Internet Appendix for: "Which global tennis rating better measures player skill? Evidence from the 2022 USTA Junior National Championships" ***The Sport Journal, 26 (e08252023-2): 1-9***

In the manuscript, we assess statistically significant differences in predictive accuracy of UTR versus WTN via overlap in the 95% confidence intervals around our estimates. While prior papers in the tennis match forecasting literature rely on 95% confidence intervals to draw inferences (e.g. del Corral and Prieto-Rodiguez 2010), it is well known that assessing statistical significance as we do in the manuscript is excessively conservative (Greenland et al. 2016). That is, if the confidence intervals do not overlap, equality is rejected at the 5% significance level but overlapping 95% confidence intervals can still potentially reject equality. We utilize a conservative approach to assessing statistical significance because when predicted values are very highly correlated across measures (as they are in our sample given the very high correlation between UTR and WTN), AUC values become highly correlated and over-rejection of the null hypothesis of AUC equality can occur (Robin et al. 2011).

In Table IA.1 below, we provide a less conservative statistical test and compare AUC values via bootstrapping with 1,000 iterations. We also assess statistical significance for overall accuracy and Brier scores for completeness. Because AUC and Brier scores are derived from the logistic regression estimates, consistent inferences between AUC and Brier scores would lend comfort that any statistically significant difference in AUCs is not simply an artifact of using AUC as our preferred measure of predictive accuracy.

In Panel A we first consider accuracy based on whether the player with the higher (lower) UTR (WTN) won the match. The proportion of matches in which UTR correctly predicts the outcome is not statistically different from WTN in the overall sample or in any subsample at the 5% significance level. In terms of magnitudes, accuracy differences are small and range from -0.009 (i.e. WTN is superior) to 0.002 (i.e. UTR is superior). Of course, these accuracy results reflect how well the binary difference in ratings performs in identifying a match winner and does not accommodate the full information contained in the rating differences between players. Our logistic regressions consider the magnitude of the rating differences between players.

Panels B and C contain the AUC values and Brier scores, respectively. In Panel B, we first present the correlation between predicted values that result from the logistic regression estimation. The correlations are quite high in the overall sample suggesting the possibility of over-rejecting the null hypothesis. In all subsamples it is above 0.68 except in the small WTN subsample. We find in the overall sample that pools together all four events, UTR provides a statistically higher AUC than WTN (p=0.01) by 3.5% (73.9% versus 70.4%). The superiority of UTR is corroborated with the Brier score difference in Panel C (p=0.04). However, no statistically significant result is observed when we decompose the overall tournament by gender, with an AUC (Brier score) p-value of 0.10 (0.07) for boys and 0.07 (0.34) for girls. In a separate study, Im and Lee (2023) examine only the boys division and also find a higher AUC for UTR versus WTN, but also find the difference to be statistically insignificant. When we decompose the overall sample by age, we find that UTR outperforms WTN statistically (p=0.01) for the 16u division, a finding that is corroborated with Brier scores (p=0.03). Decomposing the overall sample by main draw and consolation, UTR outperforms WTN in the consolation draw (p=0.04) but not the main draw when considering AUC. These findings are not corroborated with Brier scores, however, as Brier scores are statistically equivalent for both the main draw and consolation draw. Using AUCs, we find UTR also outperforms WTN in the large UTR difference subsample (p=0.03), the large WTN subsample (p=0.00)

and the small WTN subsample (p=0.00). However, Brier scores for these subsamples are not statistically significant in favor of UTR over WTN.

Overall, of the eleven comparisons of AUC we undertake, we find UTR to be statistically superior to WTN in six cases. Of those six cases, two are consistent with results based on Brier scores. On balance, there is more evidence consistent with UTR and WTN being statistically equivalent than being statistically different predictors of match outcomes. We do note, however, that given the consistency in the overall results and 16u subsample results in terms of AUC and Brier scores, one possibility is that UTR is a superior predictor solely for the 16u division, and this finding is what drives statistical significance in the overall sample. What makes the 16u subsample special is unclear and an important area of future inquiry.

Internet Appendix Table IA1: Additional Accuracy, AUC and Brier Score Comparisons

	1	2	3	4	5	6	7	8	9	10	11
Sample	**Overall**	**Boys Only**	**Girls Only**	**18U Only**	**16U Only**	**Main Draw**	**Consolation Draw**	**Large UTR Difference**	**Small UTR Difference**	**Large WTN Difference**	**Small WTN Difference**
# Observations	**1532**	**739**	**793**	**713**	**819**	**786**	**746**	**758**	**774**	**768**	**764**
Panel A: Accuracy											
FAV_UTR_WIN	0.754	0.760	0.748	0.753	0.755	0.782	0.724	0.890	0.620	0.875	0.632
FAV_WTN_WIN	0.757	0.758	0.757	0.756	0.758	0.781	0.732	0.893	0.620	0.878	0.636
Difference	-0.003	0.002	-0.009	-0.003	-0.003	0.001	-0.008	-0.003	0.000	-0.003	-0.004
p-value of difference[a]	0.69	0.15	0.43	0.82	0.74	0.90	0.52	0.32	0.80	0.31	0.80
Panel B: AUC											
Correlation of predicted *UTR* and *WTN*	0.80	0.76	0.83	0.82	0.77	0.80	0.77	0.70	0.77	0.68	0.38
AUC_{UTR}	0.739	0.749	0.730	0.731	0.748	0.754	0.719	0.695	0.598	0.730	0.629
AUC_{WTN}	0.704	0.714	0.694	0.713	0.696	0.726	0.676	0.639	0.558	0.667	0.548
Difference	0.035	0.035	0.036	0.018	0.052	0.028	0.043	0.056	0.040	0.063	0.081
p-value of difference[b]	**0.01**	0.10	0.07	0.38	**0.01**	0.16	**0.04**	**0.03**	0.12	**0.00**	**0.00**
Panel C: Brier scores											
$BRIER_{UTR}$	0.162	0.158	0.166	0.164	0.160	0.148	0.177	0.093	0.229	0.101	0.223
$BRIER_{WTN}$	0.167	0.165	0.169	0.166	0.168	0.153	0.182	0.093	0.232	0.104	0.230
Difference	-0.005	-0.007	-0.003	-0.002	-0.008	-0.005	-0.005	0.000	-0.003	-0.003	-0.007
p-value of difference[c]	**0.04**	0.07	0.34	0.56	**0.03**	0.09	0.21	0.88	0.45	0.19	0.08

[a] Two-sided p-value of McNemar test of equal proportions for paired data

[b] Two-sided p-value of bootstrap tests of AUC equivalence using 1,000 iterations

[c] Two-sided p-value of paired t-test

Values in **bold** statistically significant at 5% level