SCIENTIFIC JOURNAL OF SPORT AND PERFORMANCE

ORIGINAL ARTICLE

Comparative performance analysis of rating systems in junior and collegiate tennis

Seth Richey. Department of *Sport Management. Florida State University. Tallahassee, United States of America.*
Nathan David Pifer . Department of *Sport Management. Florida State University. Tallahassee, United States of America.*
James Du. Department of *Sport Management. Florida State University. Tallahassee, United States of America.*
Ryan Rodenberg. Department of *Sport Management. Florida State University. Tallahassee, United States of America.*

ABSTRACT

This paper analyzes the performance of two tennis rating systems—Universal Tennis Rating (UTR) and World Tennis Number (WTN)—in forecasting match outcomes in junior and collegiate tennis tournaments occurring after the September 2024 update to WTN. Results indicate no statistically significant difference in performance accuracy between UTR and WTN in junior tennis. However, in men's collegiate tennis, UTR demonstrates an advantage over WTN in predictive performance. In addition, interaction terms in select models show that the positive effects of a rating advantage on a player's probability of winning are mitigated in matches featuring higher-rated competitors. These findings highlight the importance of using evaluation metrics that are most relevant to a specific context and underscore the role of sports analytics in promoting fairness in competitive tennis.

Keywords: Performance analysis, Tennis player ratings, Performance evaluation, Head-to-head matches.

Cite this article as:
Richey, S., Pifer, N. D., Du, J., & Rodenberg, R. (2026). Comparative performance analysis of rating systems in junior and collegiate tennis. *Scientific Journal of Sport and Performance*, *5*(2), 308-321. https://doi.org/10.55860/BGFB1603

Corresponding author. *Department of Sport Management – Florida State University – Rm. 1014, Tully Gymnasium, 139 Chieftan Way, Tallahassee, FL 32306. United States of America.*
E-mail: david.pifer@fsu.edu
Submitted for publication December 03, 2025.
Accepted for publication February 03, 2026.
Published February 21, 2026.
Scientific Journal of Sport and Performance. ISSN 2794-0586.

doi: https://doi.org/10.55860/BGFB1603

INTRODUCTION

In competitive tennis, traditional metrics like wins and losses or points earned through ranking systems have long been used to gauge player performance. Such metrics aim to produce more competitive matches, improve fairness in tournament entry and seeding, and provide college coaches with tools for recruitment (Mayew & Mayew, 2023). However, these performance-based analytical metrics are often insufficient for capturing the true skill levels of players, particularly because factors like quality of opponents and match competitiveness are not considered.

In junior tennis in the United States, players earn ranking points based on performance in sanctioned tournaments, with increasingly higher points awarded as the rounds of a tournament progress. Collegiate tennis does not utilize ranking points but rather opts for an algorithm that determines player ranking based on win points and loss points, where a player's best *n* wins are averaged against every loss (ITA, 2023b). Rankings have historically been used to measure skill. However, in the last decade, performance-based rating systems have become increasingly popular, particularly in the United States. These rating algorithms, specifically Universal Tennis Rating (UTR) and World Tennis Number (WTN), provide frequently updated ratings that account for factors like strength of opponents and the competitiveness of matches. The aim of this paper is to compare the classification accuracy of UTR and WTN in junior and collegiate tennis.

UTR is a number representing a player's skill level on a scale from 1.00 to 16.50, with a higher number indicating a higher skill rating (UTR, 2023). It is a modified Elo rating system (UTR, 2023). For each match, the algorithm calculates two components: match rating and match weight (UTR, 2023). Match rating is determined by the difference in UTR between opponents and the competitiveness of the match, which is measured by the percentage of total games won (UTR, 2023). Match weight reflects the importance of the match and is calculated using the UTR difference between opponents, match format, and how recently the match occurred (UTR, 2023).

WTN is a numerical rating on a 40-point scale, with a lower number indicating higher skill (ITF, 2022). For a given matchup, the WTN algorithm predicts an expected outcome and updates ratings based on the difference between actual and expected results. Unlike UTR, WTN evaluates matches at a set level, taking individual set results as outcomes rather than the entire match (ITF, 2022). For example, if a match finishes two sets to one in favor of Player A, then Player A's WTN is updated with two set wins and one set loss. WTN utilizes every match since 2016 that is provided by participating tennis associations (ITF, 2022). Additionally, WTN offers a confidence level that measures the accuracy of a player's rating (WTN, n.d.). This confidence level implements the Glicko-2 rating system, an update to the Elo rating system (WTN, n.d.).

Prior to the launch of WTN in 2020 by the International Tennis Federation (ITF), UTR was the premier rating system for predicting player performance (Bodo, 2018). In 2022, the United States Tennis Association (USTA) announced its adoption of WTN, citing future seeding, selection, and flighting in USTA ranked tournaments as the rationale (USTA, 2022). One year later, WTN officially partnered with the Intercollegiate Tennis Association (ITA), the governing body of collegiate tennis (ITA, 2023a). The partnership led to the use of players' WTNs for entry and seeding in major collegiate tennis tournaments (ITA, 2024b).

Although some elements of both rating systems' algorithms are provided by their respective organizations, the specifics of both algorithms are largely kept secret. For instance, both rating systems claim to be on the same scale regardless of age, gender, or skill level (UTR, 2023; WTN, n.d.), but neither organization explains how this is achieved. Furthermore, the revisions applied to both algorithms lack transparency. UTR does not

give notifications of algorithmic updates but does acknowledge that these updates exist (UTR, 2024). The WTN algorithm has likely been updated several times, but only two announcements about changes to the WTN algorithm have been made (ITF, 2024; WTN, 2023).

Several rating systems have been developed to quantify performance and predict outcomes in sports. Adjustive rating systems like the Elo rating system are commonly used in chess and international soccer, where FIFA applies an Elo model to rate teams based on match outcomes, match importance, and opponent strength (FIFA, 2025). Stefani (2011) established a foundational overview of sport rating systems, emphasizing the distinction between "*ratings*" as numerical measures of performance and "*rankings*" as ordinal positions based on ratings. The primary difference between ratings and rankings is that ratings quantify performance based on results and other factors, while rankings simply order competitors. Some rating systems, like the Elo rating system, allow for direct comparison between competitors on a continuous scale where the magnitude of a rating difference conveys information about relative performance. Rankings only indicate the order of competitors.

To assess and improve the utility of performance metrics, Franks et al. (2016) introduced the concept of meta-metrics to address the "*clutter of metrics*" being created by new developments in sports analytics and to pinpoint which metrics provide the most useful information. Franks et al. (2016) define three specific meta-metrics to evaluate player performance measures: stability (i.e., the consistency of a metric over time); discrimination (i.e., the ability to distinguish between players); and independence (i.e., the extent to which a metric provides new information). It was also suggested that metrics be relevant, serving as "*a quantitative summary of the causal or predictive relationship between the metric and an outcome of interest, like wins …*" (Franks et al., 2016, p. 162). Indeed, the ability to make accurate predictions is an important feature of competitive sport, as highlighted by the efforts of prior research to more accurately predict match outcomes in various settings (Oliva-Lozano et al., 2025).

The theoretical frameworks for the rating algorithms used by UTR and WTN can be traced back to chess and the development of the Elo rating system. Elo introduced a probabilistic rating model to rate and compare chess players by updating player ratings based on the difference between expected and actual game outcomes (Elo, 1978). Recognizing shortcomings of the Elo system, Glickman introduced the Glicko and Glicko-2 rating systems, which incorporate measures of rating uncertainty and rating volatility (Glickman, 1995; Glickman, 2012).

In tennis, prior studies have evaluated UTR vis-à-vis WTN. Mayew and Mayew (2023) analyzed 1,532 matches from the 2022 USTA Junior National Championships by utilizing separate logistic regression models for UTR and WTN to estimate the probability of a player winning based on the difference between their rating and their opponent's rating. The classification accuracies were then assessed using both the area under the receiver-operating-characteristic curve (AUC) and Brier score, and the study found no significant difference in the classification accuracies of UTR and WTN. Im and Lee (2023) performed a similar analysis, only including the boys' divisions of the 2022 USTA Junior National Championships. This study also found no significant difference in the classification accuracies of logistic models using UTR compared to logistic models using WTN. Following a known 2023 algorithmic update to WTN, Krall et al. (2024) replicated the methodology of Mayew and Mayew using the 2023 USTA Junior National Championships and the 2022 USTA Junior National Championships, incorporating restated WTN ratings to reflect the update. Their findings showed no systematic improvement in WTN's classification accuracy relative to UTR.

In summary, prior studies evaluating UTR and WTN have applied similar, basic methodologies (i.e., logistic regression models based on the difference in player ratings) to datasets limited to American junior tennis, and have found that there is no statistical difference between the classification accuracies of UTR and WTN (Mayew & Mayew, 2023; Im & Lee, 2023; Krall et al., 2024). This paper extends the literature in three main ways. First, we incorporated collegiate-level match data to examine the performance of UTR and WTN at a higher level of competition. Second, we introduced logistic models with an interaction term to investigate whether the effects of rating differences on players' odds of winning varied with match quality. Third, we conducted regularized ridge regressions to ensure the robustness of our findings. Importantly, this paper evaluates and compares the classification accuracies of UTR and WTN in junior and collegiate tennis following an announced September 2024 WTN algorithmic update.

MATERIAL AND METHODS

We evaluate and compare the classification accuracies of UTR and WTN using established statistical techniques. The design of the research was informed by three prior studies analyzing the accuracy of UTR and WTN using match data from junior tennis tournaments (Mayew & Mayew, 2023; Im & Lee, 2023; Krall et al., 2024). Such methodologies have been adopted and extended in this paper to new datasets in both junior and collegiate tennis contexts. Furthermore, the new datasets exclusively include tournaments that were played after the September 2024 WTN update, allowing for evaluation of the most current version of WTN. Including the junior context provides continuity from the previous literature, while the collegiate context represents a novel extension.

Data collection procedures remain consistent across all components of the study. Analyses focus on match-level results, or outcomes of individual matches, which form the basis for all accuracy calculations. Accordingly, the dataset contains match results along with UTR and WTN ratings for each player. The match results, UTR ratings, and WTN ratings were collected for two junior tournaments and one individual collegiate tennis tournament, with each tournament occurring after the September 2024 WTN algorithm change. UTRs were collected from the UTR website (UTR, 2025). WTNs were collected directly from the public draw sheets when available; otherwise, WTNs were collected from the WTN website (WTN, 2025). Results were collected from the public draw sheets (USTA, 2024a; USTA, 2024b; USTA, 2024c; ITA, 2024a).

The junior tennis sample consists of two tournaments: the 2024 Boys' 16 & 18 USTA National Indoor Championships and the 2024 Boys' 16 & 18 USTA National Winter Championships (USTA, 2024a; 2024b; 2024c). These events were selected based on their similarity in age range, skill level, geographic diversity, and number of matches to prior tournaments used in comparable studies. Both main draw and consolation matches were included, yielding 746 matches from 384 unique players. For the collegiate sample, data were drawn from the 2024 ITA Men's All-American Championships (ITA, 2024a). The tournament included three stages—prequalifying (128 players), qualifying (64 players), and a main draw (64 players). By design, prequalifying and qualifying round draws were not fully completed. This dataset comprised 223 matches from 224 unique players and represented a diverse geographic and age distribution, broader than the junior tennis sample.

After collecting all player ratings and match results, the resulting datasets were cleaned. Matches were removed if players had identical WTNs or UTRs, if the match was not started, or if the match was not played to completion. In the Boys' 16s division, 18 matches were removed because of withdrawals or retirements due to injury or illness, and four matches were removed because the players either had identical WTNs or UTRs. In the Boys' 18s division, 28 matches were removed because of withdrawals or retirements due to

injury or illness, and two matches were removed because the players either had identical WTNs or UTRs. The Boys' 16s division and Boys' 18s division contained 348 and 342 matches in the final dataset, respectively. In the collegiate dataset, seven matches were discarded because of withdrawal or retirement due to injury or illness, while one match was removed because players either had identical WTNs or UTRs. One additional match was removed because a player did not have a listed UTR in the week preceding the collegiate tournament. The final collegiate sample contained 214 matches.

All data cleaning, analysis, and modeling were conducted primarily using R statistical software version 4.5.0. For each match, one player was randomly chosen to serve as the reference player using row-wise swaps in R's "*sample*" function. The primary predictor variable of interest was the difference in the players' respective UTR and WTN ratings, calculated with the reference player as the baseline. Importantly, WTN values were adjusted so that a higher rating indicated higher skill, while the underlying scale remained the same. Separate logistic regression models estimated the likelihood that the randomized reference player won the match (Win = 1; Loss = 0) based on the difference in the two players' ratings.

Each of the initial logistic regression models was then assessed for classification accuracy based on three measures: Brier score, logarithmic loss (log loss), and AUC. A lower Brier score, lower log loss, and higher AUC indicated better classification accuracy on out-of-sample data. Reported evaluation metrics were the measures' averages following a 5-fold cross-validation that utilized random 80% and 20% splits of the observations into training and testing sets, respectively. The general form of the logistic regression model was given as:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot X_i \quad (1)$$

where the probability that the randomly selected reference player won the match (p_i) was determined by the difference in rating between opponents (X_i), the intercept term, β_0, and the logit coefficient, β_1. Converting the log odds to a probability that the reference player wins a matchup was then given as:

$$p_i = (\mathrm{Y}_i = 1 \mid X) = \frac{\exp(\beta_0 + \beta_1 \cdot X_i)}{1 + \exp(\beta_0 + \beta_1 \cdot X_i)} \quad (2)$$

After carrying out the logistic regressions and the 5-fold cross-validations, the p-values from three tests determined if there were statistically significant differences between the classification accuracies of UTR and WTN. A McNemar test of equal proportions was used to compare the percentages that the favored player won the match (McNemar, 1947). Delong's method assessed differences in the classification accuracies based on AUC (Delong, et al., 1988). Paired t-tests were used to compare the classification accuracies based on Brier score and log loss. A *p*-value < .05 was considered statistically significant.

Additionally, extended logistic regression models including an interaction term between the rating difference and the average rating of the two players in each matchup (e.g., if one player had a rating of 12 and the other a rating of 14, the average rating was recorded as 13) were estimated to examine whether the effects of rating differences on match outcomes varied across different levels of match quality. These were included as:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3 (X_{1i} \cdot X_{2i}) \quad (3)$$

where β_3 estimated the interactive effect between the rating difference (X_{1i}) and the average rating of the two players (X_{2i}). The models incorporating the interaction terms were not evaluated for classification accuracy, but they were used to evaluate whether the effect of rating difference on the reference player's likelihood of winning varied depending on the average skill of the matched competitors. To this end, converting the log odds to a probability that the reference player will win a match was given as:

$$p_i = (\mathrm{Y}_i = 1 \mid X) = \frac{\exp(\beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3(X_{1i} \cdot X_{2i}))}{1 + \exp(\beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3(X_{1i} \cdot X_{2i}))} \quad (4)$$

Because the player ratings operate on different scales, standardized coefficients were also calculated in conjunction with the unstandardized coefficients for the models in Equations 1 and 3 to allow for more direct comparisons between the effect sizes of UTR and WTN. Ridge regression was further used to control for overfitting and multicollinearity and to optimize variable selection in the model iterations containing the interaction terms. In general, regularization adds bias to reduce variance. Ridge regression, or L^2-regularization can be understood as the point where the L^2-ball intersects with the minimum-valued contour of the unpenalized cost function (Eq. 6). Thus, the λ, or the regularization parameter (see Eq. 5), can be interpreted as a Lagrange multiplier. In Ridge regression, the L^2-gradient decreases towards 0 as the weight of a variable moves toward 0. Consequently, L^2-regularization moves any weight toward 0 but never reaches a weight of 0. Because we were not aiming to enforce sparsity in the models, L^1 regularization was not considered. The minimization equation for the weights of coefficients was given as:

$$\mathrm{w}^* = \arg\min_{\mathrm{w}} Q(X, \mathrm{y}; \mathrm{w}) + \lambda \mathcal{R}_2(\mathrm{w}), \quad (5)$$

where

$$\mathcal{R}_2(\mathrm{w}) := \frac{1}{2}||w||_2^2 \quad (6)$$

Thus, ridge regression works by shrinking the influence of less important variables rather than testing them with traditional p-values or measures of statistical significance, meaning variables are evaluated by their contribution to overall model fit rather than by hypothesis testing.

RESULTS

Table 1 presents the descriptive statistics for the WTN and UTR ratings across the Boys' 16s, Boys' 18s, and combined junior divisions. Recall that to facilitate comparison between WTN and UTR, WTN values have been inverted so that a higher number consistently indicates higher skill across both WTN and UTR. The sample includes 192 players in the Boys' 16s division and 192 players in the Boys' 18s division, resulting in a total of 384 junior players. On average, players in the Boys' 18s division demonstrated higher WTNs and UTRs than players in the Boys' 16s division, indicating higher skill. Table 1 further provides descriptive statistics of WTN and UTR for the collegiate sample. Recall that WTN values have been inverted so that higher values reflect higher skill levels. Compared to the junior ratings in the previous section, matches in the collegiate sample showcased higher skill, as indicated by the 12.963 increase in WTN and 2.101 increase in UTR.

Table 2 presents the proportions in each junior group and subgroup where the player favored by a respective metric won the match. Across each category, the favored player based on WTN won the match between 72.99% (Boys' 16s) and 73.98% (Boys' 18s) of the time, while the favored player based on UTR won between 74.14% (Boys' 16s) and 73.39% (Boys' 18s) of his matches. Notably, these proportions do not consider the magnitude by which the favored player was favored. The Boys' 18s division also contained 60 total matches (17.54%), nearly twice as many as the Boys' 16s division, in which WTN and UTR disagreed on who the favored player was. The *p*-values from tests of equal proportions indicated no significant difference between the classification accuracies of WTN and UTR in any subgroup or group, suggesting that WTN and UTR perform similarly in classifying matches based solely on the proportion of matches won by the favored junior player.

Table 1. Descriptive statistics for junior & collegiate tennis players.

	Boys' 16s (n = 192)		**Boys' 18s (n = 192)**		**Junior (n = 384)**		**Collegiate (n = 224)**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
WTN	18.936	1.699	21.934	2.010	20.435	2.389	33.398	1.913
UTR	10.400	0.584	11.161	0.608	10.780	0.707	12.881	0.435

Table 2. Favored players in junior & collegiate tennis matches.

	Boys' 16s (n = 348)		**Boys' 18s (n = 342)**		**Junior (n = 690)**		**Collegiate (n = 214)**	
	#	%	#	%	#	%	#	%
WTN favourite wins	254	72.99%	253	73.98%	507	73.48%	124	57.94%
UTR favourite wins	258	74.14%	251	73.39%	509	73.77%	141	65.89%
p-Value of tests of equal proportions	-	.731	-	.862	-	.903	-	.091
Favorito differs between WTN and UTR	34	9.77%	60	17.54%	94	13.62%	41	19.16%

Table 2 also shows the results for the proportions calculated in the collegiate sample. WTN correctly predicted the winner of the match for 57.94% of matches, while UTR correctly predicted the winner in 65.89% of matches. UTR and WTN identified different favored players in 41 instances (19.16% of matches) in the collegiate sample. This was the highest percentage of disagreements observed across the junior and collegiate samples. The p-value of the test of equal proportions ($p = .091$) indicated marginal significance ($p < .10$) that UTR outperformed WTN when choosing the favored player to win in college matches.

Shifting to the outputs of the logistic regression models, Table 3 shows the unstandardized (B) and standardized (β) coefficients for the UTR and WTN difference variables, respectively, derived from the junior sample. The standardized coefficients were included so that cross-metric effects could be compared despite inherent differences in the scales of UTR and WTN. The initial iterations (3.1 and 3.3) display the coefficients, standard errors, and *p*-values for the base models that did not include the average rating of the two competitors and the interaction with rating difference, while the second iterations (3.2 and 3.4) reveal the results of the models that contained those terms.

In isolation, both the UTR difference and WTN difference variables had significant and positive coefficients, suggesting that higher UTR differences (B = 2.07, SE = 0.175, $z = 11.862$, $p < .001$) and higher (inverted) WTN differences (B = 0.645, SE = 0.055, $z = 11.796$, $p < .001$) were associated with increased odds of winning a junior match. The standardized coefficients indicated that a 1 standard deviation advantage in UTR difference was associated with a 1.54 standard deviation increase in a player's log-odds of winning, compared to a slightly lower 1.52 standard deviation increase following an identical advantage in WTN. The

interaction term was not significant in the UTR model, but it was significant and negative in the WTN model (B = -0.056, SE = 0.025, z = -2.244, p = .025). This indicated that, in matches between higher-rated junior players, the positive effect of WTN difference on the likelihood of winning was reduced, suggesting a weaker relationship between rating difference and match outcomes at higher levels of WTN. The UTR base model had a Nagelkerke R^2 of 37.1%, and the second iteration of the WTN model that included the significant interaction term had a Nagelkerke R^2 of 37.7%. Consistent with these results, the ridge regression yielded similarly positive coefficients for the UTR and WTN difference variables (left side of Table 4), indicating that the penalization of the coefficients did not substantially alter the strength or direction of their effects in the junior sample.

Continuing, Table 5 displays the parallel results from the logistic regressions applied to the collegiate sample. These outputs show similar effects for ratings differences on the outcomes of collegiate matches, with both UTR difference (B = 1.845, SE = 0.387, z = 4.766, p < .001) and WTN difference (B = 0.230, SE = 0.076, z = 3.019, p = .003) displaying positive and significant coefficients in the base models (5.1 and 5.3). Further, the standardized coefficients revealed that a 1 standard deviation increase in UTR coincided with a 0.804 standard deviation increase in a player's log-odds of winning, whereas a 1 standard deviation increase in WTN was associated with a smaller 0.468 standard deviation increase in a player's log-odds of winning. The interaction terms were not statistically significant in either of the collegiate models (5.2 and 5.4), implying that the effects of UTR and WTN rating differences on match outcomes were not moderated by matchup quality in the college context. The fits of the base models, as measured by Nagelkerke R^2, suggested that 16.2% (UTR) and 5.26% (WTN) of the variance in collegiate match outcomes were explained by the ratings differences.

The ridge regression output displayed on the right side of Table 4 also revealed positive coefficients for UTR and WTN in the collegiate sample, implying that the coefficient shrinkage had little impact on the initial findings regarding the effects of rating differences on college match outcomes. Ultimately, across both the junior and collegiate samples, the ridge regressions produced coefficients that were similar to the unpenalized logistic regressions, suggesting that the conclusions drawn from the base models regarding ratings differences and their interactions with match quality were robust. However, negative, non-zero interaction terms were retained across both contexts and across both ratings in the ridge regressions, highlighting a diminishing impact of ratings differences on match outcomes involving highly rated competitors that was applicable to both contexts.

Table 3. Logistic regression coefficients for junior sample using UTR and WTN metrics.

	3.1				**3.2**			
	B	SE B	*p*-Value	β	B	SE B	*p*-Value	β
UTR difference	2.07**	0.175	<.001	1.54**	0.928	3.34	.781	1.55**
Average UTR					-0.089	0.163	.563	-0.054
UTR Diff. × Avg. UTR					0.105	0.308	.733	0.045
	3.3				**3.4**			
	B	SE B	*p*-Value	β	B	SE B	*p*-Value	β
WTN difference	0.645**	0.055	<.001	1.52**	1.82**	0.536	<.001	1.54**
Average WTN					-0.036	0.046	.419	-0.057
WTN Diff. × Avg. WTN					-0.056*	0.025	.025	-0.267*

*Note. UTR = Universal Tennis Rating; WTN = World Tennis Number. * p < .05; ** p < .01.*

Table 4. Ridge regression coefficients for junior & collegiate samples.

Variable	UTR (Junior)	WTN (Junior)	UTR (College)	WTN (College)
	β	β	β	β
Rating difference	1.11	1.22	0.678	0.263
Average rating	-0.053	-0.108	0.233	0.132
Rating difference × Average rating	-0.005	-0.182	-0.059	-0.113
λ	0.025	0.026	0.033	0.131

Table 5. Logistic regression coefficients for collegiate sample using utr and wtn metrics.

	5.1				5.2			
	B	SE B	*p*-Value	β	B	SE B	*p*-Value	β
UTR difference	1.845**	0.387	<.001	0.804**	7.65	11.762	.515	0.817**
Average UTR					0.609	0.405	.133	0.224
UTR Diff. × Avg. UTR					-0.446	0.904	.621	-0.073
	5.3				**5.4**			
	B	SE B	*p*-Value	β	B	SE B	*p*-Value	β
WTN difference	0.230**	0.076	.003	0.468**	1.89	1.67	.260	0.437**
Average WTN					0.103	0.089	.248	0.157
WTN Diff. × Avg. WTN					-0.050	0.050	.320	-0.159

*Note. UTR = Universal Tennis Rating; WTN = World Tennis Number. * p < .05; ** p < .01.*

In addition to analysing the ratings' effects on match outcomes through regression coefficients and measures of fit, each base model was subjected to a 5-fold cross-validation to assess its predictive performance through the classification accuracy metrics of Brier score, log loss, and AUC. Table 6 shows the results. Across the entire junior sample, WTN boasted a slightly lower Brier score and a higher AUC, while UTR had a slightly lower log loss. The only test that implied even a marginal difference ($p < .10$) in the models' predictive accuracies for any of the junior categories was log loss in the Boys' 16s, where it was seen that UTR performed slightly worse than WTN (0.547 > 0.528, $p = .094$). However, the cross-validation metrics within the collegiate sample revealed striking differences between the two metrics' predictive performances. In this setting, UTR significantly ($p < .05$) outperformed WTN across all three measures (lower Brier score, lower log loss, and higher AUC), alluding to a propensity for UTR to more accurately classify out-of-sample college match results relative to WTN.

Table 6. Classification accuracies for base logistic regression models in junior and collegiate tennis.

	Boys' 16s		Boys' 18s		Junior		Collegiate	
	EST	95% CI	EST	95% CI	EST	95% CI	EST	95% CI
WTN brier	0.177	(0.154 – 0.196)	0.181	(0.161 – 0.202)	0.178	(0.164 – 0.193)	0.241	(0.228 – 0.256)
UTR brier	0.184	(0.157 – 0.198)	0.177	(0.157 – 0.200)	0.179	(0.164 – 0.194)	0.225	(0.205 – 0.245)
p-value (T-test)	.153		.455		.738		.017*	
WTN log loss	0.528	(0.477 – 0.579)	0.550	(0.498 – 0.603)	0.533	(0.497 – 0.571)	0.678	(0.648 – 0.712)
UTR log loss	0.547	(0.496 – 0.595)	0.525	(0.474 – 0.578)	0.532	(0.499 – 0.570)	0.639	(0.592 – 0.688)
p-value (T-test)	.094		.149		.957		.017*	
WTN AUC	0.812	(0.766 – 0.856)	0.803	(0.759 – 0.853)	0.811	(0.780 – 0.842)	0.625	(0.542 – 0.700)
UTR AUC	0.796	(0.749 – 0.842)	0.811	(0.767 – 0.852)	0.806	(0.774 – 0.838)	0.683	(0.614 – 0.756)
p-value (Delong's)	.118		.555		.613		.025*	

*Note. *p < .05; **p < .01.*

DISCUSSION

This study investigated the effects and classification accuracies of two tennis rating systems—WTN and UTR—as they related to match outcomes in junior and collegiate tennis following an announced September

2024 WTN algorithm update. In the total junior sample, similar patterns found in prior literature emerged (Im & Lee, 2023; Krall et al., 2024; Mayew & Mayew, 2023), as no significant ($p < .05$) differences in classification accuracy were observed between UTR and WTN; however, notable trends within specific divisions were evident. In the Boys' 16s division, WTN slightly outperformed UTR in each performance metric, despite UTR correctly predicting more matches outright. The exact opposite was true of the Boys' 18s division, where WTN was a better predictor at face value, but UTR displayed a better Brier score, log loss, and AUC.

These trends may reflect surface differences, as the Boys' 16s division was played on Har-Tru clay courts, which are often regarded as favouring a more strategic and consistent style of tennis, emphasizing point construction over power. Alternatively, WTN may be better at predicting matches that have an overall lower skill level, a characteristic that is generally true in the younger divisions of tennis. UTR universally saw better classification accuracies in the Boys 18s division, which did not have any matches played on clay. This may also suggest that UTR serves as a marginally stronger predictor of match outcomes as player skill increases. The significant and negative interaction term for WTN further supported this, showing that matches featuring players of a higher skill diminished the effect of WTN differences on winning.

In the collegiate sample, two main findings emerged. First, both UTR and WTN performed worse than in the junior context, as seen by the models' weaker fits and the cross-validation accuracy metrics being higher (for Brier score and log loss) or lower (AUC) relative to the same metrics in the junior sample. Selecting the favoured player based on WTN in the college sample yielded results only marginally better than chance, with a classification accuracy of 57.94%. This stood in stark contrast to the 73.48% classification accuracy in the junior sample by selecting the better player according to WTN. The favoured player according to UTR in the collegiate sample (65.89%) also exhibited lower classification accuracy compared to the junior sample (73.77%). These reduced accuracies may reflect the greater international diversity of the collegiate sample. Unlike American junior tennis, collegiate tennis is characterized by its majority international presence (NCAA, 2023), which likely brings greater variation in playing styles, developmental pathways, and competitive experiences. Such heterogeneity may make it more difficult for rating differences alone to capture the factors influencing match outcomes, thereby reducing classification accuracy in this context.

Second, UTR significantly outperformed WTN across nearly every cross-validation metric at the collegiate level. This further underscores the rating's utility for match forecasting and player evaluation at the higher collegiate level of play, where variability in player profiles can obscure predictive relationships. To the extent that a relevant metric is the one more capable of making accurate predictions on an outcome of interest (Franks et al., 2016), UTR certainly seems to hold an advantage over WTN in forecasting the winners of higher-skilled college competitions.

Furthermore, logistic regression (in the junior sample) and ridge regression models including an interaction term between rating difference and the average combined rating of the two players revealed that the ratings differences' effects tended to decrease as the average rating of the players increased. This suggests that both rating systems may become less effective at predicting outcomes among higher-rated players. Such reduced efficacy may be more prominent in WTN, as the only significant interaction term in the base models came from the WTN model using the entire junior sample. This, combined with its weaker performance in college matches, may be the plausible rationale for WTN including a "*WTN ProZone*" label that conceals the exact WTN values of certain top professional tennis players.

Ultimately, these findings have meaningful implications for coaches, tournament organizers, and governing bodies. More specifically, they highlight the value of robust sports analytics in promoting fairness and

competitive balance in tennis by enabling more accurate and consistent assessments of player ability. By demonstrating that UTR more reliably predicts collegiate match outcomes relative to WTN, the findings underscore how refined rating systems can help ensure that recruitment decisions, tournament seeding, and player development plans are based on the most informative performance indicators. In turn, such evidence-based approaches can reduce bias, create more equitable competition, and allow coaches and administrators to identify and nurture talent with greater precision across diverse playing backgrounds.

Outside of tennis, the methodology developed in this paper provides a general way to compare different rating systems in other sports. By utilizing logistic and ridge regression models to predict winners based on the ratings and assessing the predictive accuracies of these models based on relevant classification metrics, a robust system for identifying which systems are most accurate in predicting game outcomes and identifying stronger competitors was outlined. This framework can be adapted to other sports and games where multiple rating systems exist.

This study additionally stands as a direct empirical application of Stefani's (2011) distinction of ratings and rankings by assessing the effectiveness of UTR and WTN to quantify player performance. The primary analytics focus is also directly aligned with the ideas of discrimination (i.e., the idea that better metrics are able to distinguish between players) and relevancy (Franks et al., 2016). Through logistic models featuring an interaction between average rating and difference in rating, as well as ridge regressions that reinforced feature selection, we found that UTR showed stronger discriminatory ability and more relevance for predicting collegiate winners than WTN.

Nonetheless, even with the significant and insightful findings, this study was not without limitations. To start, the junior and collegiate samples were constrained by sample size and player diversity, particularly in the collegiate dataset, where only one tournament was used. In the collegiate context, results may not generalize to matches played in a team format as opposed to an individual tournament. Additionally, due to varying tournament policies, surface types and match formats were not uniformly controlled across divisions, potentially influencing the classification accuracies of the rating systems. Lastly, future research could explore girls' divisions, lower-tier tournaments, and international datasets to better assess the consistency of our findings across different contexts. Investigating classification performance in team formats and analysing the magnitude of rating-based upsets could also provide valuable insights. Expanding to professional tennis remains challenging due to WTN data restrictions via concealments among certain top-level professional players, but improved transparency could enable similar analyses at the highest levels of the sport.

CONCLUSIONS

By examining player and match data from junior and collegiate tennis, we compared the classification accuracies of WTN and UTR following the 2024 WTN algorithm update. The classification performances of both WTN and UTR across junior and collegiate samples were analysed using logistic regression along with three measures of classification accuracy. In junior tennis, there was no significant difference in the classification performance of the two rating systems. However, there was evidence that UTR outperformed WTN in collegiate samples. Further findings showed that the effect of a rating advantage on a player's probability of winning was diminished in matches featuring highly rated competitors. Such results emphasize the importance of accuracy and reliability in rating systems as they are increasingly being used as a selection criterion for tennis tournaments. Additionally, the methodology applied in this paper provides a systematic way to compare ratings in other sports and games that utilize multiple player-rating systems.

AUTHOR CONTRIBUTIONS

Richey: original idea; data collection; methodologist; original writer. Pifer: corresponding author; editing; method oversight; final draft. Du: method oversight; editing. Rodenberg: idea generation; original draft assistance; editing.

SUPPORTING AGENCIES

No funding agencies were reported by the authors.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

REFERENCES

Bodo, P. (2018, April 4). A look at a possible new, streamlined tennis ranking system. ESPN. Retrieved from [Accessed 2026, 12 February]: https://www.espn.com/tennis/story/_/id/23027866/a-look-possible-new-streamlined-tennis-ranking-system

Delong, E. R., Delong, D. M., and Clarke-Pearson, D.L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837-45. https://doi.org/10.2307/2531595

Elo, A. E. (1978). The rating of chess players: Past and present. Ishi Press.

Fédération Internationale de Football Association (FIFA). (2025, July 10). Latest men's world ranking. Inside FIFA. Retrieved from [Accessed 2026, 12 February]: https://inside.fifa.com/fifa-world-ranking/men

Franks, A. M., D'Amour, A., Cervone, D., & Bornn, L. (2016). Meta-analytics: Tools for understanding the statistical properties of sports metrics. Journal of Quantitative Analysis in Sports, 12(4). https://doi.org/10.1515/jqas-2016-0098

Glickman, M. (1995). The Glicko system. Boston University.

Glickman, M. (2012). Example of the Glicko-2 system. Boston University.

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

Intercollegiate Tennis Association (ITA). (2023a, January 5). Intercollegiate Tennis Association adopts ITF World Tennis Number as exclusive official rating for college tennis. ITA. Retrieved from [Accessed 2026, 12 February]: https://wearecollegetennis.com/2023/01/05/wtn-named-official-rating-of-college-tennis/

Intercollegiate Tennis Association (ITA). (2023b, November 16). ITA Rankings explained. ITA. Retrieved from [Accessed 2026, 12 February]: https://wearecollegetennis.com/ita-rankings/rankings-explained/

Intercollegiate Tennis Association (ITA). (2024a, September 21). DI (65 or more players): 2024 ITA Men's All-American Championships. ITA. Retrieved from [Accessed 2026, 12 February]: https://colleges.wearecollegetennis.com/Competitions/ITA/Tournaments/draws/D1CB8036-87C4-4CBD-B968-929D4500BD0E

Intercollegiate Tennis Association (ITA). (2024b, November 25). ITA Division I Sectional Championships. ITA. Retrieved from [Accessed 2026, 12 February]: https://wearecollegetennis.com/2024/11/25/2024-ita-division-i-sectional-championships/

International Tennis Federation (ITF). (2022, September 5). The science behind ITF World tennis number. World Tennis Number. Retrieved from [Accessed 2026, 12 February]: https://worldtennisnumber.com/eng/news/the-science-behind-itf-world-tennis-number

International Tennis Federation (ITF). (2024, August 19). Enhancement to the ITF world tennis number calculation. Retrieved from [Accessed 2026, 12 February]: https://worldtennisnumber.com/eng/news/enhancement-to-the-itf-world-tennis-number-calculation

Krall, N., Maroulis, N., Mayew, R., & Mayew, W. (2024). 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. https://doi.org/10.52383/itfcoaching.v32i94.559

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2):153-7. https://doi.org/10.1007/BF02295996

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. Retrieved from [Accessed 2026, 12 February]: https://thesportjournal.org/article/which-global-tennis-rating-better-measures-player-skill-evidence-from-the-2022-usta-junior-national-championships/

National Collegiate Athletic Association (NCAA). (2023). Trends in the participation of International Student-Athletes in NCAA Divisions I and II. Retrieved from [Accessed 2026, 12 February]: https://ncaaorg.s3.amazonaws.com/research/demographics/2023RES_ISATrendsDivSprt.pdf

Oliva-Lozano, J. M., Vidal, M., Yousefian, F., Cost, R. & Gabbett, T. J. (2025). Predicting the Match Outcome in the 2023 FIFA Women's World Cup and Analysis of Influential Features. Journal of Human Kinetics, 98, 169-182. https://doi.org/10.5114/jhk/195563

Stefani, R. (2011). The methodology of officially recognized international sports rating systems. Journal of Quantitative Analysis in Sports, 7(4). https://doi.org/10.2202/1559-0410.1347

United States Tennis Association (USTA). (2024a, November 29). Level 1: Boys 16's National Indoor Championships. Retrieved from [Accessed 2026, 12 February]: https://playtennis.usta.com/Competitions/xs-tennis-inc/Tournaments/overview/22FDB77E-53A1-41EA-87FD-0746C6C4BF89

United States Tennis Association (USTA). (2024b, November 29). Level 1: Boys' 18 USTA National Indoor Championships. Retrieved from [Accessed 2026, 12 February]: https://playtennis.usta.com/OPRC/Tournaments/overview/425A9B9D-5D9C-480B-9FD5-02F8837F34FD

United States Tennis Association (USTA). (2024c, December 28). Level 1: USTA National Winter Championships (BG16-18). Retrieved from [Accessed 2026, 12 February]: https://playtennis.usta.com/Competitions/USTANatlcampus/Tournaments/overview/A7F1E626-903A-4787-88C4-7C6B2044427A

United States Tennis Association (USTA). (2022, June 9). USTA launches ITF World Tennis Number widget online. Retrieved from [Accessed 2026, 12 February]: https://www.usta.com/en/home/stay-current/national/usta-launches-itf-world-tennis-number-widget-online.html

Universal Tennis Rating (UTR). (2023, December 19). How UTR rating works. UTR Sports. Retrieved from [Accessed 2026, 12 February]: https://www.utrsports.net/blogs/news/how-utr-works

Universal Tennis Rating (UTR). (2024, September 18). FAQ: UTR Rating algorithm. UTR Sports. Retrieved from [Accessed 2026, 12 February]: https://support.universaltennis.com/en/support/solutions/articles/9000233354-faq-utr-rating-algorithm

Universal Tennis Rating (UTR). (2025). Retrieved from [Accessed 2026, 12 February]: https://app.utrsports.net/home

World Tennis Number (WTN). (n.d.). Frequently asked questions. ITF World Tennis Number. Retrieved from [Accessed 2026, 12 February]: https://worldtennisnumber.com/eng/faq

World Tennis Number (WTN). (2023, July 7). Notification of enhancement to the ITF World Tennis Number algorithm calculation. ITF World Tennis Number. Retrieved from [Accessed 2026, 12 February]: https://worldtennisnumber.com/eng/news/notification-of-enhancement-to-the-itf-world-tennis-number-algorithm-calculation

World Tennis Number (WTN). (2025). Search Players. Retrieved from [Accessed 2026, 12 February]: https://worldtennisnumber.com/eng/player-search/