A Comparative Analysis of Rating Systems in the US Junior Tennis Development Pathway

Koray S. Abramson
Science Research Program
Pine Crest School

Abstract

The United States Tennis Association (USTA) has historically used point-per-round rankings to determine competitive tournament entry and seeding, but this system often rewards participation over quality of play and can be distorted by random draw effects. Alternative systems such as Universal Tennis Rating (UTR) and World Tennis Number (WTN) use algorithmic predictive modeling based on prior head-to-head results to estimate player ability across gender, age, and geography. Although previous studies (Im, 2023; Kiely, 2025; Krall, 2025; Mayew, 2023) have evaluated predictive accuracy between these two systems using smaller, elite-level samples, large-scale analyses spanning all competitive levels of U.S. junior tennis remain limited. This study addresses that gap through a comprehensive, multi-level analysis of 70,822 USTA junior matches (scraped from January–July 2024), evaluating UTR, WTN, and USTA rankings for both accuracy and bias. Overall, UTR predicted 78.5%, WTN 74.2%, and USTA 70.1% of matches correctly, respectively, with statistically significant differences. Geographic bias was evident across systems, favoring players from less-competitive sections. In matches between similarly rated opponents, players from stronger sections won 61.7% (USTA), 59.0% (WTN), and 53.9% (UTR), indicating systematic underestimation of those cohorts. By combining a large-scale comparative analysis with the first known bias assessment of these systems, this study extends prior evaluations and contextualizes newer findings. The results demonstrate that UTR is the most accurate and least-biased predictor of match outcomes, supporting the adoption of algorithmic, data-driven rating frameworks such as UTR over traditional point-per-round ranking systems in junior tennis.

1. Introduction

Tennis is a sport increasingly analyzed by various systems assessing player performance. While player advancement within tournaments is determined by wins against competing players, the competitiveness of an individual match can be analyzed, allowing for the skill-level of a player to be estimated with greater accuracy. Knowing player skill-level is useful for a wide range of applications: players looking for other players to train with; college coaches assessing a player they might be recruiting; or tournament directors accepting and seeding players in tournaments. This study analyzes various rating and ranking systems utilized by players and coaches in the USTA junior development pathway.

1.1 History of Ranking Systems in Tennis

Since tennis' inception, player strength has been primarily assessed by some variation of a ranking system. The USTA, the ATP (Association of Tennis Professionals), the WTA (Women's Tennis Association), and the ITF (International Tennis Federation) (ITF, 2023; USTA, 2020; USTA, 2022; Wilson, 2023) utilize rankings to determine which players gain entry into tournaments, as well as the seeding of players within a draw, based on the idea that the stronger the level of the player, the better the player's ranking is. Most rankings use a point-per-round (PPR) system. The PPR system first attributes points to a tournament; a tournament with more points attributed to it generally attracts more competitive players than one with less points. For example, in the USTA, the winner of an L1 (i.e., highest-level tournament) receives 3000 points compared to an L5 winner who receives 300 points. Points are awarded based on how far (i.e., how many "rounds") a player advances in a tournament and are aggregated on a rolling 12-

month basis, with a player's ranking based off only the best 6 tournaments of the year for each of singles and doubles (USTA, 2020; USTA, 2022).

There are some significant limitations and flaws to ranking systems. One limitation is the influence of random factors, commonly called the "luck of the draw", which relate to the random nature of a tournament's draw. For example, one player may play the top seed in the first round, while another similarly-leveled player in the same tournament may randomly obtain a much easier path to the later rounds, consequently allowing the "luckier" player to gain more ranking points. Another example could be an injury of an opponent leading to forfeiture that then gives points to a player who did not even compete. Additionally, since USTA rankings are based on a player's best six matches (USTA, 2022), it can reward quantity of play more than quality (e.g., a player competing in eight tournaments a year will have a harder time achieving six great results based on the "luck of the draw" than a player who plays 24 tournaments a year).

1.2 Introduction of Rating Systems to Tennis

More recently, different organizations have started comparing the levels and status of players through new models attempting to create a more accurate system than traditional rankings. Most of these algorithms are variations of the Elo system utilized in chess, such that in head-to-head matches, it is a zero-sum system where the gain in the rating of one player must be offset equally by the loss in the rating of the opponent (Chess.com; Vernon, 2024).

1.2.1 Universal Tennis Rating

In 2008, Universal Tennis Rating ("UTR") was introduced. UTR is a "universal" rating system, which means it attempts to put all players on a single rating scale from 1.00 to 16.50 across all demographics, including gender, geography, and age (UTR Sports, 2023). UTR's algorithm relies on the percentage of games won relative to the rating of an opposing player and is based on a weighted average of a player's last 30 matches, with more recent matches receiving more weight. Unlike rankings, it assesses the competitiveness of a match to determine a "rating" relative to another rated opponent (UTR Sports, 2023). For example, if a player loses 0-6, 0-6 to a 10-UTR competitor, UTR assumes the losing player is significantly below a 10-UTR. In contrast, if the losing player loses 6-0, 6-7, 6-7, UTR will assign a rating for that match greater than the 10-UTR winner, indicating the losing player is stronger; while the losing player lost 2 out of 3 sets, he or she won 56% of the games (18 out of 34). UTR does not care about who wins a match and only looks at percentage-of-games-won relative to the rating of the competitor (UTR Sports, 2023).

1.2.2 World Tennis Number

During the COVID-19 shutdown, UTR gained traction as there were fewer tournaments occurring and naturally rankings became less meaningful. Some tournaments started using UTR for seeding and entry. UTR's parent company also started running its own UTR Tournaments, thus competing with the USTA.

In 2021, World Tennis Number (WTN) was created as an alternative to UTR by the International Tennis Federation (ITF) (ITF, 2024b), which is affiliated with the USTA. WTN, like UTR, applies a rating by assessing the competitiveness of a match and is also meant to be universal. WTN's algorithm differs from UTR's in that it is based only on the percent of sets won; consequently, ratings improve more by winning in straight sets rather than splitting sets in a match (ITF, 2024a). WTN operates on a 40-point scale, with lower numbers denoting higher skill levels, which is the opposite of UTR's convention (USTA, 2023).

1.2.3 Comparison of UTR and WTN

UTR and WTN are correlated (r^2>0.9) for both gender divisions (Figure 1). As skill level increases variability decreases, suggesting greater alignment of the systems for advanced players.

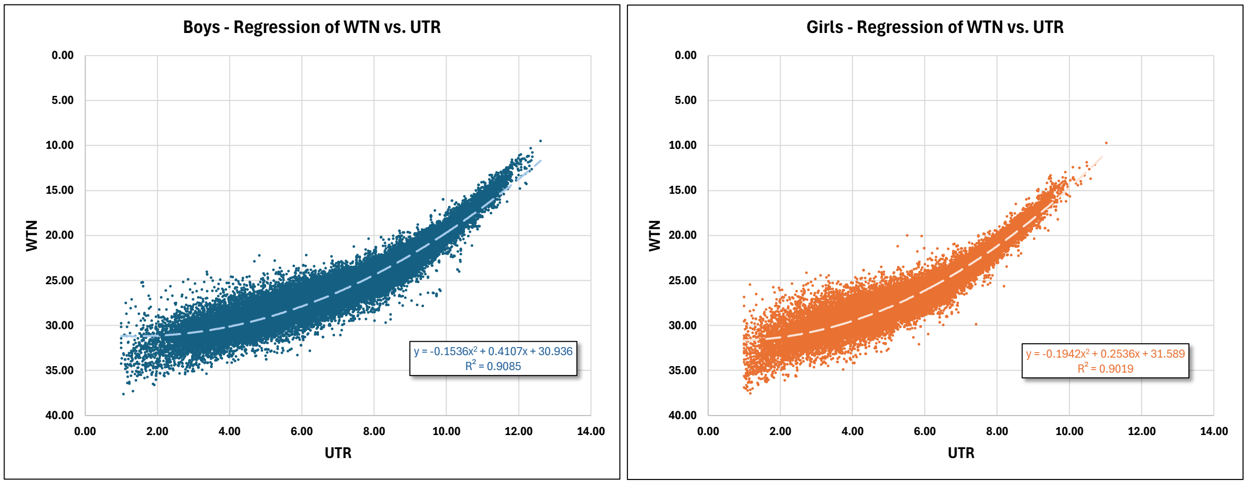


Figure 1: Scatterplots showing a gender-separated dataset of WTN (inverted scale) vs. UTR for 17,278 unique players. The correlation between the two rating scales is calculated with a quadratic line of best fit, with r^2 values larger than 0.9.

1.3 Existing research investigating rating systems in tennis

Previous research on the comparative accuracies of UTR, WTN, and USTA rankings addressed each system's ability to predict match outcomes, although with limited datasets and only at elite level of play (Im, 2023; Kiely, 2025; Krall, 2025; Mayew, 2023).

The first such study analyzed comparative accuracy between UTR and WTN and found these systems to be statistically comparable (Mayew, 2023). The study analyzed 1,532 matches from the USTA National Championships (i.e., elite-level players), spanning two age divisions (16s and 18s). Consequently, the results were limited in that they did not address system performance for younger and developing players in the developmental pathway, thus excluding the majority of junior players. The authors of this study then performed a follow-up investigation (Krall, 2025) using a dataset twice as large, but still limited to the championship level, to assess the effect of a 2023 WTN algorithm change; this study also concluded that neither system had a statistical advantage. Another recent study (Im, 2023) compared UTR, WTN, and USTA rankings and validated previous conclusions indicating that WTN and UTR have similar predictive accuracy; however, across its sample size of approximately 800 matches, it also demonstrated the superior predictive performance of both UTR and WTN relative to USTA rankings.

Most recently, a more comprehensive analysis (Kiely, 2025) from the authors of the initial study compared the predictive accuracy between WTN and UTR within international competition by analyzing 585 matches from the ITA All-American Championships (N.B. international players are a significant portion of collegiate tennis players) and 3,142 matches from various international championship level tournaments for the 12s and 14s division (e.g., Junior Orange Bowl, Les Petits As Mondial). While their initial studies showed comparable performance for UTR and WTN when applied to US-only players within championship-level play, once international competition was a significant part of the dataset, UTR statistically outperformed WTN; the authors surmised that this was potentially due to other countries not being as fully onboarded to WTN as with UTR.

1.4 Purpose of Study

This study seeks to improve upon previous efforts to assess the predictive accuracy of UTR, WTN, and USTA rankings for match outcomes, as described in section 1.3 above. Specifically, the analysis investigates results from 70,822 junior USTA matches scraped from the USTA official website from January through July 2024, combined with rating metrics for 17,278 unique players recorded weekly over this period. The large size of the dataset used in this study permits an evaluation of each system’s ability to predict match results across skill level, gender, and other sub-categories at a statistically significant level. This is also the first study to analyze rating-system universality by quantifying geographic bias across USTA regional sections, identifying whether rating systems systematically under- or over-estimate player ability. Through this combination of large-scale, multi-level data and bias evaluation, this study provides a comprehensive assessment of rating-system performance and practical implications for equitable seeding, tournament placement, and advancement within the US junior tennis pathway.

2. Methodology

To evaluate the predictive accuracies of three tennis rating/ranking systems relative to each other across various player levels and gender, and to determine if any internal geographic bias exists in what are supposed to be universal ratings, a large dataset of match results with corresponding player attributes (e.g., gender, ranking/ratings, level, geography) was required.

2.1 Data Collection

UTR, WTN, and USTA Rankings were scraped weekly from the USTA-affiliated *matchtennisapp.com* website (Match Tennis App; Octoparse). Because player ratings/rankings continually adjust for all players to include the most recent results, data was captured each Thursday in advance of weekend matches; consequently, the dataset contains weekly historical player ratings that are not readily available to the public.

Data was collected for every player competing in Boys' and Girls' Divisions for L1 through L5 tournaments in the 12s, 14, 16s, and 18s from January through July 2024. If a match did not contain complete pre-match fields for both players (i.e., current rating, ranking, name, division, section, gender, match date, and tournament level), it was excluded from the dataset. In total, 83,403 unique player profiles were captured across 17,278 unique players (i.e., individual players that competed in multiple tournaments through the 7-month recording period) with ratings and rankings captured at the time of each match. Match results were then collected from the official USTA website. In total, 70,822 matches had complete player profiles for both competitors, after removing matches between players with an identical rating for UTR or WTN.

2.2 Calculation of Predictive Accuracy for Each Rating / Ranking System

In any given match, UTR, WTN, and USTA rankings all predict a winner based on which player has a higher-level rating or ranking. The predicted result of each system was then evaluated compared to actual match results. Understanding predictive accuracy across multiple skill levels was of interest as previous studies were limited to only the highest skill levels and age groupings. This study allows for a cross-sectional analysis across all skill levels from intermediate to elite juniors.

To analyze predictive ability across different skill levels, matches were grouped into ten evenly spaced decile cohorts based on the average UTR rating of the competitors, independently determined for Boys' and Girls' Divisions. The dataset was further filtered to look at matches between closer-leveled competitors, which was defined as matches between players with a small differential in UTR (between 0.05-0.25) or WTN (0.13-0.65) rating, yielding 19,772 matches to analyze with significant sample sizes within each skill-level cohort (Table 1). This filter attempts to remove the matches that are easy to predict and artificially boost the accuracy of each rating system, as a significant portion of the full dataset contains matches, often in early rounds of tournaments, between players of very different abilities.

Matches Binned by Skill Cohort (For "All Matches" and Filtered for Matches "Between Closely Rated Players") *(Closely Rated Players defined as: UTR Differential 0.05-0.25 or WTN Differential 0.13-0.65)*											
	Matches 0-10%	**Matches 10-20%**	**Matches 20-30%**	**Matches 30-40%**	**Matches 40-50%**	**Matches 50-60%**	**Matches 60-70%**	**Matches 70-80%**	**Matches 80-90%**	**Matches 90-100%**	**All Matches**
Boys UTR Range	0.00-4.31	4.31-5.20	5.20-5.94	5.94-6.59	6.59-7.21	7.21-7.83	7.83-8.43	8.43-9.03	9.03-9.83	9.83-16.00	
Girls UTR Range	0.00-3.20	3.20-4.00	4.00-4.60	4.60-5.12	5.12-5.62	5.62-6.12	6.12-6.64	6.64-7.22	7.22-7.98	7.98-16.00	
All Matches	7,073	7,048	7,100	7,046	7,097	7,104	7,082	7,083	7,084	7,105	70,822
Btwn. Closely Rated	2,181	2,192	2,179	2,148	1,962	1,809	1,810	1,755	1,713	2,023	19,772

Table 1: 70,822 matches were segmented into decile cohorts (>7,000 matches per cohort) based on the average UTR of the two competitors. Higher cohorts represent more advanced junior players (e.g., in the top decile, while this dataset is for USTA juniors under the age of 18, this UTR range would be typical for an NCAA Division 1 college player). The dataset was also filtered to matches between closely rated players, defined as having a small differential between how the competitors were rated by UTR (>=0.05, <=0.25) or WTN (>=0.13, <=0.65).

2.3 Determination of Geographic Bias

To analyze potential geographic bias within rating/ranking systems, which has not been previously studied, matches between similarly-leveled players from “more competitive” regional sections and “less competitive” sections were analyzed; if a system is geographically universal, a similarly-rated player from a less competitive section should have an equal chance of beating a player from a more competitive section. Section competitiveness was determined by analyzing USTA sectional quote data for the 17 geographic sections (USTA, 2024) and was based on a 60%/40% weighting of: (i) sections having the largest player number ranked in the top 150 nationally and (ii) the percentage of section registrants in the top 150 nationally. The most competitive sections (Florida, Southern California, Southern, Northern California, and Eastern) are some of the larger sections, and contain almost 50% of all players nationally (Figure 2).

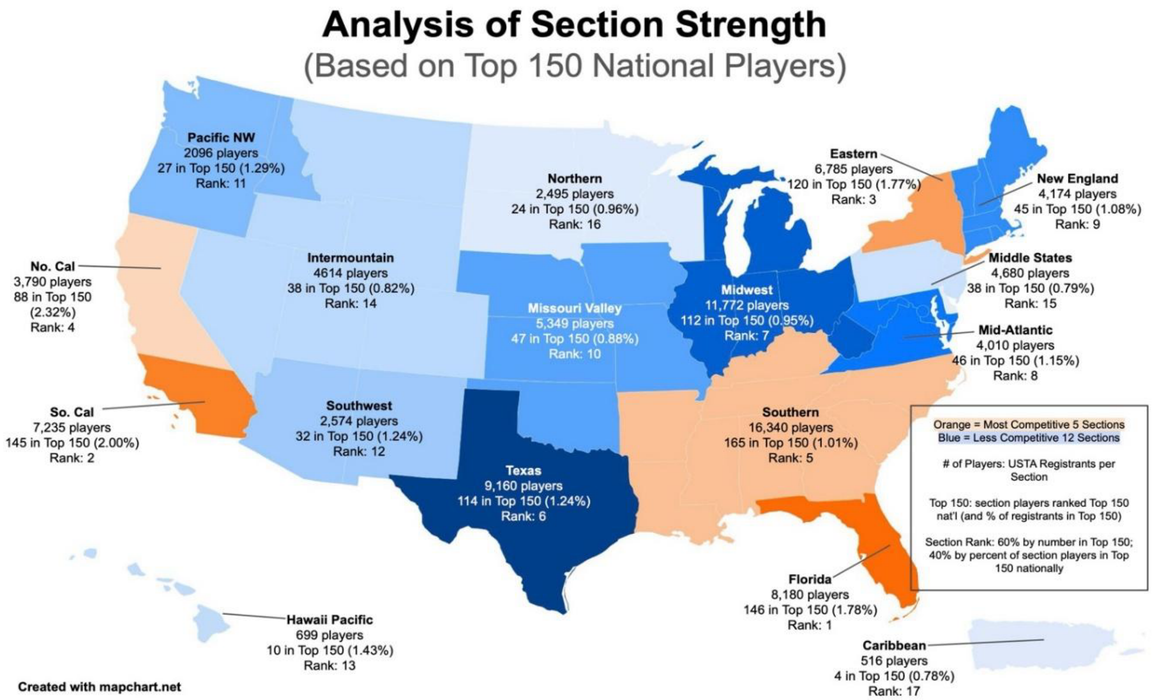


Figure 2: Map of the 17 USTA Sections (i.e., geographic groupings) separated by section competitiveness, determined by analyzing the USTA quota data for entry into the national championship level tournaments. "Most Competitive Sections" are orange on the map and represent 45% of total players nationally. The darker the shading of each color reflects the relative strength of a section with the "Most" and "Less" categories.

Matches from the top 5 skill-level deciles were considered, as this is the most relevant intersectional play; there were 8,096 matches between players from a competitive section and a less competitive section. A subset of "toss-up" matches was analyzed, defined as a differential of 0.25 in UTR, 0.65 for WTN, or 50 spots in ranking. The predictive rating/ranking average was computed for these matches, with a differential of near-zero for all systems (Table 2). The difference of results from the null "50-50" parity hypothesis is interpreted as geographical bias.

Intersection Matches between Near-Equivalent Players from More and Less Competitive Sections *(Player Differentials: UTR<0.25, WTN<0.65, Ranking<50)*				
		UTR	**WTN**	**PPR**
Total Matches		1,633	1,606	1,643
Player Avg. Rating/Ranking	**Most Competetive Sections (MCS)**	8.51	21.59	257
	Less Competitve Sections (LCS)	8.51	21.59	255
Ranking / Rating Advantage to MCS		0.00	0.00	-2

Table 2: Table represents matches between similarly rated/ranked players from one of the more competitive sections ("MCS") competing against a player from a less competitive section ("LCS"). The near equivalence for UTR and WTN (i.e., no differences to the reported precision of 0.00 rating) suggests a player from either section type should have an equal chance of winning. The ranking differential of -2 spots minimally favors the LCS players.

2.4 Statistical Assumptions and Modeling

Throughout this study, conservative binomial assumptions and uncertainties were used to estimate *p*-values and statistical significance. Given the large size of the dataset, in most of the subcategories, the *p*-values for the accuracy difference between the rating systems were negligible, and the statistical significance was consequently extremely high. In subsets segmented by skill level, in addition to *p*-values computed using conservative binomial statistics, McNemar's test was used to quantitatively assess each system's performance given the same set of match outcomes, as it focuses on only discordant prediction pairs (i.e., where only one rating system predicts the correct outcome).

3. Results & Analysis

3.1 Predictive Accuracy of Each Rating/Ranking System

Across all 70,822 matches, UTR's predictive accuracy is highest at 78.5%. WTN and USTA rankings also obtain high accuracy levels of 74.2% and 70.1% respectively (Figure 3). The relative differences in accuracy are at high levels of statistical significance, with *p*-values of effectively zero, demonstrating a clear difference in the performance quality between the three systems (Table 3). While each system exhibits greater predictive performance for Girls' Divisions vs. Boys', this differential is small and not statistically significant (Figure 3).

UTR performs the best across all skill-level deciles, with outperformance greatest at the lower and middle skill-level deciles. At the higher skill-level deciles, UTR's superior performance relative to WTN diminishes, and for the top two deciles UTR's predictive accuracy is only 0.4% above WTN's. However, at higher skill levels, USTA Ranking becomes far less predictive relative to both UTR and WTN (Figure 3).

Overall, these results display considerable and comprehensive evidence for different predictive performance across the three rating systems. UTR consistently outperforms WTN (although marginally at the highest skill-levels), while both systems outperform USTA rankings.

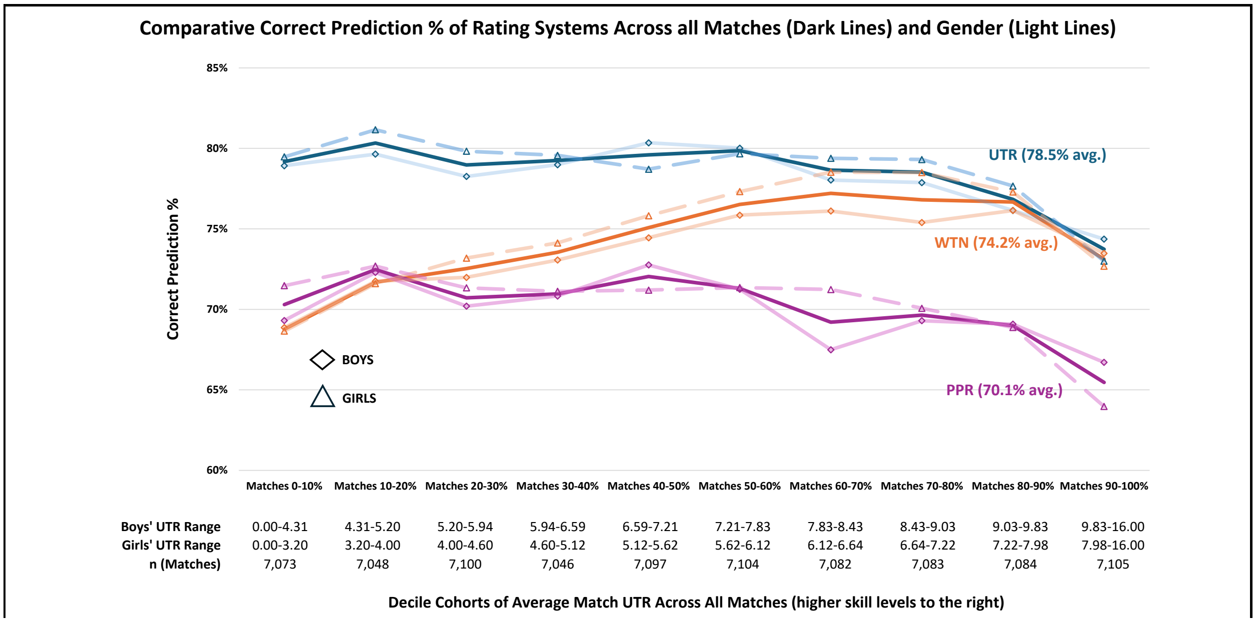
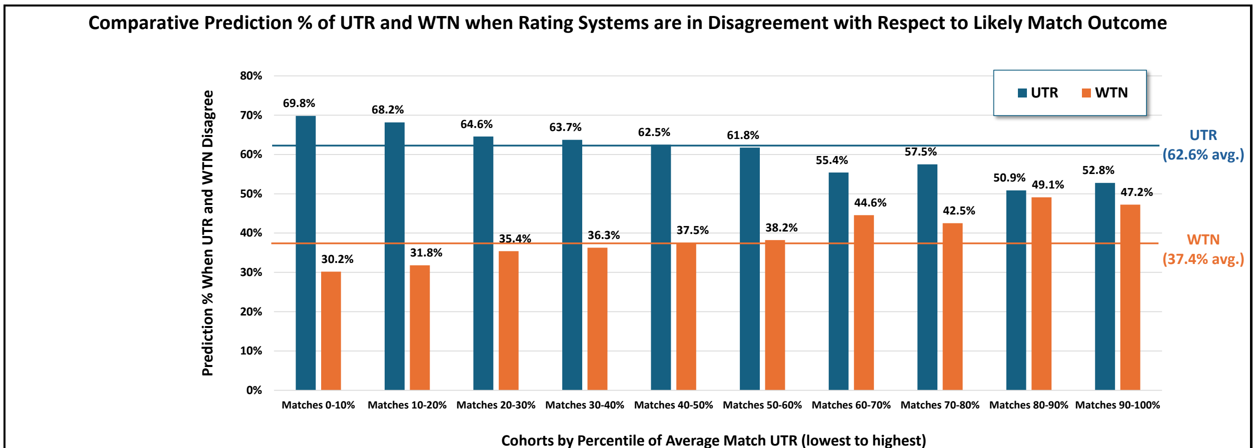


Figure 3: Comparative predictive accuracy for match outcomes of UTR, WTN and USTA Rankings. UTR (78.5% accurate) in aggregate outperformed WTN (74.2%) and USTA (70.1%). At lower skill levels UTR has the greatest differential in performance. While it outperforms WTN at the highest skill-level cohorts, the separation is minimal (Table 3). USTA Ranking becomes even less predictive at higher skill levels. The lighter-shaded lines represent predictive accuracy by gender at each skill-level cohort.

Using McNemar's test, which analyzes the disagreement subset (i.e., isolating outcomes where one algorithm is correct while the other is not), UTR is statistically more accurate when considering all matches, with a p-value near zero. At high levels of statistical significance, UTR outperformed WTN in all skill-level cohorts except in the top two deciles (Table 3). When UTR and WTN disagreed in their prediction of the winner, UTR was correct 62.6% of the matches vs. 37.4% for WTN, with the greatest differential in the lower and intermediate skill-levels (Figure 4). UTR also statistically outperforms USTA Rankings in all cohorts with p-values near zero. Finally, WTN statistically outperforms USTA Rankings when considering all matches, and in all cohorts except for the lower-skilled players comprising cohort 2 (Table 3).

Comparison of UTR, WTN and Rank Predictive Performance Across All Matches and by Skill-Level of Competitors												
Shaded p-values are statistically significant		Matches 0-10%	Matches 10-20%	Matches 20-30%	Matches 30-40%	Matches 40-50%	Matches 50-60%	Matches 60-70%	Matches 70-80%	Matches 80-90%	Matches 90-100%	All Matches
	Boys UTR Range	0.00-4.31	4.31-5.20	5.20-5.94	5.94-6.59	6.59-7.21	7.21-7.83	7.83-8.43	8.43-9.03	9.03-9.83	9.83-16.00	
	Girls UTR Range	0.00-3.20	3.20-4.00	4.00-4.60	4.60-5.12	5.12-5.62	5.62-6.12	6.12-6.64	6.64-7.22	7.22-7.98	7.98-16.00	
	Total Matches	7,073	7,048	7,100	7,046	7,097	7,104	7,082	7,083	7,084	7,105	70,822
	UTR Correct	79.2%	80.3%	79.0%	79.3%	79.6%	79.9%	78.7%	78.5%	76.8%	73.7%	78.5%
	WTN Correct	68.8%	71.7%	72.5%	73.5%	75.1%	76.5%	77.2%	76.8%	76.7%	73.1%	74.2%
	Rank Correct	70.3%	72.5%	70.7%	71.0%	72.0%	71.3%	69.2%	69.6%	69.0%	65.5%	70.1%
UTR vs. WTN	Z-test p-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.039	0.014	0.811	0.403	0.000
	McNemar's p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.674	0.127	0.000
UTR vs. Rank	Z-test p-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	McNemar's p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
WTN vs. Rank	Z-test p-Value	0.049	0.293	0.016	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	McNemar's p-value	0.036	0.248	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 3: UTR statistically outperformed WTN across all 70,822 matches with a *p*-value near zero. It also statistically outperformed WTN in all skill-level cohorts except for the two highest-level deciles. Both UTR and WTN statistically outperform USTA Rankings with *p*-values near zero.



Percentile Cohort by UTR Rating of Matches	Matches 0-10%	Matches 10-20%	Matches 20-30%	Matches 30-40%	Matches 40-50%	Matches 50-60%	Matches 60-70%	Matches 70-80%	Matches 80-90%	Matches 90-100%	All Matches
Matches where ratings systems disagree	1,856	1,678	1,567	1,464	1,285	1,007	940	814	686	794	12,091
Total matches in cohort	7,073	7,048	7,100	7,046	7,097	7,104	7,082	7,083	7,084	7,105	70,822
% of matches where UTR & WTN disagree	26.2%	23.8%	22.1%	20.8%	18.1%	14.2%	13.3%	11.5%	9.7%	11.2%	17.1%
Boys' UTR Range	0.00-4.31	4.31-5.20	5.20-5.94	5.94-6.59	6.59-7.21	7.21-7.83	7.83-8.43	8.43-9.03	9.03-9.83	9.83-16.00	
Girls' UTR Range	0.00-3.20	3.20-4.00	4.00-4.60	4.60-5.12	5.12-5.62	5.62-6.12	6.12-6.64	6.64-7.22	7.22-7.98	7.98-16.00	

Figure 4: When UTR and WTN disagree in predicted outcomes, UTR is the more predictive system across all skill levels. As skill level increases, so does the relative performance of WTN, although it lags UTR in all cohorts.

While each rating/ranking systems had high levels of predictive accuracy (i.e., all above 70%), this is not surprising given tournament construct placing stronger players in different parts of the bracket such that they play head-to-head in later rounds; as a result, in early rounds where a significant number of matches occur, competitors are often at different levels. Across the entire dataset, 50% of matches had UTR differentials of greater than 0.71 (on a 16.50 rating scale) and WTN differentials of 1.58 (on a 40-point rating scale) (Table 4); these differentials imply a meaningful difference in the skill level of opponents, and the higher the differential in rating between players, the easier it is to predict the outcome (Mayew, 2023). For example, for matches with a separation greater than a 0.71 in UTR in competitor rating (which is the median differential across all matches), UTR was correct in predicting the winner 91.4% of the time; WTN was correct 86.8% of the time for matches with a separation greater than 1.58 (Table 4).

Percentiles of Rating Differential for UTR and WTN Competitors									
	Percentile								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
UTR Differential	0.13	0.27	0.40	0.55	**0.71**	0.90	1.14	1.45	1.97
WTN Differential	0.29	0.57	0.89	1.22	**1.58**	1.99	2.51	3.17	4.22
					UTR Predictive Accuracy - 91.4% WTN Predictive Accuracy - 86.8%				

Table 4: Matches were placed in deciles based on competitor rating differential; the smaller the differential, the more competitive a match should be. Over 50% of total matches are between players with differentials greater than 0.71 for UTR and 1.58 for WTN, suggesting that a large percentage of matches should be easy to predict since there are significant disparities in opponent skill levels.

To directly analyze matches between competitors of similar skill levels to exclude easily-predicted contests, matches where competitors were within 0.25 in UTR differential or 0.65 in WTN were analyzed. When considering these closely-rated matches, UTR again statistically outperformed WTN, and did so at all skill levels with the exception of decile nine (Table 5). This finding corroborates the statistical significance determined in 3.1.1 when looking at the dataset in its entirety, with UTR outperforming WTN, and contrasts with conclusions of some previous studies (Im, 2023; Krall, 2025; Mayew, 2023) while corroborating the conclusion of the most recent paper published (Kiely, 2025). The superior performance of UTR is most pronounced between lower- and middle-level competitors but is still apparent using McNemar's test among the highest-skilled players.

Comparison of UTR and WTN Performance Filtered for Expected Close Matches by Skill Cohort (Defined as: UTR Differential 0.05-0.25 or WTN Differential 0.13-0.65)												
Shaded p-values are statistically significant		**Matches 0-10%**	**Matches 10-20%**	**Matches 20-30%**	**Matches 30-40%**	**Matches 40-50%**	**Matches 50-60%**	**Matches 60-70%**	**Matches 70-80%**	**Matches 80-90%**	**Matches 90-100%**	**All Matches**
	Boys UTR Range	0.00-4.31	4.31-5.20	5.20-5.94	5.94-6.59	6.59-7.21	7.21-7.83	7.83-8.43	8.43-9.03	9.03-9.83	9.83-16.00	
	Girls UTR Range	0.00-3.20	3.20-4.00	4.00-4.60	4.60-5.12	5.12-5.62	5.62-6.12	6.12-6.64	6.64-7.22	7.22-7.98	7.98-16.00	
	Total Matches	2,181	2,192	2,179	2,148	1,962	1,809	1,810	1,755	1,713	2,023	19,772
	UTR Correct	66.5%	68.2%	65.8%	65.1%	64.0%	62.1%	61.2%	61.3%	59.0%	57.8%	63.3%
	WTN Correct	55.2%	56.2%	55.0%	55.6%	54.7%	55.8%	57.5%	56.3%	58.7%	55.3%	56.0%
UTR vs. WTN	**Z-test p-Value**	0.000	0.000	0.000	0.000	0.000	0.000	0.023	0.003	0.862	0.113	0.000
	McNemar's p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.000	0.860	0.045	0.000

Table 5: Analysis of matches between closely rated players (as defined above) demonstrates high levels of UTR outperformance in predictive accuracy with statistical significance overall and within all skill-level cohorts with the exception of the 9th cohort.

3.2 Geographical bias

In analyzing the potential for geographical bias, the study filtered the dataset to matches that were considered, at least on paper, to be close to a "toss-up" (competitor differential of UTR <= 0.25, WTN <= 0.65, Ranking <= 50) and between players from a more competitive section and a less competitive section. There were 1,633 toss-up matches measured by UTR, 1606 by WTN, and 1643 by USTA rankings (Table 2). Under the no-bias hypothesis that can be utilized due to the near-zero average differential across every model's subset, the even expectation for a match winner is 50%, assuming the systems are universal (Table 2, Figure 5).

When analyzing match results, however, the "competitive section" player won 53.9% for UTR, 59.0% for WTN, and 61.7% for USTA Ranking (Figure 5). Under the null hypothesis of 50-50 parity between sections, the *p*-value for the UTR-even matches in this subset is 0.0009, and the *p*-values for WTN and USTA Rankings were effectively zero. Thus, a significant level of bias was observed for all three systems; UTR exhibited the least bias, and USTA rankings were the most biased. (Note: USTA Rankings are not intended to be universal by sectional geography, which is why USTA uses a quota system for entry into some national tournaments.)

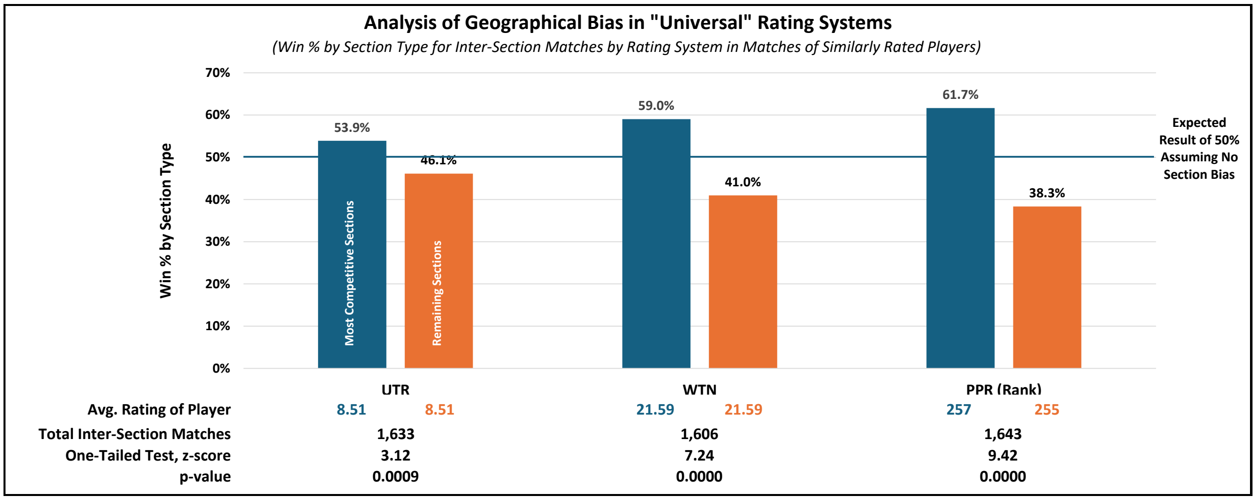


Figure 5: The chart depicts the correct prediction percentage for "toss-up" matches (i.e., near equivalent rating/ranking) between the more competitive sections and less competitive sections (defined per Methodology section) for each system. The difference between the 50-50 expected results and actual percentage of matches won for the more competitive sections demonstrates statistical geographical bias within each of the models, with a *p*-value of 0.0009 for UTR and effectively zero for WTN and USTA rankings.

4. Discussion

This study improved upon and reached different conclusions than previous studies investigating predictive accuracy of tennis rating/ranking systems, particularly when applied to the US junior development pathway. Overall, both UTR and WTN outperformed USTA rankings, validating conclusions in the *ITF Coaching & Sport Science Review* (Im, 2023) study that showed that head-to-head rating models are superior in predictive accuracy. When comparing UTR and WTN, this study demonstrates that UTR outperforms WTN significantly when looking at the entire junior developmental pathway, which includes younger and not-yet-elite-level players. Even at the most elite level of play in the study (i.e., skill-level cohort 10), UTR statistically outperformed WTN when removing the dilution from easier-to-predict matches between competitors with larger rating differentials.

The results of this study diverge from prior research analyzing US-based player datasets (Im, 2023; Krall, 2025; Mayew, 2023), which collectively concluded that UTR and WTN do not exhibit statistically significant differences in predictive accuracy. The most recent investigation (Kiely, 2025), which studied international competition at both the collegiate and 12s and 14s age divisions, concluded that UTR statistically was more predictive than WTN, and surmised that this finding was potentially due to the lack of homogeneity in international competition where WTN is less prevalent, contradicting their previous study on an only US-based player dataset,.

This study, with a dataset encompassing 70,822 matches between US-based players across all competitive levels, demonstrates that even when removing the international element, UTR is still the superior system. By expanding the framework to decompose predictive performance by skill tier, match parity, and geographic region, the study determined that WTN's relative accuracy declines progressively with lower player levels and that UTR sustains stronger predictive consistency across divisions. Even within the highest-skill cohort, once easily predicted matches are excluded, UTR demonstrates statistically significant superiority, underscoring the model's robustness under more stringent predictive conditions.

UTR's superior accuracy could be due to multiple factors. For example, UTR uses more granular inputs as it is based on games within sets, while WTN only considers the winner of sets without considering internal game scores. UTR also has a richer dataset since it aggregates more match sources than WTN (e.g., UTR-only tournaments, high school matches, etc.).

4.1 Limitations

While demonstrating geographic bias for UTR and WTN, this study could not evaluate other elements of universality – specifically as it relates to age and gender. There was no ability to capture the age of a player as the division they are playing in is not representative of their birth year. For gender, while there is no real recorded competition between genders that could result in a meaningful dataset, analysis would suggest that one or both of UTR and WTN is not actually universal across gender. If both systems were universal, the regression of WTN against UTR would produce similar results for both Boys' and Girls' Divisions. As illustrated in Figure 6, this is not the case, indicating that at least one of the rating systems is not universal across gender.

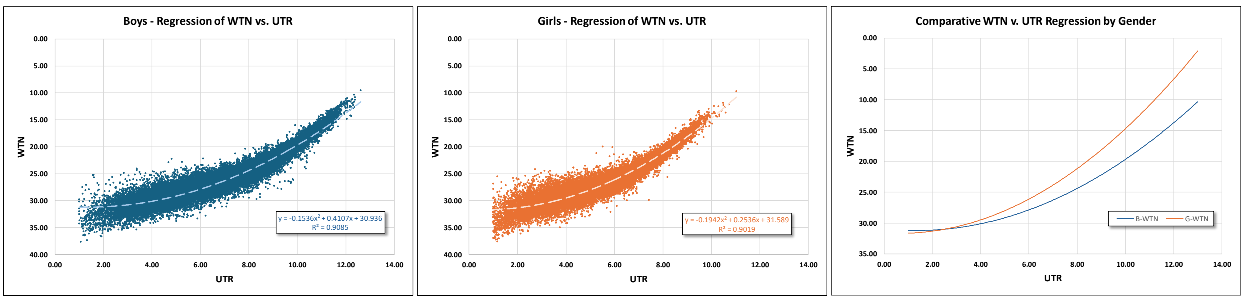


Figure 6: Figure 1 is demonstrated here with the addition of a separate graph displaying just the regression lines for boys and girls UTR vs WTN values. The regression line difference demonstrates that either one or both of the two ratings cannot be truly universal, as the correlation between values of UTR and WTN starkly differentiates between genders as skill level increases.

Additionally, the dataset used in this study predates the September 2024 WTN algorithm update. The ITF stated that their expectation for the outcomes of this revision is that it would have the most significant benefit at the more junior levels (ITF, 2024; Kiely, 2025); this would be of significant importance as WTN underperformance is most pronounced at lower skill levels. Future research incorporating post-update junior data could further validate this interpretation.

5. Conclusions

UTR, in both predictive accuracy and geographical bias, had the best performance of the rating systems studied, at high levels of statistical significance. WTN is also statistically more predictive than USTA Rankings. UTR's outperformance diminishes as skill levels of competitors increase, but when directly selecting matches with closely-rated players (i.e., removing the diluting effect of easily-predicted matches), UTR outperforms WTN across all matches, and does so statistically in nine out of the ten skill-level cohorts.

This study also demonstrates that UTR and WTN are not truly universal when considering geography (i.e., USTA regional sections) as bias was observed. If the systems applied a single scale across all players as they were designed, near equal-rated players from one section would win nearly 50% of the time when competing with a player from another section. This was not the case, and these win-rates deviate significantly (*p*-values near zero) from the 50% parity that is expected when assuming no geographical bias. However, UTR's bias is lower than both WTN's and USTA's, again implying that UTR is the better measurement of skill level.

In summary, while all models analyzed exhibit limitations in their evaluation of player skill level, UTR consistently outperforms both WTN and USTA rankings in both predictive accuracy and in the degree of regional bias over almost all subsets of the dataset, including across skill level and gender. Future studies across additional dimensions and algorithmic design could shed more light on the underlying differences between the predictive performance of these systems.

5.1 Application in Sport

This analysis is applicable to all aspiring tennis players and coaches in the USTA junior development pathway, as it includes data from intermediate-through-advanced skill-level tournaments. While USTA rankings, and at times WTN, are used by the USTA for tournament entry and seeding, these are the two least-predictive systems for assessing skill level compared to UTR; this is especially pronounced for players earlier in their development (i.e., at lower skill level). One recommendation would be for the USTA to preferentially utilize UTR (or even WTN) in granting entry to tournaments, as it is most predictive at assessing player skill level.

Youth tennis has a very significant burnout rate (Gould, 1993) in large part due to the required frequency of play and travel necessary to build a ranking. A majority of players from top college teams previously attended “alternative education” systems (e.g., online schools, tennis academy schools, etc.), as national and ITF tournaments do not align with regular school schedules, with tournaments extending beyond the weekend. Furthermore, players participating in ITF tournaments travel weeks at a time, which adds significant expense to the process.

If USTA and/or ITF utilized UTR (or a similar rating system) for tournament acceptance, burnout rates should decrease. The most-skilled players could then play local tournaments in older divisions against higher-rated players to build their rating with successful outcomes, allowing players to avoid the necessity of travel and excessive tournament play that is currently required to gain ranking points for entry to national-level and ITF tournaments.

6. Acknowledgements

I would like to thank Jed Biesiada, PhD, for his invaluable help through his continued mentorship throughout this project. I very much appreciate his education on the research process, commitment to detail, and his guidance on how to overcome the obstacles that were encountered. I also want to thank my Pine Crest Science Research teachers, Dr. Ganden and Mrs. Gordinier, for both building my foundation of knowledge for creating this project, and their much-appreciated suggestions they provided throughout the research process.

7. References

1. Chess.com. ELO rating system. https://www.chess.com/terms/elo-rating-chess
2. Gould, D., Tuffey, S., Udry, S., & Loehr, J. (1993). Burnout in competitive junior tennis players. https://doi.org/10.1123/tsp.10.4.322
3. Im, S., & Lee, C.-H. (2023). World Tennis Number: The new gold standard, or a failure? *ITF Coaching & Sport Science Review, 31*(91), 6-12. https://doi.org/10.52383/itfcoaching.v32i91.371
4. International Tennis Federation (ITF). (2023). ITF World Tennis Ranking points tables. https://www.itftennis.com/media/9074/itf-points-tables-2023.pdf
5. ITF. Frequently asked questions: What is the ITF World Tennis Number? https://worldtennisnumber.com/eng/faq
6. ITF. (2024a). Enhancement to the ITF World Tennis Number calculation. https://worldtennisnumber.com/eng/news/enhancement-to-the-itf-world-tennis-number-calculation
7. ITF. (2024b). Taking centre court: The rise of the World Tennis Number. https://www.itftennis.com/en/news-and-media/articles/taking-centre-court-the-rise-of-the-world-tennis-number
8. Kiely, L. A., Mayew, R. L., & Mayew, W. J. (2025, April 10). An updated assessment of the predictive accuracy of World Tennis Number and Universal Tennis Ratings. Unpublished manuscript.
9. Krall, N., Maroulis, N., Mayew, R., & Mayew, W. (2025). Initial evidence on the impact of the 2023 World Tennis Number algorithm change for predicting match outcomes. *ITF Coaching & Sport Science Review, 32*(94), 52–58.
10. Match Tennis App. *Match! Tennis.* https://web.matchtennisapp.com
11. Mayew, R. L., & Mayew, W. J. (2023). Which global tennis rating better measures player skill? Evidence from the 2022 USTA Junior National Championships. *The Sport Journal, 26*(2).
12. Octoparse. *Easy web scraping for anyone.* https://www.octoparse.com
13. USTA. (2024). Quota for 2024 USTA National Championships. https://www.usta.com/content/dam/usta/2024-pdfs/2024-quota-final.pdf
14. USTA. (2023). Manual and automatic seeding. https://customercare.usta.com/hc/en-us/articles/360053180492-Manual-and-Automatic-Seeding
15. USTA. (2020). USTA Junior Tournaments Ranking System. https://www.usta.com/content/dam/usta/pdfs/junior-tournaments-ranking-system.pdf
16. USTA. (2022). USTA National Junior Rankings Overview. https://docs.google.com/document/d/1QhbtKvMAM5ZQvcwKm6cw-iF7l8sZpIqL/edit#heading=h.30j0zll
17. UTR Sports. (2023). How UTR rating works. https://www.utrsports.net/blogs/news/how-utr-works

18. Vernon, J. (2024). Understanding the algorithm – complete summary. *UTR Sports.* https://support.universaltennis.com/support/solutions/articles/9000151830-understanding-the-algorithm-complete-summary

19. Wilson. (2023). Tennis rankings explained. https://www.wilson.com/en-us/blog/tennis/how-tos/tennis-rankings-explained