URS™: Universally Better Than Elo

The URS™ rating algorithm was designed and developed by our research team, which consists of Mr. Maxime Rischard, Dr. J. Isaac Miller, Dr. Mark Glickman, and Mr. Jeff Sonas. The work has been funded for the last two years through a collaborative research project funded by the Grand Chess Tour, the Kasparov Chess Foundation, and the Chess Club and Scholastic Center of Saint Louis.

There are many differences between the URS™ and the FIDE Elo systems. The most striking difference is that the URS™ calculates only one rating for each player, informed by their results at all rates of play from Classical to Blitz (5 minutes per game). This published rating is the URS™ system's assessment of each player's strength at Classical chess (defined as a rate of play where each player has at least 2 hours for their first 60 moves).

We expect some people to challenge the notion that games played at slow time controls can be mixed together with faster games within a single rating system. One commonly-held (though admittedly subjective) belief is that classical chess is categorically different from rapid chess and even more different from blitz chess and the three types of chess ought to be kept separate.

There is another way to think about this, however. What if classical and rapid and blitz aren't that different from each other? What if they all reveal information about a player's universal chess ability, with the understanding that games become more chaotic and less informative as the rate of play speeds up?

If you accept this concept, then perhaps there is a way to effectively combine over-the-board games from all time controls into a single rating system, to use a single pool of data for analysis, and to create a single "universal" rating for each player. How could we tell, objectively rather than subjectively, whether this is a step in the right direction, or a step in the wrong direction?

If we believe that having three separate rating systems (and hence three separate ratings for each player) is a better approach than having one universal rating system (and one universal rating for each player), then wouldn't that suggest that the FIDE Elo Standard ratings, calculated only from games played at slow time controls, are a purer and superior measure of playing strength at classical chess than a Universal Rating that has been tainted by games at faster time controls? Similarly, would we not expect that the FIDE Elo Rapid Ratings (calculated only from rapid games) are better at measuring players' skill at rapid chess than that same Universal Rating which mixes the faster games with the slower games that supposedly require different skills for success? And the same for Blitz? How should we decide which ratings work better?

There are several ways to assess the accuracy of a rating system, but we propose one approach that is as simple and straightforward a method as you could imagine. We asked one simple question...

"When a game ends in a decisive result (not a draw), did the higher-rated player or the lower-rated player win?"

If players' ratings were completely random and bore absolutely no relationship to true chess strength, then exactly 50% of decisive games would be won by the higher-rated player. If, on the other hand, players' ratings were perfectly accurate, then theoretically 100% of all decisive games would be won by the higher-rated player. While this is clearly an unattainable standard, 75% - 80% is a more reasonable goal, and we believed it was possible to design a rating system that would accurately predict the results of decisive games (discarding drawn results) at a better prediction rate than existing rating systems.

Once the models underlying the URS™ were built, we then decided to put our theory to the test. We started by retroactively calculating URS™ Ratings for the past several years on a month by month basis. This generated results which could be directly compared against the three sets of monthly FIDE Elo ratings to see which ratings (from the start of the month when the game was played) better predicted the outcome of decisive games.

We used the same set of URS™ ratings to determine the URS™ rating favorite in all games. On the other hand, we used the FIDE Standard ratings to determine the FIDE Elo rating favorite in standard games, and the FIDE Rapid ratings to determine the FIDE Elo rating favorite in rapid games, and the FIDE Blitz ratings to determine the FIDE Elo rating favorite in blitz games. Since the FIDE Rapid and Blitz rating systems only came into effect in 2012, we decided to give these ratings a one year grace period to settle, and we therefore started comparing results for all months between January 2013 and December 2016.

An illustrative example of the process that was followed is recreated below. This illustration is based on the results at the recently completed World Blitz Championships that were held in Doha from 29 – 30 December 2016.

For the sake of simplicity, we can look at just a partial cross-table which includes just the nine players who were rated 2800+ on the 1 December 2016 FIDE Blitz rating list. We would then sort these players both by their FIDE Blitz ratings and by their URS™ Ratings as of 1 December 2016. This generates the following two tables:

FIDE Open World Blitz Championship 2016 (using FIDE Blitz Elo ratings)										
			2873	2847	2842	2830	2830	2823	2813	2800	2800
FIDE Blitz Elo			1	2	3	4	5	6	7	8	9
2873	1	Carlsen Magnus	X								
2847	2	Artemiev Vladislav		X							
2842	3	Nakamura Hikaru			X						
2830	4	Aronian Levon				X					
2830	5	Nepomniachtchi Ian					X				
2823	6	Vachier-Lagrave Maxime						X			
2813	7	Mamedyarov Shakhriyar							X		
2800	8	Karjakin Sergey								X	
2800	9	Radjabov Teimour									X

FIDE Open World Blitz Championship 2016 (using Universal Ratings)										
			2836	2781	2772	2772	2770	2768	2760	2718	2714
URS™ Rating			1	2	3	4	5	6	7	8	9
2836	1	Carlsen Magnus	X								
2781	2	Nakamura Hikaru		X							
2772	3	Nepomniachtchi Ian			X						
2772	4	Vachier-Lagrave Maxime				X					
2770	5	Karjakin Sergey					X				
2768	6	Aronian Levon						X			
2760	7	Mamedyarov Shakhriyar							X		
2718	8	Radjabov Teimour								X	
2714	9	Artemiev Vladislav									X

Comparing the tables shows clear differences. For example, GM Vladislav Artemiev was seeded ahead of GM Hikaru Nakamua based on their FIDE Elo Blitz ratings before the event but well behind Nakamura on the URS™ rating list.

Once the actual game results are available, we populate the cross-tables and compare the results. We simply ignore everything below and to the left of the diagonal line since this is a mirror image of the information in the top right. We also ignore drawn games and matchups where the players have identical ratings, since in these rare cases there are no "higher-rated" or "lower-rated" players.

This generates a table where anything shown as 1 in the area to the right and above the diagonal reflects a correct prediction, where the higher-rated player won. Anything that is a zero in this same area is a missed prediction. All of the cells we are disregarding, we have shown in gray, including the decisive results shown to the left and below the diagonal. The correct predictions (the "1" values) are shown in blue and the missed predictions (the "0" values) in red:

FIDE Open World Blitz Championship 2016 (using FIDE Blitz Elo ratings)											
			2873	2847	2842	2830	2830	2823	2813	2800	2800
FIDE Blitz Elo			1	2	3	4	5	6	7	8	9
2873	1	Carlsen Magnus	X		½			1		0	1
2847	2	Artemiev Vladislav		X	0						
2842	3	Nakamura Hikaru	½	1	X			0		1	
2830	4	Aronian Levon				X	X				
2830	5	Nepomniachtchi Ian				X	X	X		0	
2823	6	Vachier-Lagrave Maxime	0		1		X	X	1	½	½
2813	7	Mamedyarov Shakhriyar						0	X	0	
2800	8	Karjakin Sergey	1		0		1	½	1	X	X
2800	9	Radjabov Teimour	0					½		X	X

4 instances of 1 (correct predictions)
\+ 5 instances of 0 (missed predictions)
9 decisive games (where ratings not equal)

Prediction rate = 4 out of 9 = 44.4%

So when we use the FIDE Blitz Elo ratings, Magnus Carlsen's two wins (against the lower-rated Maxime Vachier-Lagrave and Teimour Radjabov) were correct predictions while his loss to Sergey Karjakin (also lower-rated) represents a missed prediction. Overall there were four correct predictions and five misses, for an overall prediction rate (across this tiny sample of nine games) of 44%. Of particular note were Artemiev's loss to the lower-rated Nakamura and Mamedyarov's loss to the lower-rated Karjakin. Also note the extra "X" marks to remind us to disregard any Aronian-Nepomniachtchi and Karjakin-Radjabov results, where the players had the same FIDE ratings, or Nepomniachtchi-Vachier-Lagrave results, where the players had the same URS™ ratings.

When we do the same analysis using the URS™ ratings, the results are as follows:

FIDE Open World Blitz Championship 2016 (using Universal Ratings)											
			2836	2781	2772	2772	2770	2768	2760	2718	2714
URS™ Rating			1	2	3	4	5	6	7	8	9
2836	1	Carlsen Magnus	X	½		1	0			1	
2781	2	Nakamura Hikaru	½	X		0	1				1
2772	3	Nepomniachtchi Ian			X	X	0	X			
2772	4	Vachier-Lagrave Maxime	0	1	X	X	½		1	½	
2770	5	Karjakin Sergey	1	0	1	½	X		1	X	
2768	6	Aronian Levon			X			X			
2760	7	Mamedyarov Shakhriyar				0	0		X		
2718	8	Radjabov Teimour	0			½	X			X	
2714	9	Artemiev Vladislav		0							X

	6	instances of	1	(correct predictions)
+	3	instances of	0	(missed predictions)
	9	decisive games		(where ratings not equal)

Prediction rate = 6 out of 9 = 66.7%

From the URS™ perspective the Nakamura win over Artemiev represents a correct prediction, as does the win by Karjakin over Mamedyarov. So for this portion of the cross-table, the URS more successfully categorized the players, with a 67% prediction rate. While the dataset is clearly far too small to be drawing conclusions from, the example above should serve to illustrate how we can objectively compare the accuracy of two different rating lists that apply to the same games.

The results clearly only start having significance once we start looking at far larger data-sets. We consequently applied the same methodology to all four groups, and all players, at those recently completed World Rapid and Blitz Championships (Open Rapid, Open Blitz, Women's Rapid and Women's Blitz). We found that the URS™ ratings worked better than the FIDE Blitz ratings at predicting the blitz game results and also worked better than the FIDE Rapid ratings at predicting the rapid games.

Below is a high level summary of the results:

	# decisive games	% won by Elo favorite	% won by URS favorite	% games URS is better
World Rapid Championship 2016 (Open)	487	69.8%	70.4%	**+0.62%**
World Rapid Championship 2016 (Women)	124	67.7%	71.8%	**+4.03%**
World Blitz Championship 2016 (Open)	839	65.7%	67.0%	**+1.31%**
World Blitz Championship 2016 (Women)	217	67.7%	72.8%	**+5.07%**
Totals (all four events):	**1,667**	**67.3%**	**69.1%**	**+1.80%**

In the table above, the rightmost column has a color gradient applied so that numbers near zero are white, while more positive numbers are a deeper / darker blue, and negative numbers (had there been any) would have been red. The deeper blue colors illustrate where the superiority of the URS™ is more pronounced.

Still, that is only 1,667 decisive games. What if we cast a wider net and looked at more games? What if we looked at all blitz games, and all rapid games and all classical games, across the entire four-year period stretching from 2013 to 2016?

We did that and here are the results:

Classical Chess 120+ minutes each Jan 2013 to Dec 2016	2013 only	2014 only	2015 only	2016 only	**All rated games 2013-16**
# of decisive games:	542,646	577,003	693,755	762,219	**2,575,623**
% won by Elo favorite:	74.8%	74.9%	75.8%	75.9%	**75.4%**
% won by URS favorite:	75.6%	75.9%	76.6%	76.6%	**76.2%**
% games URS is better:	**+0.78%**	**+0.93%**	**+0.75%**	**+0.72%**	**+0.79%**

Rapid Chess 11-119 minutes each Jan 2013 to Dec 2016	2013 only	2014 only	2015 only	2016 only	**All rated games 2013-16**
# of decisive games:	129,508	239,740	287,651	410,262	**1,067,161**
% won by Elo favorite:	75.7%	75.4%	75.5%	75.8%	**75.6%**
% won by URS favorite:	76.9%	76.6%	76.9%	77.2%	**77.0%**
% games URS is better:	**+1.15%**	**+1.23%**	**+1.42%**	**+1.43%**	**+1.35%**

Blitz Chess 5-10 minutes each Jan 2013 to Dec 2016	2013 only	2014 only	2015 only	2016 only	**All rated games 2013-16**
# of decisive games:	83,910	146,550	186,150	270,275	**686,885**
% won by Elo favorite:	73.8%	74.4%	74.1%	74.3%	**74.2%**
% won by URS favorite:	74.7%	75.2%	75.3%	75.6%	**75.3%**
% games URS is better:	**+0.93%**	**+0.83%**	**+1.24%**	**+1.29%**	**+1.13%**

On a consistent basis, from year-to-year, and across all three rating categories, the URS™ rating engine consistently predicted the results better.

By now, you can probably see where we are going with this. Our findings indicate that that URS™ Ratings are better at identifying who is going to win a classical chess game than the FIDE Standard ratings. The (same) URS™ Ratings are better at identifying who is going to win a rapid chess game than the FIDE Rapid ratings, and the (same) URS™ Ratings are better at identifying who is going to win a blitz chess game than the FIDE Blitz ratings.

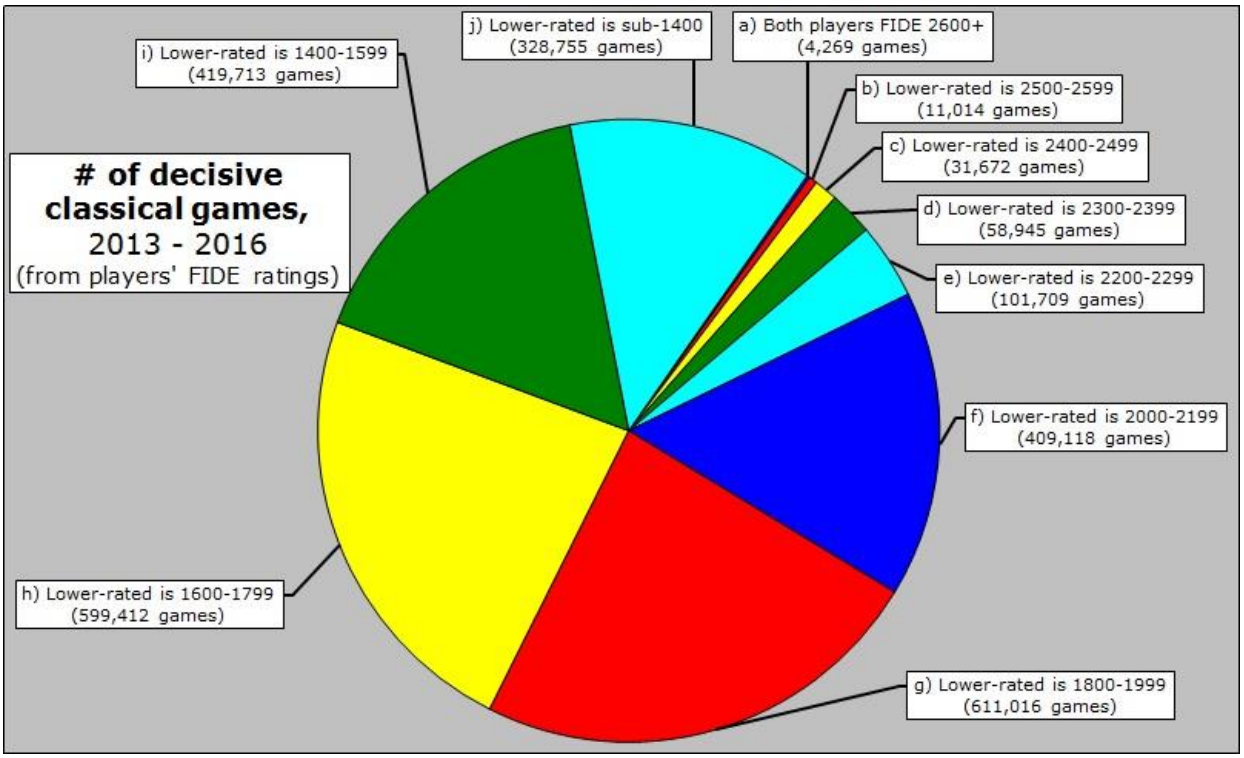
What does this say about the argument that the three types of chess should be kept in isolation within separate rating systems?

These results suggest that URS™ Ratings are, in fact, universally better than Elo ratings at identifying who is going to win a given game of chess. We would further consider this to be objective evidence in favor of the conclusion that ratings from the URS™ are more accurate across the spectrum of time controls than the Elo ratings from the separate rating lists maintained by FIDE.

From a statistical point of view, it is important to notice whether the results from 2016 were just as successful as those from 2013 - 2015. This is important, because when we optimized the inner workings of the URS™ in 2016, we adjusted a very small number of system parameters (approximately ten) to appropriate values. We did this using a statistical methodology that involved predicting the results of actual games played in the period from 2013 to 2015 and then seeing how well our rating system did at making the relevant predictions. The game result data from 2016 was only used as "out-of-sample" data, meaning that it was never run as part of any comparison exercise until we had completed our full and final rating system design. The behavior and results in 2016 can thus be viewed as being the final test. We will of course continue to monitor the behavior of the URS™ into 2017 and beyond.

The analysis above has only looked at overall numbers across the entire pool of players. However, perhaps the URS™ works well for one segment of the rating pool but not for all of it? For instance, the Elo system is known to work much better when players have a large game history, face each other often, and play more consistently. It therefore tends to function better for the top of the rating pool when compared to the entire pool.

Of course, the top of the rating pool includes only a tiny portion of the games played today. This is illustrated by the pie chart below which indicates the relative frequency of games played between players of different strengths, based on the FIDE standard rating of the lower-rated player in each game.



During the four-year period under consideration, there were barely 4,000 decisive games played where both players were rated 2600+. In fact, there were more than 600 decisive games played by lower-rated players, for every 1 decisive game played between 2600+ rated players. The slice is so small that you can barely see the blue slice marked as “a) Both players FIDE 2600+” in the upper-right of the chart.

We checked each of these ten groups of games, ranging from the elite games played among players 2600+, all the way down to games involving at least one player rated below 1400. We then compared how well the URS™ system did at predicting the winners of all the decisive games played when compared to the same players’ FIDE Standard ratings.

Classical Chess 120+ minutes each Jan 2013 to Dec 2016										
FIDE Standard rating (lower-rated player):	2600+	2500-2599	2400-2499	2300-2399	2200-2299	2000-2199	1800-1999	1600-1799	1400-1599	Below 1400
# of decisive games:	4,269	11,014	31,672	58,945	101,709	409,118	611,016	599,412	419,713	328,755
% won by Elo favorite:	63.0%	69.4%	72.5%	74.2%	74.2%	73.8%	74.2%	75.2%	76.8%	79.8%
% won by URS favorite:	63.6%	69.5%	72.5%	74.4%	74.2%	74.3%	74.9%	76.2%	77.8%	80.8%
% games URS is better:	**+0.56%**	**+0.12%**	**+0.03%**	**+0.25%**	**+0.04%**	**+0.53%**	**+0.70%**	**+1.01%**	**+1.02%**	**+1.01%**

Regardless of whether you analyze the small slice representing the elite games, or the larger slice with the weakest players, or anywhere in between, the cells are all blue across the board. This means that at every level of player strength the URS™ better predicted the results than the standard Elo ratings. In some categories, the results were only slightly better, but they were never worse. Not in one single category.

And even though the URS™ is specifically optimized to measure a players' strength at classical chess, it is in fact at rapid and blitz chess that the URS™ truly shows off its superiority. By including classical results within the ratings that are used to predict rapid and blitz games, we enable our rating system to make better predictions, up and down the rating list:

Rapid Chess 11-119 minutes each Jan 2013 to Dec 2016										
Applicable FIDE rating (lower-rated player):	2600+	2500-2599	2400-2499	2300-2399	2200-2299	2000-2199	1800-1999	1600-1799	1400-1599	Below 1400
# of decisive games:	1,788	3,311	6,234	12,788	24,629	106,180	198,365	243,516	217,208	253,142
% won by Elo favorite:	63.9%	66.9%	68.0%	71.6%	72.6%	73.8%	74.8%	75.6%	76.1%	77.5%
% won by URS favorite:	64.9%	67.6%	69.0%	72.0%	73.8%	75.1%	76.1%	76.9%	77.6%	78.9%
% games URS is better:	**+1.01%**	**+0.69%**	**+0.96%**	**+0.43%**	**+1.18%**	**+1.31%**	**+1.30%**	**+1.32%**	**+1.47%**	**+1.41%**

Blitz Chess 5-10 minutes each Jan 2013 to Dec 2016										
FIDE Blitz rating (lower-rated player):	2600+	2500-2599	2400-2499	2300-2399	2200-2299	2000-2199	1800-1999	1600-1799	1400-1599	Below 1400
# of decisive games:	3,466	5,368	9,988	19,672	35,146	126,088	174,346	147,288	97,199	68,324
% won by Elo favorite:	61.0%	65.3%	66.6%	69.2%	70.1%	72.0%	73.6%	75.5%	77.2%	78.8%
% won by URS favorite:	61.7%	66.4%	67.1%	69.5%	70.5%	72.8%	74.9%	76.7%	78.4%	80.7%
% games URS is better:	**+0.75%**	**+1.08%**	**+0.52%**	**+0.25%**	**+0.38%**	**+0.73%**	**+1.21%**	**+1.29%**	**+1.26%**	**+1.91%**

You may observe that even across four years of results, some of the columns are sparsely populated, having only a few thousand games. This is not actually that surprising when we consider how small the slices were for the highest-strength games, in the pie chart presented earlier in this article.

It may also prove interesting to do a more detailed check of player strength versus more specific rates of play, to see if there were any areas where in fact the FIDE Elo ratings were working better than the URS™ at predicting game results. To get sufficient data to look at this in two dimensions, we combined the strongest categories into one larger "Both rated 2000+" category so that we would have five roughly equal-sized groups of games. We could then see if there were any overall groups of players and particular time controls (or ranges of time controls) where the universal ratings were indeed inferior. The most obvious target would be the slowest time controls, for the strongest players, as that is generally the place where the Elo system works best. Games played at this level are typically less random and most players have stable strengths and face each other a lot. It was hence not surprising when it proved that this was indeed the place where the FIDE Elo ratings held up relatively best. Nevertheless, the cells remained consistently blue, with some areas deeper than others, suggesting strongly that the URS™ ratings are in fact universally superior to the FIDE Elo ratings at predicting game results:

Rate of play (maximum # of minutes for each player's first 60 moves)	# decisive rated games 2013-2016	FIDE Rate of Play	Both players rated 2000+	Weaker player rated 1800 -1999	Weaker player rated 1600 -1799	Weaker player rated 1400 -1599	Weaker player rated below 1400
150+ minutes (slowest classical):	902,250	Standard	**+0.22%**	**+0.61%**	**+0.95%**	**+1.04%**	**+0.97%**
135 minutes:	80,585	Standard	**+0.12%**	**+0.64%**	**+0.49%**	**+0.64%**	**+0.90%**
120 minutes (regular classical):	1,537,773	Standard	**+0.60%**	**+0.79%**	**+1.08%**	**+1.05%**	**+1.05%**
90-110 minutes:	213,886	Standard / Rapid	**+1.65%**	**+0.99%**	**+0.89%**	**+1.09%**	**+1.06%**
40-80 minutes:	46,562	Rapid/ Standard	**+1.63%**	**+2.75%**	**+1.95%**	**+2.73%**	**+1.14%**
30-35 minutes:	64,742	Rapid	**+1.59%**	**+1.70%**	**+2.05%**	**+2.18%**	**+1.91%**
25 minutes:	172,767	Rapid	**+1.00%**	**+1.45%**	**+1.93%**	**+1.92%**	**+2.63%**
20 minutes:	83,114	Rapid	**+0.70%**	**+1.33%**	**+1.76%**	**+1.54%**	**+1.83%**
15 minutes:	318,168	Rapid	**+1.38%**	**+1.33%**	**+1.25%**	**+1.66%**	**+2.16%**
11-14 minutes (fastest rapid):	21,677	Rapid	**+2.33%**	**+1.33%**	**+1.49%**	**+1.79%**	**+1.38%**
6-10 minutes (slowest blitz):	85,410	Blitz	**+0.94%**	**+1.43%**	**+1.11%**	**+1.98%**	**+3.49%**
5 minutes:	599,932	Blitz	**+0.59%**	**+1.18%**	**+1.32%**	**+1.13%**	**+1.68%**