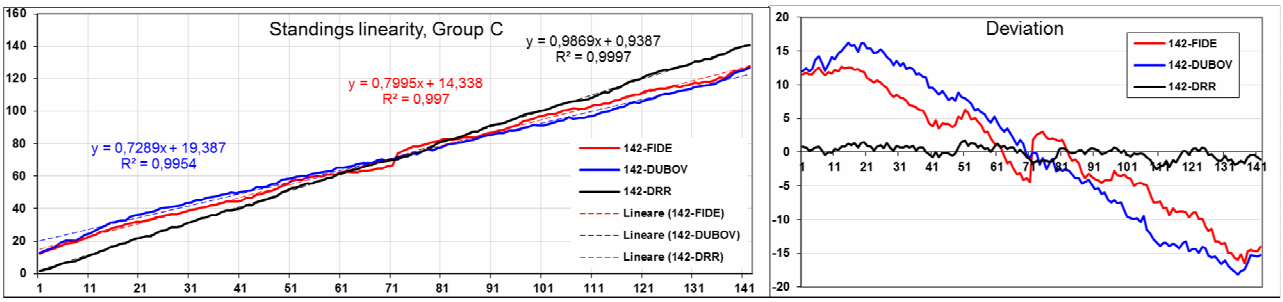


Figure 13: Standings linearity (Group C). Left: correlation between initial ranking (x) and final standing (y). Right: Deviation of final standing from intial ranking.

[8] *See "*The simulations*", page 3.*

In all sample sets we found a remarkable correspondence between initial ranking and final position. Only in outmost scoregroups there is a mild divergence from linearity, caused by the limiting effects of the edges[9]. This causes the observed partial loss of linearity, which is unavoidable. This effect is just barely visible in the standings obtained from round robin tournaments, (although it can still be seen), presumably because of the *very* large number of rounds.

The deviation from ideal (see graphs above, right side) shows this distortion in more detail – it is therefore mainly positive for the upper part of standings, and negative for the lower part. The above systematic error, which is intrinsic in all systems, is more evident in Swiss systems because of the smaller number of rounds.

Its standard deviation gives an immediate estimate of the uncertainty in the yielded standings – comparing the obtained values is rather interesting (see table below).

	Group A (Masters)			Group B (Open)			Group C (large Masters)		
	FIDE Swiss	DUBOV	DRR	FIDE Swiss	DUBOV	DRR	FIDE Swiss	DUBOV	DRR
σ_{DEV}	4,6	5,0	1,7	4,9	6,0	1,1	8,3	11,3	0,9

For example, with the Dubov system, player #10 in Group A obtains, an average position equal to 13.2, and a ≈63% probability of ending up in positions 8÷18; with the FIDE Swiss system the average positioning is 12.2 (that is, one standing position up) and the player has a ≈63% probability of ending up in positions 8÷17. However, in both cases, the player could well end up eight ranks – with a non-negligible probability of ≈18%.

In Group A (Master tournaments with comparatively few players), simulations results are similar for both systems. The difference is far more evident in Group B and especially in Group C.

The FIDE Swiss system shows a very peculiar 'swing' exactly at the centre of the standings. This is especially evident in Group C, between players #71 and #72. Those are respectively the last of the first half and the first of the second half of the initial ranking. On average, those two players are six positions apart from each other in final standings, while all other consecutive players are, on average, no more than two positions apart. Hence, deviation 'jumps' (this is also visible in scores – hence, it is not an artifact of tiebreaks). This behaviour, even if not always evident, is intrinsic to the FIDE Swiss system[10], and cannot be eliminated; however, as hinted above, standings deviation from ideal is globally smaller for FIDE Swiss than for Dubov.

The lack of linearity in the formation of standings poses of course the question if final positions are consistent with playing strengths; we can answer this question by means of the *correlation index*[11].

---

[9] *For example, the position of player #50 can be in a wide range of standings centred on the 50th and, in the average positioning, we obtain a position approximately nearing the initial ranking. Conversely, for a top (or bottom) player, the interval is necessarily unbalanced towards lower (or, respectively, higher) positions.*

[10] *The reason for this behaviour is that, in the pairing process, the centre of the bracket is a discontinuity point. Here we find an abrupt transition from the lowest players of the upper subgroup (who are paired to weaker opponents), to the upper players of the lower subgroup (who are paired with stronger opponents).*

[11] *The correlation index* $\boldsymbol{\rho}$ *(this is the lowercase Greek letter 'ro') between two or sometimes more variables is a quantity that varies in the interval ±1, where +1 and -1 mean a perfect linear dependency, while a null value means total uncorrelation. The latter often means that the interested variables are reciprocally independent – although there are examples of uncorrelated variables that <u>are</u> dependent (e.g.* $y=x^2$*). As a rule of thumb, the correlation is loose for* $|\rho| < 0.3$*, average for* $|\rho| < 0.7$ *and good for* $|\rho| \geq 0.7$

We therefore computed the correlation index between position in standings and initial ranking (i.e., ideal position) for all sample sets (see table below).

Correlation index	Group A (Masters)	Group B (Open)	Group C (large Masters)
Dubov	96.59%	99.34%	99.54%
FIDE Swiss	96.16%	99.45%	99.70%
Double RR	99.34%	99.93%	99.97%

All the results show an almost perfect correlation, thus indicating a very strict functional dependence between ideal and achieved positions. The correlation index is essentially equal to unity both for Dubov and FIDE Swiss systems, thus indicating that the standings obtained in both systems represent playing strengths very well and with similar reliability.

To scrutinize the matter in even some more depth, let's once again examine in detail the deviation of standings from ideal (see Figures 11, 12, 13) – here, we define deviation as the difference between achieved and ideal (initial ranking) positions. The analysis of such deviations shows that, with respect to the Dubov system, FIDE Swiss always yields standings that are slightly nearer to ideal – however, as expected, round robin gives the best (by far!) result. Since maximum (and most unwelcome) deviations show up in top scoregroups, let's focus briefly on top standings (see graphs below).

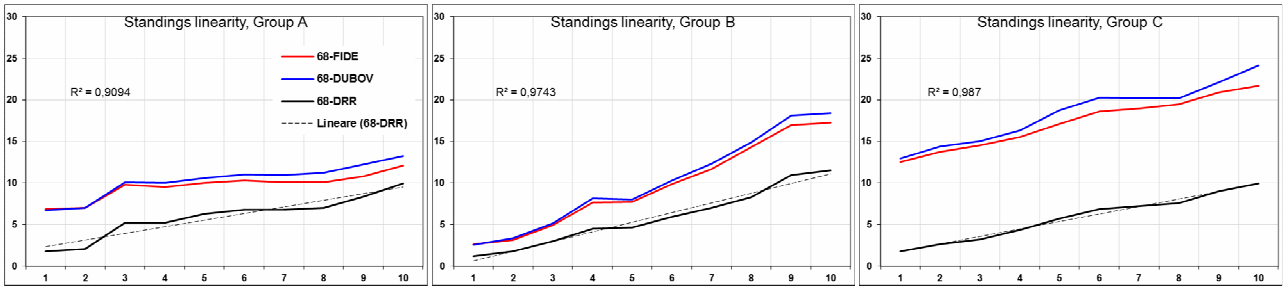


Figure 14: Detail of standings linearity in top scoregroups (from left to right: Group A; B; C.).

As before, we can see that the FIDE Swiss system Swiss yields standings that are slightly nearer to ideal than Dubov does; in Group B, which is distinguished by a far wider variability of ratings, both the Swiss systems show a clear tendency to assign the very first standings positions with better precision – in other words, *the presence of weak players helps the correct finding of the stronger*. (After all, this behaviour should not be unexpected.) In top standings of Group A, where the players are not many and the ratings are distributed in a narrower range, the correlation decreases noticeably. Even for round robin, it goes down to around 90%, which is still a very good value, but definitely much worse than the global results, which are over 99%.

Finally, we verified the correlation of the standings obtained from each system among themselves (see table below). The very high values show a strong reciprocal dependence of the results, and thus a significant agreement among standings.

Correlation index	Group A (Masters)	Group B (Open)	Group C (large Masters)
FIDE vs. DRR	99.16%	99.81%	99.85%
Dubov vs. DRR	99.29%	99.75%	99.81%
FIDE vs. Dubov	99.90%	99.94%	99.66%

We conclude the analysis of standings positions with a look to the standard deviation, which characterises the uncertainty of the position and therefore its typical error.

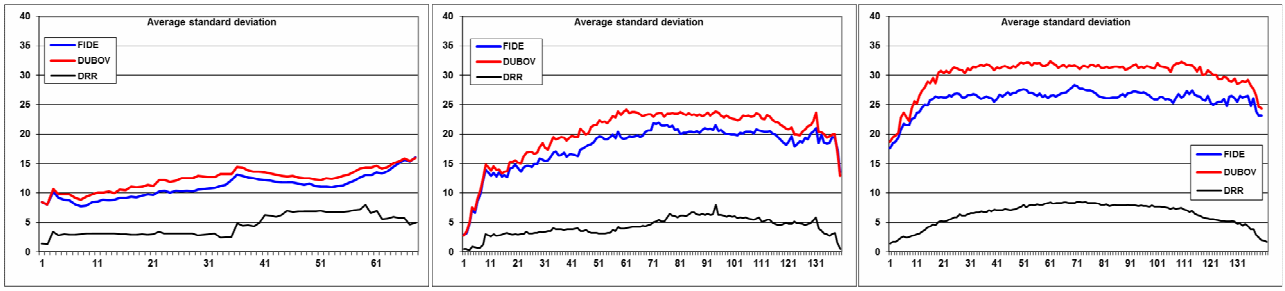


Figure 15: Average standings standard deviation (from left to right: Group A; B; C.).

Once again, we find that FIDE Swiss is more precise than Dubov in assigning the standings positions. We can also notice that even double-round robin shows a far for negligible standard deviation, which seems to decrease, but does not vanish, for top standings.

Podium selection

To examine how players are chosen for the podium, we studied the first ten players' standings. In order to be able to compare the scores yielded by round robin tournaments with those of Swiss tournaments, we normalised them to a maximum of 9 points.

Group B		FIDE Swiss				DUBOV				Double RR			
#	Elo	Avg Rank	Avg score	σ_{Rank}	estimated expected standing	Avg Rank	Avg score	σ_{Rank}	estimated expected standing	Avg Rank	Norm Score	σ_{Rank}	estimated expected standing
1	**2599**	2,64	7,46	2,89	1÷6	2,56	7,54	2,79	1÷5	1,23	8,58	0,49	1÷2
2	**2572**	3,19	7,32	3,17	1÷6	3,39	7,33	3,45	1÷7	1,82	8,51	0,51	1÷2
3	**2510**	4,97	7,00	4,85	1÷10	5,15	6,98	4,97	1÷10	3,01	8,30	0,35	3÷3
4	**2446**	7,65	6,68	7,17	1÷15	8,16	6,65	7,59	1÷16	4,52	8,04	0,88	4÷5
5	**2443**	7,73	6,68	6,64	1÷14	8,01	6,68	7,18	1÷15	4,68	8,03	0,80	4÷5
6	**2405**	9,86	6,52	8,53	1÷18	10,34	6,50	8,95	1÷19	5,96	7,84	0,72	5÷7
7	**2374**	11,68	6,41	9,75	2÷21	12,32	6,38	10,44	2÷23	6,99	7,68	0,68	6÷8
8	**2339**	14,35	6,26	11,93	2÷26	14,90	6,24	12,19	3÷27	8,32	7,48	1,19	7÷10
9	**2303**	16,99	6,14	13,88	3÷31	18,08	6,10	14,89	3÷33	10,97	7,25	3,09	8÷14
10	**2298**	17,23	6,13	13,56	4÷31	18,38	6,08	14,46	4÷33	11,51	7,22	2,88	9÷14

The above table shows the situation for Group B, which gave the best results among our simulations. The expected standings from Swiss systems vary in a very wide interval (computed as a ±σ confidence interval) – although it is slightly narrower for FIDE Swiss than for Dubov.

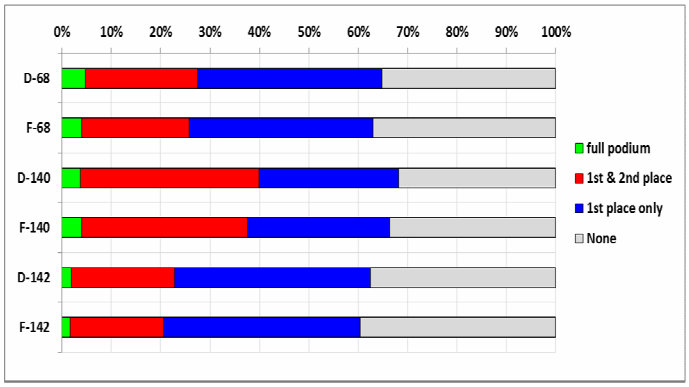
Of course, the double-round robin system yields a much smaller uncertainty – likely because of the huge number of rounds.

The worst situation is that found in Group C (see table below), showing a very wide confidence interval, in which the first players of the initial ranking can achieve definitely unsatisfactory final standings.

Group C		FIDE Swiss				DUBOV				Double RR			
#	Elo	Avg Rank	Avg score	σ_{Rank}	estimated expected standing	Avg Rank	Avg score	σ_{Rank}	estimated expected standing	Avg Rank	Norm Score	σ_{Rank}	estimated expected standing
1	**2615**	12,54	6,33	17,56	1÷30	12,97	6,35	18,67	1÷32	1,83	6,94	1,44	1÷3
2	**2605**	13,78	6,27	18,53	1÷32	14,41	6,26	19,56	1÷34	2,66	6,84	1,68	1÷4
3	**2600**	14,51	6,24	18,68	1÷33	15,06	6,23	19,75	1÷35	3,21	6,79	1,72	1÷5
4	**2590**	15,52	6,19	19,42	1÷35	16,32	6,16	20,20	1÷37	4,39	6,68	2,05	2÷6
5	**2580**	17,10	6,13	20,76	1÷38	18,73	6,06	22,72	1÷41	5,73	6,58	2,40	3÷8
6	**2572**	18,59	6,05	21,69	1÷40	20,24	6,00	23,63	1÷44	6,87	6,50	2,63	4÷10
7	**2570**	18,94	6,05	21,52	1÷40	20,22	5,99	22,98	1÷43	7,22	6,48	2,48	5÷10
8	**2568**	19,44	6,03	21,56	1÷41	20,17	5,99	22,26	1÷42	7,57	6,46	2,51	5÷10
9	**2559**	20,90	5,97	22,51	1÷43	22,12	5,93	24,27	1÷46	9,01	6,36	2,69	6÷12
10	**2553**	21,69	5,94	22,79	1÷44	24,12	5,85	25,57	1÷50	9,96	6,31	2,79	7÷13

To complete the comparison of the two systems, we want to estimate their ability to identify with certainty the assignees of the first three positions in standings (podium) – and in particular the tournament winner. Simulation data show the percentage of tournaments in which the standings could be drawn up without help from tiebreaks (see table and picture to the right). The first white column represents the percentage of tournaments in which the topmost scoregroup contained only one player, while the subsequent scoregroup contained several players – and thus the winner was univocally identified, but the runners-up were not. The second column represents the percentage of tournaments in which we could determine with certainty the first and second place; while the third column refers to tournaments in which all podium players were identified. The fourth column represents the percentage of tournaments in which tiebreaks were necessary event to determine the winner.

Group	Sample	1st place only	1st & 2nd place	full podium	None
A	D-68	37,4%	22,7%	4,7%	35,2%
	F-68	37,1%	21,8%	4,0%	37,0%
B	D-140	28,2%	36,3%	3,7%	31,9%
	F-140	28,8%	33,5%	4,0%	33,6%
C	D-142	39,7%	20,9%	1,9%	37,5%
	F-142	39,9%	18,9%	1,6%	39,6%



It is readily evident that the choice of tiebreakers is always very important, because it determines the podium composition, completely or in part, in almost all tournaments (in our simulations, more than 95%). The percentages, however widely different from samples set to samples set, are always fairly similar between Dubov and FIDE Swiss.

CONCLUSIONS

The composition of scoregroups, and the corresponding players' distribution, produced by Dubov and FIDE Swiss systems are essentially similar, with only a slight tendency of the latter to better select the outmost (top and bottom) scoregroups.

The evolution of ARO versus scoregroups is also comparable between the two systems, with a just mild tendency of FIDE Swiss to create a slightly more difficult path for winners.

The Dubov system shows a better – but still not perfect – equalisation of AROs in central (i.e., middle standings) scoregroups. However, this behaviour seems to vanish in top scoregroups, where FIDE Swiss gives comparable results – or even better, as is the case for Group B. This seems to show that tournament winners encounter a superior and more uniform opposition with FIDE Swiss than with Dubov.

The linearity of standings seems to be fairly similar between the two systems – however, standings yielded by FIDE Swiss show slightly smaller deviation from ideal and slightly less uncertainty.

The two systems yield similar results also in the composition of the podium – sometimes one system is marginally better, sometimes the other.

In conclusion, the goal of ARO equalisation is at least partly achieved. However, this seems not to imply a better reliability of the yielded standings, which on the contrary seem to be slightly better for FIDE Swiss. This seems to call into question the basic assumption of the Dubov system – that equalising ARO can ensure a fairer pathway or result in the tournament.