Assessing Bias by Athletic Conference in Rating Percentage Index (RPI) Rankings for NCAA Division I Men's Soccer Teams

Conrad Lee

May 7, 2025

Abstract

This undergraduate thesis examines the hypothesis that the Rating Percentage Index (RPI) metric systematically overvalues teams from stronger conferences in NCAA Division I Men's Soccer. Alternative RPI weightings are proposed and tested, first using Bradley-Terry-Davidson simulations and then using real data from the 2024 NCAA season, to determine whether they reduce conference bias while improving predictive accuracy. Simulations show a clear bias in RPI by athletic conference; when compared to a baseline predictive Bradley-Terry-Davidson model using Bayesian inference, RPI is less accurate and systemically favors stronger-conference teams. In addition, alternative weightings of the RPI formula reduce this bias while increasing predictive accuracy of win-loss outcomes. However, when data from the 2024 season is examined, the default RPI weighting achieves greater accuracy than selected alternatives despite its preference for stronger-conference teams. This is proposed to be a consequence of the tendency for teams from stronger conferences to play more games with home advantage, an advantage which skews the RPI calcula-

tion. The paper concludes with suggestions for the NCAA to modify the Men's Soccer RPI formula to reduce home advantage bias and conference bias. Taken as a whole, this paper provides a nuanced picture of various inefficiencies in the RPI metric for NCAA D1 Men's Soccer and motivates further research, particularly with empirical data from prior seasons.

1 Introduction

This undergraduate senior thesis project seeks to investigate whether the Ratings Percentage Index (RPI), a metric of pairwise comparison for selecting teams to the NCAA Division 1 Men's College Soccer Championship, systematically over-values teams from stronger Division 1 conferences. The project begins with a formal mathematical analysis of the RPI formula as used by NCAA Men's Soccer and the intuition behind its potential bias towards teams from stronger Division 1 conferences. Then, soccer matches are simulated using a Bradley-Terry-Davidson model to identify potential reweightings of the RPI formula. The best potential reweightings are selected for further analysis to compare preference for stronger-conference teams as well as predictive accuracy. After that, similar methods are applied to real NCAA college soccer data to examine how simulation patterns hold with historical results. Lastly, a detailed analysis of findings provides insights to the efficacy of the RPI as a metric for evaluating NCAA Men's Soccer teams across conferences in the NCAA Championship selection process.

2 Background and Motivation

The National Collegiate Athletic Association (NCAA) is the chief administrator for varsity college athletics in the United States. The body oversees 24 different

sports, including soccer (also known as association football outside of the United States). The largest universities compete at the Division I (D1) level in Men's Soccer. As of the 2024 season, there were 212 NCAA Men's Soccer teams at the D1 level, competing against each other in 22 conferences as well as in non-conference matchups. Official NCAA matches are conducted from August to December.

Around midway through each season, the NCAA begins to publish its Adjusted Rating Percentage Index (RPI) rankings for the 212 D1 Men's Soccer teams. RPI provides a single metric to evaluate all teams based solely on their record and their strength of schedule. The NCAA updates these rankings weekly until the NCAA D1 Men's Soccer Championship begins.

The NCAA provides a D1 championship for Men's Soccer as outlined by the annual NCAA Division I Manual, following the conclusion of all regular-season matchups and conference championships. As of the 2024 edition, the Men's Soccer championship is intended "to provide national-level competition among the best eligible student-athletes and teams of member institutions" [1]. The procedure for selection to the Championship is specified by the NCAA's 2024-25 DI Men's Soccer Championship Prechampionship Manual. The Championship is formatted in a 48-team, single-elimination bracket. The winners of each of the 22 conferences receive automatic qualification. The 26 remaining teams are selected at-large by an 8-member selection committee in a process detailed by the NCAA. RPI plays a key role in this process, as demonstrated in the following section.

2.1 Selection Procedure

According to the DI Men's Soccer Championship Prechampionship Manual, at-large selection to the Championship ["selection"] will be determined on the basis

of three criteria, which selection for all NCAA Championships are determined by:

- "Won-lost record;
- strength of schedule; and
- eligibility and availability of student-athletes for NCAA championships" [2].

In addition, a team must have a DI won-lost-tied record of .500 or above to be eligible for selection. (The committee's deliberations are private and rationale for decisions is not officially revealed.)

As outlined in the Prechampionship Manual, the Men's Soccer selection committee is required to consider other criteria specific to Men's Soccer in the selection process. This includes the Adjusted RPI, which includes the following:

- Division I winning percentage with ties calculated as 1/3 of a win ["Win Percentage" or "WP"] (25%)
- Strength of schedule ["Opponent's Win Percentage" or "OWP"] (50%)
- Opponents' strength of schedule ["Opponent's Opponent's Win Percentage" or "OOWP"] (25%)
- A bonus/penalty adjustment, based on outstanding results when accounting for home-field advantage [2]

In addition, the committee must consider the following criteria:

- Head-to-head competition
- Results versus common opponents
- Strength and results against nonconference opponents

- Result against teams already selected (including automatic qualifiers with an RPI of 1-75)
- Late-season performance in last eight games (strength and results)
- Strength and result against conference opponents (regular-season and post-season) [2]

These are broad standards for selection and leave the potential for the selection committee to justify the selection of any team they see fit. For example, suppose the selection committee is deciding between two teams for the final at-large bid in the NCAA Championship, Team A and Team B. Both teams have similar RPIs, similar schedules and compete in the same conference. They tied in their head-to-head matchup. Team A has a record of 10-5-1 (6-0-0 in non-conference play and 4-5-1 in conference play), while Team B has a record of 10-5-1 (1-5-0 in non-conference play and 9-0-1 in conference play). In this example, the selection committee could be justified in selecting Team A because of its superior non-conference record. However, the selection committee could also be justified in selecting Team B because of its superior in-conference record.

As shown in this example, the selection committee's arbitrary criteria can (and has) lead to controversy in practice [3]. Without a clear overarching system or metric, many teams can make a case for selection. However, it is accepted in practice that the committee selects the teams with the highest RPI, with limited deviations [4]. Teams with high RPIs that fail to win their conference tournament can generally be confident that they will qualify for the NCAA Championship.

Thus, although the RPI metric does not provide a completely deterministic means of selection to the NCAA tournament, it is reasonable to assume that the RPI serves as a strong *baseline ranking system* to which adjustments can be made according to the NCAA's arbitrary conception of the "best... teams of

member institutions" [1].

2.2 After Selection

After the 26 at-large teams are selected, the selection committee seeds the top 16 teams and gives them a first-round bye. The remaining teams are paired according to geographical proximity in matchups that avoid potential intra-conference matches in the first two rounds of the championship. [2]. Notably, RPI does not play an official role in this process; the committee does not need to justify its choices based on any official seeding criteria. Chris Thomas notes on his site "RPI for Division I Women's Soccer" that this is likely intentional. As he describes for the Women's Soccer Championship, which follows a similar procedure to that for the Men's Soccer Championship:

"An authoritative source has told me that it is 'not unintentional' that the Manual does not state how the Committee is to seed teams. Thus, although the Committee almost certainly considers the at large selection criteria in doing the seeding, the Committee is not bound by the criteria, nor is it foreclosed from giving weight to other considerations" [5].

After the selection committee chooses the bracket for the championship, games are played until a winner is determined. After the conclusion of the season, coaches and athletics staff for the 212 NCAA D1 teams work to set their non-conference schedules in an open market with other universities. Results from the previous season and the final RPI rankings may impact how coaches select their non-conference matches for the following season, particularly for teams with larger budgets who can entice potential opponents with financial incentives. As RPI places a strong weight on strength of schedule, a team's prior results may affect who chooses to play against them in non-conference play in the following season.

Measuring the effect of prior RPI rankings or financial resources on the matchup selection process is outside the scope of this paper. Instead, this thesis focuses on the impact of RPI on selection to the NCAA Men's Championship. The relationships between RPI and team seeding or RPI and team schedule selection is an opportunity for further research.

3 RPI

3.1 Overview

The following section explains the RPI metric in depth and explores claims about potential bias towards teams from stronger conferences. Note that this thesis does not seek to determine the rationale for the NCAA's choice of the RPI metric as a baseline ranking system, despite the existence of other, more established pairwise-ranking systems such as the Bradley-Terry-Davidson model. (The NCAA did stop using the RPI metric in Men's Basketball in 2018, replacing it with the NCAA Evaluation Tool (NET). Women's Basketball switched to NET in 2020 [6].)

3.2 Standard RPI

As mentioned above, the standard form of RPI is defined as

$$\text{RPI} = 0.25 \times \text{WP} + 0.50 \times \text{OWP} + 0.25 \times \text{OOWP}$$

where WP = Winning Percentage, OWP = Opponents' Winning Percentage, and OOWP = Opponents' Opponents' Winning Percentage.

Note that results against the team in question are not included in calculations of OWP. However, the team in question is included in calculations of OOWP.

For the purposes of this undergraduate thesis, it makes sense to think about the RPI metric in two parts: the part which is directly influenced by the team's performances (WP) and the part which is mostly not (OWP and OOWP). Thus, we will think about the RPI formula as such:

$$\text{RPI} = \text{Performance} + \text{Context}$$

where

$$\text{Performance} = 0.25 \cdot \text{WP}$$

and

$$\text{Context} = 0.5 \cdot \text{OWP} + 0.25 \cdot \text{OOWP}.$$

The RPI formula has recently been updated for Men's Soccer; now, ties count as 1/3 of a win in the calculation of WP. This matches the weighting of the points-based ranking system used in professional soccer, where a win is 3 points and a tie is 1 points. However, ties are still valued as 1/2 of a win in the calculations for OWP and OOWP.

3.3 Generalized RPI

As Thomas notes in his Women's College Soccer blog, the selection of weights in standard RPI metric are arbitrary [7]. We can generalize the standard RPI formula, so that it becomes RPI = Performance + Context, where

$$\text{Performance} = c_1 \cdot \text{WP}$$

and

$$\text{Context} = c_2 \cdot \text{OWP} + c_3 \cdot \text{OOWP}$$

where c_1, c_2 and c_3 are nonnegative constants such that $c_1 + c_2 + c_3 = 1$.

Note that c_1 then reflects the percentage of RPI weight on performance while $c_2 + c_3 (= 1 - c_1)$ reflects percentage of weight on context.

Note that this thesis will also present RPI weightings using the shorthand notation (c_1, c_2, c_3).

3.4 Adjusted RPI

It is important to note that the metric in its current form does not factor in home-field advantage. Likely for this reason, the NCAA uses an Adjusted RPI ranking when ranking teams. The Adjusted RPI used by the NCAA for D1 Men's Soccer rankings can be defined as follows:

$$\text{RPI} = 0.25 \times \text{WP} + 0.50 \times \text{OWP} + 0.25 \times \text{OOWP} + \text{A}$$

where A is a cumulative bonus or penalty for good wins/ties or bad losses/ties, determined by the RPI rank of the opponent. The full table of adjustments is outlined in the Prechampionship Manual on page 26; it will not be restated here [2]. Note that the penalty or bonus adjustment is calculated after the standard RPI calculation and added to the standard RPI values. This ensures that the final Adjusted RPI is not dependent on the order in which bonuses are applied.

The bonuses in the Adjusted RPI are only given to away teams, while the penalties are structured to hurt teams if they lose at home (rather than away). This is intended to reward teams for playing on the road and to factor home-field advantage into the ranking system.

The NCAA notes that the weights of the adjustment are in part calculated

empirically based on the previous seasons to standardize bonuses and penalties with respect to the increment roughly necessary to move a team up one position in the rankings [7]. However, the methodology behind the different performance thresholds for obtaining a bonus or penalty, as well as the relative weighting of thresholds relative to each other, is not detailed. It is safe to assume that these increments are arbitrarily selected.

With this methodology, the Adjusted Generalized RPI then becomes

$$\text{Adjusted Generalized RPI} = \text{Performance} + \text{Context} + \text{A}$$

$$= c_1 \cdot \text{WP} + c_2 \cdot \text{OWP} + c_3 \cdot \text{OOWP} + \text{A}$$

For standardization and interpretability, the cumulative bonus/penalty A will be calculated using the same formula regardless of the weights selected for c_1, c_2 and c_3. An improved A for generalized RPI weights is a topic for further study in future research.

3.5 RPI Bias Intuition

It has been claimed that RPI systematically overvalues teams from stronger conferences who have more opportunities to play stronger teams and therefore increase their strength of schedule [8]. In fact, the idea is so pervasive that it is included on the Wikipedia page for RPI without citation [9].

The mathematical intuition comes from the following. Pairwise rankings should naturally account for strength of schedule (to ensure that teams with harder schedules are not punished), but the official RPI formula places a clear majority of its weight on results outside of a team's control. This weighting makes it so that losses can result in an increase in RPI (if against a good team) and so that winning can decrease a team's RPI (if against a bad team).

This seems counterintuitive; it is justified to assume that a metric designed to evaluate performance should not reward a team when they play poorly.

Counterintuitive examples of RPI are provided below.

3.5.1 Counterintuitive RPI Examples

We provide example cases of counterintuitive RPI adjustments (using unadjusted RPI).

Consider a league with 9 teams: A1, A2, A3, B1, B2, B3, C1, C2, C3. Suppose that lower numbers always beat higher numbers (A1 always beats A2, ex.) and teams closer to the beginning of the alphabet always beat teams closer to the end (A1 always beats B2, ex.). Suppose that teams have played 4 games: 2 against the other teams in their conference, as denoted by letter (A, B, or C) and 2 against the teams that share their number (1, 2 or 3).

The league table after these 4 matches is constructed as shown in Table 1.

Team	**RPI**	**WP**	**OWP**	**OOWP**
A1	0.8125	1.0000	0.8333	0.5833
A2	0.6562	0.7500	0.6667	0.5417
A3	0.5000	0.5000	0.5000	0.5000
B1	0.6562	0.7500	0.6667	0.5417
B2	0.5000	0.5000	0.5000	0.5000
B3	0.3438	0.2500	0.3333	0.4583
C1	0.5000	0.5000	0.5000	0.5000
C2	0.3438	0.2500	0.3333	0.4583
C3	0.1875	0.0000	0.1667	0.4167

Table 1: RPI, WP, OWP, and OOWP for Teams A1–C3

Here, A1 has an RPI of 0.8125 and C3 has an RPI of 0.1875.

Now suppose team A1 plays team C3 and wins. A1's winning percentage does not change; its record remains perfect. However, measures of context drop because of C3's poor record and poor conference strength. (The opposite is true for C3.)

The league table after this match is shown in Table 2.

Team	RPI	WP	OWP	OOWP
A1	0.7167	1.0000	0.6667	0.5333
A2	0.6458	0.7500	0.6667	0.5000
A3	0.5000	0.5000	0.5000	0.5000
B1	0.6458	0.7500	0.6667	0.5000
B2	0.5000	0.5000	0.5000	0.5000
B3	0.3542	0.2500	0.3333	0.5000
C1	0.5000	0.5000	0.5000	0.5000
C2	0.3542	0.2500	0.3333	0.5000
C3	0.2833	0.0000	0.3333	0.4667

Table 2: RPI, WP, OWP, and OOWP for Teams A1–C3 (Updated Dataset)

This table shows that A1 has an RPI of 0.7167 and C3 has an RPI of 0.2833. Thus A1 beat C3 and still dropped in RPI, and vice versa for C3. This is counterintuitive; one would expect a metric to reward teams for winning games and punish them for losing games. A metric as such introduces potentially problematic incentives. For example, A1 would be rewarded for canceling their game against C3 because it would hurt their RPI ranking regardless of result.

There are also scenarios where a team can lose to a team below them in the rankings and still move up. Consider the scenario with teams A1, A2 and A3 in the A conference and B1, B2, and B3 in the B conference. Each team plays every team in its conference once, with the lower number (closer to 1) always winning. Then, out of conference, A1 beats B3, A2 beats B2 and B1 beats A3.

The RPI in this scenario is displayed in Table 3.

Team	RPI	WP	OWP	OOWP
A1	0.5833	1.0000	0.3333	0.6667
A2	0.5556	0.6667	0.5000	0.5556
A3	0.5000	0.0000	0.8333	0.3333
B1	0.5000	1.0000	0.1667	0.6667
B2	0.4444	0.3333	0.5000	0.4444
B3	0.4167	0.0000	0.6667	0.3333

Table 3: RPI, WP, OWP, and OOWP for Teams A1–B3 (Third Dataset)

Now suppose A3 plays B1 again and loses again. The RPI after this game is displayed in Table 4.

Team	**RPI**	**WP**	**OWP**	**OOWP**
A1	0.5868	1.0000	0.3333	0.6806
A2	0.5590	0.6667	0.5000	0.5694
A3	0.5052	0.0000	0.8750	0.2708
B1	0.4948	1.0000	0.1250	0.7292
B2	0.4410	0.3333	0.5000	0.4306
B3	0.4132	0.0000	0.6667	0.3194

Table 4: RPI, WP, OWP, and OOWP for Teams A1–B3 (Fourth Dataset)

Here, A3's RPI has improved from 0.5000 to 0.5052, despite their loss to a team below them in the RPI. Again, this is counterintuitive.

In both of the above examples, note that the source of counterintuitive RPI updates comes from the overweighting of OWP relative to WP. Although positive results help a team's WP, their final RPI rating is more influenced by the WP of the teams they face.

This would be less of an issue if one could assume that all teams play teams with roughly the same average win percentage. This assumption holds in leagues such as the English Premier League, where all teams in the league play each other twice. However, in NCAA Soccer, every team plays only a small fraction of the total teams in the league while playing most or all of the teams in their conference.

D1 Men's Soccer is dominated by two major conferences: the ACC and the Big Ten. These two conferences have most of the teams that traditionally win the NCAA College Cup [10]. It logically follows that teams that play in these stronger conferences have more opportunities to play against stronger teams with higher WPs (and higher OWPs). These opportunities would boost teams' RPIs regardless of match result.

Based on this intuition, it is logical to hypothesize that the RPI could overes-

timate the strength of teams from stronger conferences in pairwise comparison. Research has shown evidence of this kind of bias in NCAA Men's Basketball [11]. However, no academic research has sought to rigorously test RPI conference bias for NCAA D1 Men's Soccer from a formal mathematical standpoint. This undergraduate thesis seeks to change that.

3.6 Bias Counterargument

As shown above, the mathematical intuition behind potential bias towards stronger conferences comes from an overweighting on OWP (and OOWP) relative to WP. In other words, RPI could systematically over-value teams in stronger athletic conferences by placing more of its weight on context and less of its weight on performance. If this hypothesis were true, we would expect to see that RPI reweightings which increase the weight on performance decrease bias towards teams from stronger conferences. The above hypothesis will be tested in later sections.

Before continuing, it should be noted that the distribution of weights for RPI could be misleading. As Thomas notes in his Women's College Soccer blog, a larger variance can be expected for WP when compared to OWP or OOWP [7]. Thomas shows this empirically, but I will demonstrate theoretically. Suppose for generality that, for all teams, WP is standard uniformly distributed and all are independent of each other, i.e.

$$\text{WP} \sim \text{Unif}(0, 1).$$

The variance is then, by definition,

$$\text{Var}(\text{WP}) = 1/12$$

.

Suppose each team plays 16 games. A team's OWP will then be the average of the sum of 16 Unif$(0, 1)$ distributions. Thus OWP is roughly following the Irwin-Hall distribution where $n = 16$.

$$\text{OWP} = 1/16 \cdot \text{IH} = 1/16 \cdot (X_1 + X_2 + ... + X_{16})$$

where

$$X_1, X_2, ..., X_{16} \sim \text{Unif}(0, 1)$$

are independent and identically distributed. The variance is then roughly

$$\text{Var(OWP)} = 1/(16)^2 \cdot \text{Var(IH)}$$

$$= 1/256 \cdot (16 \cdot \text{Var(WP)})$$

$$= 1/16 \cdot 1/12$$

It is trivial to repeat the same process for OOWP by summing OWPs to get

$$\text{Var(OOWP)} = 1/16 \cdot 1/16 \cdot 1/12$$

Thus by these assumptions, although OWP is weighted twice as much as WP in the default RPI formula, we can expect the WP coefficient to quantify 8 times the data variance (or $\sqrt{8} = 2\sqrt{2}$ times as much spread in the data). Following the same logic, we can expect WP to reflect 256(!) times as much data variance as the OOWP component (or $\sqrt{256} = 16$ times as much spread in the data).

This would seem to contradict the initial intuition that the weighting on

WP is too low— if anything, it would appear to capture far more variance in the data than the OWP and OOWP. However, the assumptions made above do not hold in practice. Winning percentages are not uniformly distributed, nor should we expect them to be independent of each other. And varying conference strengths mean that we should expect more variance in OWP and OOWP than if opponents were randomly selected. (In his analysis of Women's Soccer, Thomas shows that the WP component captures roughly 50% of the spread in the data [7].)

Thus, although the 25% weighting of WP in the default RPI formula appears to underweight performance relative to context, counter-arguments can be made. Therefore, a thorough investigation is warranted. This undergraduate thesis will seek to do this through both synthetic and empirical data.

4 Bradley-Terry Model

4.1 Overview

Bradley-Terry models will be used in this undergraduate thesis for two main purposes:

1. To generate synthetic match results for simulation purposes ("Forward Bradley-Terry Model"); and

2. To generate rankings from match results (synthetic or real) ("Reverse Bradley-Terry Model").

The use of Bradley-Terry for (1) is straightforward, but (2) is not. The reverse Bradley-Terry model is intended to be used as a standard pairwise ranking metric for bias evaluation purposes. This means that Bradley-Terry (using Bayesian Inference) will be used to infer team strengths, which will be used in

turn to infer conference strengths. Conference preference in the RPI metric will then be evaluated by examining the correlation between conference strength and the RPI's deviation from these estimated rankings (using rank error, described in the "Bias Evaluation" section).

It should be explained why, in the context of simulations, RPI rankings should be compared to estimated strengths rather than the true strength parameters as generated by (1).

Ranking systems would be straightforward if match results were deterministic, meaning that the better team always won. (Sports probably would not be that fun to watch, either.) However, this is not the case. Teams don't always perform at their strength– they can (and will) overperform or underperform relative to their true abilities. This means that a team's results are only a proxy for their true strength. Ranking systems such as RPI use these performances (as well as context) in an attempt to reconstruct true team strength.

Note, however, that it is impossible to systematically undervalue weaker teams and overvalue stronger teams in a closed ranking system. For every spot a team moves up, a team above it must move down, and vice versa. Overvaluing a strong team means undervaluing a stronger team.

Thus, we should not expect a ranking system to systematically overvalue teams from stronger conferences with respect to their true strength. Stronger conferences will have stronger teams and stronger teams are more likely to underperform than overperform. Thus, we should expect ranking systems to undervalue teams from stronger conferences, and, with them, the conferences themselves.

Thus, to assess bias in a pairwise ranking system, it does not make sense to compare to true strength rankings. Rather, it makes sense to compare rankings to the maximally likely rankings (given the underlying structure of the model).

This is exactly what a Reverse Bradley-Terry Model does, using Bayesian Inference. Thus, the Reverse Bradley-Terry Model will provide a standard pairwise ranking metric for RPI bias evaluation purposes, both in simulations and in real data.

Note that Bradley-Terry is not the singular standard method for pairwise comparison. However, the Bradley-Terry model is commonly used in sports analytics and is a natural fit for this project [12].

4.2 Basic Bradley-Terry Model

The basic Bradley-Terry model gives the probability that team i beats team j as

$$P(\text{i} \quad \text{beats} \quad \text{j}) = \frac{\lambda_i}{\lambda_i + \lambda_j}$$

where λ_i and λ_j represent the respective strength parameters of teams i and j [13].

This provides a way to estimate the probability of a result given the strength coefficients of two teams. This is sufficient for the forward Bradley-Terry model and the generation of data.

However, the reverse Bradley-Terry model must derive team rankings rather than win probabilities. Therefore, one must estimate the distribution of the λ terms. These values will be derived using Bayesian inference. First, a prior distribution is set for the λ terms to reflect initial beliefs about the distribution of the data. Then, given a set of game results, the likelihood function is

$$P(\text{data} \mid \lambda) = \prod_{i<j} \left(\frac{\lambda_i}{\lambda_i + \lambda_j} \right)^{w_{ij}} \left(\frac{\lambda_j}{\lambda_i + \lambda_j} \right)^{w_{ji}}$$

where w_{ij} represents the number of times team i beat team j.

Then, using Bayes' Rule, the posterior distribution is:

$$P(\lambda \mid \text{data}) \propto P(\text{data} \mid \lambda) P(\lambda)$$

Where $P(\lambda)$ can be inferred from the prior.

The Bradley-Terry model can be re-expressed in the form of logits. This has benefits for optimization efficiency and floating point stability [14]. Thus, we will express a base Bradley-Terry model, but with π to represent strength parameters:

$$P(\text{i} \quad \text{beats} \quad \text{j}) = \frac{\pi_i}{\pi_i + \pi_j}$$

Then we apply a logit transformation $\lambda_i = \log(\pi_i)$, resulting in the following probability:

$$P(\text{i} \quad \text{beats} \quad \text{j}) = \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}}$$

The logit transformation

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

then gives us

$$\text{logit}(P(\text{i} \quad \text{beats} \quad \text{j})) = \lambda_i - \lambda_j$$

which allows us to model the relationships between team strengths linearly on the log scale.

The likelihood function then becomes

$$P(\text{data} \mid \lambda) = L(\lambda) = \prod_{i<j} \left(\frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}}\right)^{w_{ij}} \left(\frac{e^{\lambda_j}}{e^{\lambda_i} + e^{\lambda_j}}\right)^{w_{ji}}$$

As before, the posterior distribution can be expressed as

$$P(\lambda \mid \text{data}) \propto P(\text{data} \mid \lambda) P(\lambda)$$

Where $P(\lambda)$ can be inferred from the choice of prior.

Now, for the reverse Bradley-Terry model, the choice of prior must be discussed.

The choice of prior must reflect beliefs about the strengths of the teams in our field of NCAA Men's Soccer teams. It is reasonable to suggest a common prior for all teams in the field. This is because RPI rankings (and selections to the NCAA D1 Championship) are not determined by results from prior seasons.

The question then becomes about a reasonable expectation for the distribution of strengths of NCAA Men's Soccer teams. It is my opinion that the selection of a normal distribution is not unreasonable, especially given the highly-random nature of soccer results.

The normal prior is represented as such:

$$\theta_i \sim \mathcal{N}(\mu, \sigma^2)$$

The prior will set $\mu = 0$ for all teams as a baseline value.

The question then becomes picking a value for σ. Team distribution is intended to be tight (not giving outliers too much sway, given the random nature of soccer) but also not too tight (as we do not want all teams ranked right next to each other). For standardization and interpretability, a value of $\sigma = 1$ is not unreasonable for this context, giving us a standard normal Gaussian prior

$$\theta_i \sim \mathcal{N}(0, 1)$$

Thus, $P(\lambda)$ can be obtained from the probability density function (PDF) of the standard normal distribution:

$$P(\lambda) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\lambda^2}{2}\right)$$

Thus we have a base Bradley-Terry model with a prior selected for our reverse Bradley-Terry model. However, there are necessary extensions to this model that will be covered in the following sections.

4.3 Bradley-Terry-Davidson Model

The Bradley-Terry-Davidson model allows for the inclusion of ties in Bradley-Terry pairwise comparison [15].

It can be written as follows. The probability that team i beats team j is

$$P(i\ beats\ j) = \frac{\lambda_i}{\lambda_i + \lambda_j + \nu\sqrt{\lambda_i\lambda_j}}$$

and the probability that the two teams tie is

$$P(\text{draw}) = \frac{\nu\sqrt{\lambda_i\lambda_j}}{\lambda_i + \lambda_j + \nu\sqrt{\lambda_i\lambda_j}}$$

where ν is the new strength parameter for a tie.

The likelihood function then becomes:

$$P(\text{data} \mid \lambda, \nu) = \prod_{i<j} \left(\frac{\lambda_i}{\lambda_i + \lambda_j + \nu\sqrt{\lambda_i\lambda_j}}\right)^{w_{ij}}$$

$$\left(\frac{\nu\sqrt{\lambda_i\lambda_j}}{\lambda_i + \lambda_j + \nu\sqrt{\lambda_i\lambda_j}}\right)^{t_{ij}} \left(\frac{\lambda_j}{\lambda_i + \lambda_j + \nu\sqrt{\lambda_i\lambda_j}}\right)^{w_{ji}}$$

and the posterior distribution is:

$$P(\lambda, \nu \mid D) \propto L(\lambda, \nu \mid D)P(\lambda)P(\nu)$$

for two priors $P(\lambda)$ and $P(\nu)$.

Again, we will use a log transformation. We have both

$$\theta_i = e^{\lambda_i}, \quad \eta = e^{\nu}$$

Leaving the following equations:

$$P(i\,beats\,j) = \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}$$

$$P(\text{draw}) = \frac{e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}{e^{\lambda_i} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}$$

and likelihood function

$$P(\text{data} \mid \lambda, \nu) = \prod_{i<j} \left(\frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}\right)^{w_{ij}}$$

$$\left(\frac{e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}{e^{\lambda_i} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}\right)^{t_{ij}} \left(\frac{e^{\lambda_j}}{e^{\lambda_i} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}\right)^{w_{ji}}$$

with log

$$\log P(\text{data} \mid \lambda, \nu) = \sum_{i<j} w_{ij} \log\left(\frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}\right)$$

$$+ t_{ij} \log\left(\frac{e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}{e^{\lambda_i} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}\right) + w_{ji} \log\left(\frac{e^{\lambda_j}}{e^{\lambda_i} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}\right)$$

And posterior

$$P(\lambda, \nu \mid D) \propto L(\lambda, \nu \mid D)P(\lambda)P(\nu)$$

for both priors.

Maintaining a standard normal prior for λ, a prior for ν must then be selected. Although empirical data could be examined, the adjusted RPI formula makes no assumptions about the probability of ties. We seek a reverse Bradley-

Terry model that is comparable to RPI. Thus, the selection of a standard normal prior for ν, is intuitive, reasonable, consistent and generalizable. (The effect of different baseline priors in analyses is a subject for further research.)

Thus

$$P(\nu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\nu^2}{2}\right)$$

4.4 Bradley-Terry-Davidson Model With Home Advantage Weights

The value of home-field advantage must be accounted for. To do this, this thesis will use the "common" home-field-adjusted version of Bradley-Terry-Davidson Model as outlined by Jamil (2010) for soccer purposes [12].

Now we have logit transformations

$$\theta_i = e^{\lambda_i}, \quad \eta = e^{\nu}, \quad \gamma = e^{\rho}$$

where ρ is the home advantage parameter. This gives

$$P(i \text{ beats } j | i \text{ at home}) = \frac{e^{\lambda_i} e^{\rho}}{e^{\lambda_i} e^{\rho} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\rho} e^{\lambda_i} e^{\lambda_j}}}$$

$$P(\text{draw} | i \text{ at home}) = \frac{e^{\nu}\sqrt{e^{\rho} e^{\lambda_i} e^{\lambda_j}}}{e^{\lambda_i} e^{\rho} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\rho} e^{\lambda_i} e^{\lambda_j}}}$$

with likelihood function

$$P(\text{data} \mid \lambda, \nu, \rho) =$$

$$\prod_{i<j} \left(\frac{e^{\lambda_i} e^{\rho_i}}{e^{\lambda_i} e^{\rho_i} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\rho} e^{\lambda_i} e^{\lambda_j}}}\right)^{w_{ij}^{home}} \left(\frac{e^{\nu}\sqrt{e^{\rho} e^{\lambda_i} e^{\lambda_j}}}{e^{\lambda_i} e^{\rho_i} + e^{\lambda_j} + e^{\nu}\sqrt{e^{\rho} e^{\lambda_i} e^{\lambda_j}}}\right)^{t_{ij}^{home}}$$

$$\left(\frac{e^{\lambda_j}e^{\rho_j}}{e^{\lambda_j}e^{\rho_j}+e^{\lambda_i}+e^{\nu}\sqrt{e^{\rho}e^{\lambda_i}e^{\lambda_j}}}\right)^{w_{ji}^{home}} \left(\frac{e^{\lambda_i}}{e^{\lambda_i}+e^{\lambda_j}+e^{\nu}\sqrt{e^{\rho}e^{\lambda_i}e^{\lambda_j}}}\right)^{w_{ij}^{neutral}}$$

$$\left(\frac{e^{\nu}\sqrt{e^{\rho}e^{\lambda_i}e^{\lambda_j}}}{e^{\lambda_i}+e^{\lambda_j}+e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}\right)^{t_{ij}^{neutral}} \left(\frac{e^{\lambda_j}}{e^{\lambda_i}+e^{\lambda_j}+e^{\nu}\sqrt{e^{\rho}e^{\lambda_i}e^{\lambda_j}}}\right)^{w_{ji}^{neutral}}$$

(when including neutral-venue games) and log-likelihood function

$$\log P(\text{data} \mid \lambda, \nu, \rho) =$$

$$\sum_{i<j}\left(\frac{e^{\lambda_i}e^{\rho_i}}{e^{\lambda_i}e^{\rho_i}+e^{\lambda_j}+e^{\nu}\sqrt{e^{\rho}e^{\lambda_i}e^{\lambda_j}}}\right)^{w_{ij}^{home}} + \left(\frac{e^{\nu}\sqrt{e^{\rho}e^{\lambda_i}e^{\lambda_j}}}{e^{\lambda_i}e^{\rho_i}+e^{\lambda_j}+e^{\nu}\sqrt{e^{\rho}e^{\lambda_i}e^{\lambda_j}}}\right)^{t_{ij}^{home}}$$

$$+\left(\frac{e^{\lambda_j}e^{\rho_j}}{e^{\lambda_j}e^{\rho_j}+e^{\lambda_i}+e^{\nu}\sqrt{e^{\rho}e^{\lambda_i}e^{\lambda_j}}}\right)^{w_{ji}^{home}} + \left(\frac{e^{\lambda_i}}{e^{\lambda_i}+e^{\lambda_j}+e^{\nu}\sqrt{e^{\rho}e^{\lambda_i}e^{\lambda_j}}}\right)^{w_{ij}^{neutral}}$$

$$+\left(\frac{e^{\nu}\sqrt{e^{\rho}e^{\lambda_i}e^{\lambda_j}}}{e^{\lambda_i}+e^{\lambda_j}+e^{\nu}\sqrt{e^{\lambda_i}e^{\lambda_j}}}\right)^{t_{ij}^{neutral}} + \left(\frac{e^{\lambda_j}}{e^{\lambda_i}+e^{\lambda_j}+e^{\nu}\sqrt{e^{\rho}e^{\lambda_i}e^{\lambda_j}}}\right)^{w_{ji}^{neutral}}$$

with posteriors

$$P(\lambda, \nu, \rho \mid D) \propto L(\lambda, \nu, \rho \mid D)P(\lambda)P(\nu)P(\rho)$$

As before, the selection of an empirically-informed prior can be justified for precision, but is not necessary. This thesis will use a standard normal prior for ρ for simplicity and generalizability. (Note that we should expect some home field advantage, where $\mu > 0$, but to what degree is uncertain). Thus

$$P(\rho) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\rho^2}{2}\right)$$

This is our final model. We will use this version of a reverse Bradley-Terry-Davidson model for comparison with Adjusted RPI.

4.5 Other Bradley-Terry Models

There are other examples of Bradley-Terry models that were considered as baseline ranking systems for this thesis but which were eventually rejected. One example is a Bradley-Terry-Davidson model which accounts for match scores. Although match scores could be used to better quantify a team's strength, they are not considered by the selection committee. This thesis does not seek to factor metrics that are outside the committee's selection criteria. Thus, the score will be ignored in this thesis's forward and reverse Bradley-Terry-Davidson models.

The same logic can be applied to most other extensions of the Bradley-Terry model. The effect of different Bradley-Terry models on the results of this thesis is an opportunity for further research.

5 Bias Measurement

In order to measure "bias" in a ranking system, it must first be defined.

Note that it is not enough to define bias as a systematic preference for teams from stronger conferences. In fact, one should expect teams from stronger conferences to have stronger rankings. One expects them to perform better; they are expected to win more games. In other words, we expect team rankings to generally predict the results of matches. Thus, this thesis defines bias to be a systematic preference for teams from stronger conferences *at the expense of predictive accuracy.*

This thesis will evaluate conference preference by comparing the "rank error" by conference relative to a baseline ranking. The rank error will represent the average difference in each conference team's rank when compared to the baseline ranking.

The rank error formula is

$$\text{Conference Rank Error} = -\frac{\sum_{i=1}^{N}(R_i - B_i)}{N}$$

where

- R_i is the RPI ranking of conference team i;
- B_i is the baseline ranking of conference team i; and
- N is the number of teams in the conference

Note the the negative sign at the front of the equation; this ensures that positive values reflect an overvaluation by the RPI and negative values reflect an undervaluation, for interpretability.

A simple example is provided in Table 5 below. Suppose there are four teams: A1, A2, B1 and B2. Suppose that the column on the left is the maximally likely team rankings generated by the reverse Bradley-Terry-Davidson model. The column on the right then represents the projected ranking of the teams using the standard RPI metric. In the RPI rankings, A2 is ranked below B1, even though it is ranked above it by the Bradley-Terry-Davidson model. A2 then is undervalued with respect to its Bradley-Terry-Davidson ranking by 1 place, while B1 is overvalued by 1 place. The average rank error by conference (relative to the Bradley-Terry-Davidson ranking) would then be -1/2 = -0.5 points for Conference A and 1/2 = 0.5 for Conference B.

Table 5: Bradley-Terry-Davidson (BTDM) vs. RPI Rankings and Conference Rank Error

Team	Conference	BTDM Rank	RPI Rank	Rank Error	Note
A1	A	1	1	0	–
A2	A	2	3	-1	Undervalued
B1	B	3	2	+1	Overvalued
B2	B	4	4	0	–

Predictive accuracy will be computed by measuring, for the games that did not end in a tie, what percentage of results were predicted by the rankings. In other words, accuracy will reflect the percentage of decisive games that the higher-ranked team win. Bias, then, will occur when the rating system prefers teams from stronger conferences at the expense of accuracy.

I must motivate the selection of accuracy metric to evaluate predictive power. This metric was chosen because it is simple to compute and easy to interpret. In addition, it is intuitive to understand; a better ranking metric should, in theory, be able to predict game results more often.

It can be argued that accuracy might be too simple of a metric to quantify the predictive nature of something as complex as a pairwise ranking system. For example, a team which only plays against teams at the top and the bottom of the rankings could be ranked anywhere in between without affecting the computed accuracy. However, in the context of repeated simulations over time, scenarios such as this one should not be expected to occur very often, and one should generally expect large swings in a team's ranking to correspond to swings in accuracy. Thus, one should expect biases to negatively impact predictive accuracy in simulations almost surely with enough simulations.

6 Data

Data for all matches during the 2024 NCAA D1 Men's Soccer season has been scraped from the NCAA Men's Soccer Scoreboard using henrygd's NCAA API tool on Github [16][17]. Data was cleaned using Python and saved in a .csv format.

Data was cleaned using methods which created distinctions between the data used for model evaluation and the data used by the NCAA for RPI calculation. These include the following:

- Teams are allowed to play non-Division 1 teams as part of their season. Although the results of these matches do not affect a team's RPI, the NCAA penalizes a team's RPI for playing multiple matches against non-D1 opposition, in an attempt to encourage D1 competition. For interpretability, games not played against D1 opposition were removed from the dataset, which also removed the RPI penalty.

- The NCAA calculates the "adjustment" in Adjusted RPI differently depending on if a game was played in a neutral location or not. However, the API from which results could be scraped did not include this information. On further research I found conflicting information on the neutral site status of certain games (and sometimes no location information at all). Therefore, for ease of interpretability, a game was labeled as a neutral site if the location of the game differed from the location where the team typically plays. This provided a strong (but not exact) proxy for neutral site status without access to official neutral site data.

It is primarily for these factors that the Adjusted RPI calculations are very close but not exact matches to the official NCAA RPIs.

Although this thesis is motivated to investigate official NCAA Men's Soccer

results, the limited amount of season data motivates the creation of synthetic data. Thus, this thesis will begin with an analysis of simulated season data, using a Bradley-Terry-Davidson Model. This allows for more data from which the chosen methods can be tested. The methodology behind the creation of simulated data will be detailed in the following sections.

7 RPI Reweighting Selection

7.1 Simulation Overview

To test the accuracy of potential RPI reweightings, various 1000-season simulations were conducted with the following general methodology:

- Team strengths are randomly selected for each season.
- Match results are generated using a Bradley-Terry-Davidson model with a home advantage parameter.
- The predictive accuracy of every possible RPI reweighting (in which c_1, c_2 and c_3 are multiples of 0.05 and add to 1) is then measured and averaged across the seasons.
- Reweightings with the highest average accuracy are outputted.

This methodology was applied across six different sets of simulations on the Yale High-Performance Computer clusters, varying:

- Balanced vs. imbalanced baseline conference strengths
- No home advantage vs. home advantage
- Synthetic teams/matchups vs. teams/matchups from the 2024 season

Each will be described below and examined in further detail.

The code used for this undergraduate thesis is uploaded to this Github repository. The link can also be found in the References section [18].

7.2 Procedures and Results

6 rounds of simulations were completed.

Round 1 was completed using the following procedures:

First, a synthetic list of teams and matchups was created, using the following criteria:

- 22 conferences were created (letters A-V).
- 9 teams were created for each conference (A1-A9 for conference A, B1-B9 for conference B, and so on).

Matchups were then created in the following manner:

- First, a round of in-conference games was created deterministically (4 home, 4 away for each team)
- Then, all teams were randomly paired into 8 rounds of non-conference matchups, leaving a total of 16 games and 50 percent that were in conference.

This roughly simulates the structure of D1 college soccer, with 22 conferences and 212 teams that play 16-20 games in a season, with roughly half of those being conference games (depending on the conference).

Team strengths were generated using the following methods:

- Each team was given an underlying strength parameter α generated from the standard normal distribution.

- The α values are then centered ($\alpha = \alpha - \bar{\alpha}$) so that the mean α value is 0 (for standardization).

Now match results were generated using the following methodology:

- Results were simulated using a logit-adjusted Bradley-Terry-Davidson model with the logits of each α as the underlying team strengths.
- The tie coefficient γ as well as the home advantage coefficient ρ were both set to 0 (meaning no home advantage).

In addition, for each simulated season of match results, the following procedure was used to evaluate each combination of weights in the non-adjusted generalized RPI formula ($c_1 \cdot \text{WP} + c_2 \cdot \text{OWP} + c_3 \cdot \text{OOWP}$):

- c_1, c_2 and c_3 were all multiples of 1/20, i.e. elements in the set

  (0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5,0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1)

  and were all such that $c_1 + c_2 + c_3 = 1$.
- Given c_1, c_2, and c_3, every team's reweighted RPI was calculated for the simulated season.
- Teams were then ranked from highest to lowest to provide a descending ranking of reweighted team RPIs.
- Match results were then predicted by choosing the team with the higher RPI.
- The predictive accuracy of the RPI reweighting for all decisive (non-tied) games was then recorded.

All simulations were conducted in python. This project also used CMDStan in python to process Stan code. All simulations were run on 4 16Gb cores on the High Performance Computer clusters at Yale University.

The results for this simulation can be seen in **Table 10** in the Appendix.

Round 2 was completed using the same procedures as Round 1, but with the following differences:

- Each team was given an underlying strength parameter α generated from the standard normal distribution *plus* an underlying conference strength parameter generated from the standard normal distribution. This second parameter was consistent for all teams in the conference.

The results for this simulation can be seen in **Table 11** in the Appendix.

Round 3 was completed using the same procedures as Round 1, but with the following differences:

- Results were simulated such that the home strength parameter was increased to $\rho = \ln(2)$, rather than 0.
- To adjust for this, Adjusted Generalized RPI was used (rather than Standard Generalized RPI), and the Bradley-Terry-Davidson model for Bayesian Inference was modified to include a weight for home advantage.

The results for this simulation can be seen in **Table 12** in the Appendix.

Round 4 was completed using the same procedures as Round 2, but with the same differences as Round 3:

- Results were simulated such that the home strength parameter was increased to $\rho = \ln(2)$, rather than 0.
- To adjust for this, Adjusted Generalized RPI was used (rather than Standard Generalized RPI), and the Bradley-Terry-Davidson model for Bayesian Inference was modified to include a weight for home advantage.

The results for this simulation can be seen in **Table 13** in the Appendix.

Round 5 was completed using the same procedures as Round 3, but with the following differences:

- Rather than generating synthetic teams and matchups for each of the 1000 simulations, the teams and matchups scraped from the 2024 NCAA D1 Men's Soccer season were used and held consistent across simulations.

The results for this simulation can be seen in **Table 14** in the Appendix.

Round 6 was completed using the same procedures as Round 4, but with the same differences as Round 5:

- Rather than generating synthetic teams and matchups for each of the 1000 simulations, the teams and matchups scraped from the 2024 NCAA D1 Men's Soccer season were used and held consistent across simulations.

The results for this simulation can be seen in **Table 15** in the Appendix.

7.3 Analysis of Results

The most accurate weightings for each of the simulations are summarized in the table below.

Simulation	Triple	Average Accuracy
1	(0.25, 0.3, 0.45)	0.785428
2	(0.25, 0.3, 0.45)	0.808670
3	(0.45, 0.4, 0.15)	0.767450
4	(0.3, 0.25, 0.45)	0.791365
5	(0.45, 0.4, 0.15)	0.764676
6	(0.3, 0.3, 0.4)	0.785131

Table 6: Top-performing RPI reweightings by average accuracy.

In all six rounds of simulations, the most accurate reweighting of RPI weighed performance (c_1) at or above 25 percent. In fact, the only simulations in which

the highest accuracy RPI reweighting had $c_1 = 0.25$ were the first and second simulations, in which home advantage was not factored into the result generation. For the rest of simulations, which adjusted for home advantage and utilized Adjusted RPI as used by the NCAA, the most accurate rankings weighted $c_1 > 0.25$. This is an encouraging result which aligns with the initial hypothesis that the RPI could undervalue performance and overvalue contextual factors such as conference strength at the expense of predictive accuracy.

For all six rounds of simulation, the highest-accuracy RPI reweighting outperformed the standard RPI weighting ($c_1 = c_3 = 0.25, c_2 = 0.5$). Notably, this was the case even for Rounds 1 and 2, where $c_1 = 0.25$. This suggests that, even without changing the weighting on performance, a reweighting simply of the conference parameters c_2 and c_3 can lead to an improvement in predictive accuracy.

As computed in the "Generalized RPI" section, we should expect, with some strong assumptions, for the weighting for WP to encode more spread in the data that OWP or OOWP. However, in those computations, we must note that we also showed that we should expect OWP to encode far more spread in the data than OOWP:

$$\text{Var}(\text{OOWP}) = 1/16 \cdot \text{Var}(\text{OWP})$$

And thus

$$\text{SD}(\text{OOWP}) = 1/4 \cdot \text{SD}(\text{OWP})$$

With this context, results from Rounds 1 and 2 suggest that the weighting of $c_2 = 0.5$ and $c_3 = 0.25$ is biased towards OWP at the expense of predictive accuracy. An increase in c_3 relative to c_2 should intuitively help balance the

variances for OWP and OOWP. Supporting this theory, more than half of the simulations had the most accurate RPI reweighting with $c_3 > c_2$. (The only exceptions are the simulations which allow for both home advantage and balanced conferences.) The relationship between c_2 and c_3 will continue to be explored in later sections.

For Rounds 3-4 and Rounds 5-6, note the decrease in c_1 for the most accurate RPI reweighting. This is intuitively because, when conference strengths are varied, the context variables OWP and OOWP serving as proxies for opposition strength (and with it a team's conference strength) are associated the true strength of the team. Thus, when conference strengths are imbalanced, one should expect to weigh team performance c_1 less, which is exactly what is found.

With a stronger metric for conference strength, more information about a team's strength can be gleaned just by looking at their matchups. This means that more accurate predictions should occur in simulations for imbalanced conferences. This is exactly what one finds in the simulation results; there is an increase in the average accuracy for simulations with imbalanced conferences.

8 Bias Evaluation Simulations

The previous results have shown that reweightings of RPI can lead to increases in predictive accuracy for match results generated in a variety of simulations.

The next step is to examine the relationship between accuracy and conference bias for a selection of these RPI reweightings. The following simulations seek to explore this relationship.

8.1 Simulation Overview

Six RPI reweightings from the previous simulations were selected for further analysis. Selections were made to prioritize predictive accuracy while also en-

suring a diverse sample of reweightings to examine. The six weightings selected were the following:

- $c_1 = 0.25, c_2 = 0.3, c_3 = 0.45$
- $c_1 = 0.3, c_2 = 0.25, c_3 = 0.45$
- $c_1 = 0.4, c_2 = 0.3, c_3 = 0.4$
- $c_1 = 0.35, c_2 = 0.35, c_3 = 0.3$
- $c_1 = 0.4, c_2 = 0.35, c_3 = 0.25$
- $c_1 = 0.45, c_2 = 0.4, c_3 = 0.15$

Further analysis was then performed on these RPI weightings as well as the original weightings ($c_1 = 0.25, c_2 = 0.5, c_3 = 0.25$) in order to assess accuracy and conference-level bias with respect to a reverse Bradley-Terry-Davidson model. This was done through various sets of 1000-season simulations, conducted with the following general methodology:

- First, team strengths and matchups were randomly selected and kept consistent across seasons.
- For each season, match results are randomly generated using a forward Bradley-Terry-Davidson model with a home advantage parameter.
- Bayesian inference was then used with the reverse Bradley-Terry-Davidson model to create a ranking of the maximally likely team strengths.
- For each RPI weighting, team rankings were calculated.
- Then, the rank error relative to the predictive Bradley-Terry rankings was computed for each team and averaged for each conference.

- Rank error was then compared to the true underlying strength parameters of each conference (averaged by team) and correlation was calculated.
- The predictive accuracy of the rankings for each RPI weighting was also compared to that for the reverse Bradley-Terry-Davidson rankings.

This methodology was applied across 3 rounds of 6 simulation sets. As before, simulations were designed to vary conference balance, home advantage and matchup selection.

8.2 Procedures and Results

6 rounds of 3 sets of simulations were completed.

Round 1 was completed using the following procedures for each set of simulations:

First, the teams and matchups scraped from the 2024 NCAA D1 Men's Soccer season were used and held consistent across simulations.

Team and conference strengths were generated using the following methods:

- Each team was given an underlying strength parameter α generated from the standard normal distribution.
- The α values are then centered so that the mean α value is 0.
- Conference strengths were then calculated by averaging the alpha values of all of the teams in the conference.

Now, from there, 1000 seasons of match results were generated. Unlike in the "RPI Reweighting" section, matchups, team strengths and conference strengths were *consistent* within each set of simulations. Data variability came only from the variance in match results. Match results were generated using the following methodology:

- Results were simulated using a logit-adjusted Bradley-Terry-Davidson model with the logits of the αs as the underlying team strengths.
- The tie coefficient γ as well as the home advantage coefficient ρ were both set to 0 (meaning no home advantage).

Then, baseline rankings were generated with a reverse Bradley-Terry-Davidson model using the following methodology:

- A simple Bradley-Terry-Davidson model was used with the form

$$P(i\,beats\,j) = \frac{\lambda_i}{\lambda_i + \lambda_j + \nu\sqrt{\lambda_i \lambda_j}}$$

- Bayesian inference was then performed with Stan using Monte Carlo Markov Chains (MCMC) to estimate the posterior distribution for each team's α. The default settings for Bayesian inference under CMDStan were used. This means:
  - 4 chains
  - 1000 warmup iterations
  - 1000 sampling iterations
- The mean of the α samples across the 1000 simulations was then calculated as $\bar{\alpha}$.
- The $\bar{\alpha}$s were then ranked highest to lowest to create a descending ranking of predicted team strengths.

As mentioned above, the following RPI weightings were selected for analysis:

- $c_1 = 0.25, c_2 = 0.5, c_3 = 0.25$ (default weighting)

- $c_1 = 0.25, c_2 = 0.3, c_3 = 0.45$
- $c_1 = 0.3, c_2 = 0.25, c_3 = 0.45$
- $c_1 = 0.4, c_2 = 0.3, c_3 = 0.4$
- $c_1 = 0.35, c_2 = 0.35, c_3 = 0.3$
- $c_1 = 0.4, c_2 = 0.35, c_3 = 0.25$
- $c_1 = 0.45, c_2 = 0.4, c_3 = 0.15$

The following procedure was used to evaluate each combination of weights in the non-adjusted generalized RPI formula:

- Given c_1, c_2, and c_3, every team's reweighted RPI was calculated for the simulated season.
- Teams were then ranked from highest to lowest to provide a descending ranking of reweighted team RPIs.

Averages across the 1000 simulations were then calculated for the following metrics:

- Rank Error for the RPI weighting relative to the reverse Bradley-Terry-Davidson rankings (using the Rank Error methodology as described in the "Bias Evalution" section);
- Correlation coefficient between ground truth conference strength and the above rank error;
- Rank Error for the RPI weighting relative to the ground truth rankings (for comparison);
- Correlation coefficient between ground truth conference strength and the above rank error (for comparison); and

- Average predictive accuracy of the rankings.

The results across the three sets of simulation can be seen in **Tables 16-18** in the Appendix.

Round 2 was completed using the same procedures as Round 1, but with the following differences:

- Each team was given an underlying strength parameter α generated from the standard normal distribution *plus* an underlying conference strength parameter generated from the standard normal distribution. This second parameter was consistent for all teams in the conference.

The results for this simulation can be seen in **Tables 19-21** in the Appendix.

Round 3 was completed using the same procedures as Round 1, but with the following differences:

- Instead of using matchups from the 2024 season, a synthetic list of teams and matchups was created, using the same criteria as before:
  - 22 conferences were created.
  - 9 teams were created for each conference.
- Matchups were then created in the same manner as before, to roughly simulate the structure of college soccer:
  - First, a round of in-conference games was created deterministically (4 home, 4 away for each team)
  - Then, all teams were randomly paired into 8 rounds of non-conference matchups, leaving a total of 16 games and 50 percent that were in conference.
- Results were simulated such that the home strength parameter was increased to $\rho = \ln(2)$, rather than 0.

- Adjusted Generalized RPI was used instead of Standard Generalized RPI.
- The Bradley-Terry-Davidson model for Bayesian Inference was modified to include a weight for home advantage.

The results for this simulation can be seen in **Tables 22-24** in the Appendix.

Round 4 was completed using the same procedures as Round 2, but with the same differences as Round 3:

- Instead of using matchups from the 2024 season, a synthetic list of teams and matchups was created, using the same criteria as before:
  - 22 conferences were created.
  - 9 teams were created for each conference.
- Matchups were then created in the same manner as before, to roughly simulate the structure of college soccer:
  - First, a round of in-conference games was created deterministically (4 home, 4 away for each team)
  - Then, all teams were randomly paired into 8 rounds of non-conference matchups, leaving a total of 16 games and 50 percent that were in conference.
- Results were simulated such that the home strength parameter was increased to $\rho = \ln(2)$, rather than 0.
- Adjusted Generalized RPI was used instead of Standard Generalized RPI.
- The Bradley-Terry-Davidson model for Bayesian Inference was modified to include a weight for home advantage.

The results for this simulation can be seen in **Tables 25-27** in the Appendix.

Round 5 was completed using the same procedures as Round 3, but with the following differences:

- Instead of using synthetic data, the teams and matchups scraped from the 2024 NCAA D1 Men's Soccer season were used and held consistent across simulations.

The results for this simulation can be seen in **Tables 28-30** in the Appendix.

Round 6 was completed using the same procedures as Round 4, but with the same differences as Round 5:

- Instead of using synthetic data, the teams and matchups scraped from the 2024 NCAA D1 Men's Soccer season were used and held consistent across simulations.

The results for this simulation can be seen in **Tables 31-33** in the appendix.

8.3 Analysis of Results

In every simulation, RPI ($c_1 = 0.25$, $c_2 = 0.5$, $c_3 = 0.25$) rank error by conference (with respect to the reverse Bradley-Terry rankings) was positively correlated to conference strength. In other words, the default RPI weighting ranked teams from stronger conferences higher than the reverse Bradley-Terry-Davidson model. In addition, each proposed RPI reweighting (each with higher c_1) lowered this correlation while improving accuracy. This clearly shows a bias from the default $c_1 = 0.25, c_2 = 0.5, c_3 = 0.25$ RPI weighting towards teams from stronger conferences at the expense of accuracy in the context of these simulations.

For every RPI reweighting in every round of simulations, the rank error by conference with respect to the true underlying team rankings was *negatively* correlated to conference strength with $p < 0.05$. In other words, all RPI reweight-

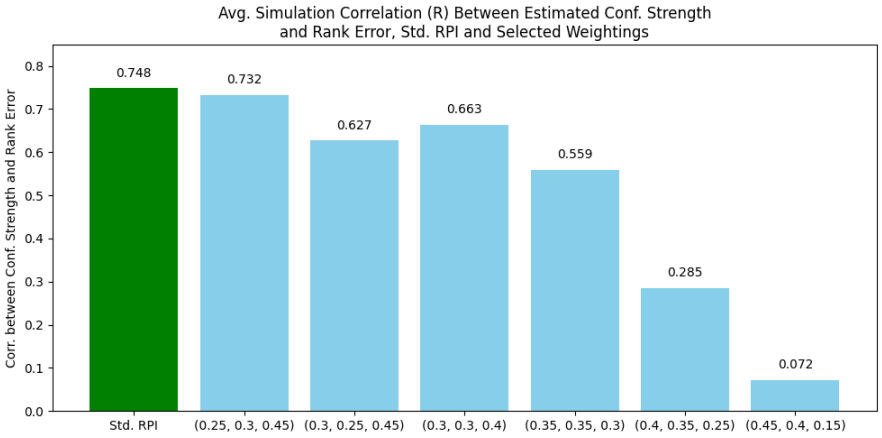

Figure 1: Average Simulation Conf. Strength Correlations

ings tended to rank teams from stronger conferences *below* their ground truth strength. Although this negative correlation seems to contradict the results in the previous paragraph, this behavior is to be expected. As explained in the motivation behind the usage of the reverse Bradley-Terry-Davidson model, teams will not perform to their true strength through a season, and we expect better teams to underperform and worse teams to overperform on average in a closed system. Thus, we expect teams from better conferences to be undervalued and teams from worse conferences to be overvalued. Thus, this result confirms our expectations and should not be taken to contradict the findings in the previous paragraph.

We can aggregate results using averages to display graphics for generalizability. These are displayed in Figures 1-2.

Note that all of the RPI reweightings reduced the average correlation between conference strength and rank error when compared to the default RPI weighting $c_1 = 0.25, c_2 = 0.5, c_3 = 0.25$. In addition, the average rank error correlation to conference strengths decreased as c_1 increased. This supports the hypothesis that conference bias in RPI rankings is a consequence of under-

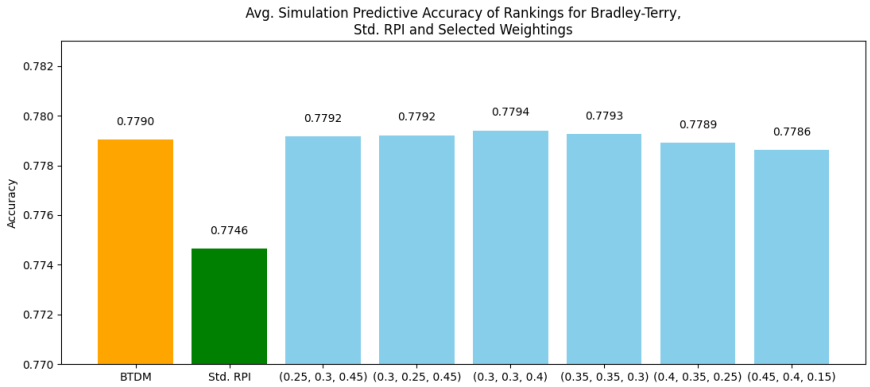

Figure 2: Average Simulation Accuracies

weighting performance.

It must be noted that not all of the reweightings had $c_1 > 0.25$. In the case of the $c_1 = 0.25$, $c_2 = 0.3$ and $c_3 = 0.45$ weighting, average accuracy significantly improved (and the average rank error v. conference correlation slightly decreased) just by increasing c_3 and decreasing c_2. This is consistent with the results of the RPI reweighting simulations in the previous section. This suggests again that an increase in c_3 with respect to c_2 provides a better measure of context.

This means that the original RPI weights could be misleading if a team's opponents play weaker than average (or stronger than average) opposition. In other words, the original context weighting of $c_2 = 0.5$ and $c_3 = 0.25$ underweights the strength of schedule of a team's opponents, leading to diminished accuracy and a slight increase in conference bias. Further analysis is needed to see if this pattern holds for a variety of weightings of c_2 and c_3 when holding c_1 constant. Further analysis is also needed to see the effects of c_2 and c_3 reweightings on the correlation between rank error and conference strength.

In sum, these results conclusively show that, across the simulations, the RPI reweightings significantly outperform the standard RPI metric in both predictive

accuracy and conference rank error (relative to reverse Bradley-Terry-Davidson rankings as the benchmark comparison). Based on these findings, it is safe to conclude that, in the context of these simulations:

- the default RPI weights shows bias towards stronger conferences at the expense of accuracy;
- RPI reweightings which emphasize performance (WP) and deemphasize context (OWP + OOWP) reduce this bias; and
- An RPI reweighting which emphasizes OOWP and deemphasizes OWP increases accuracy without increasing the conference strength rank error correlation.

These conclusions motivate the next section, which tests whether these patterns hold with real data from the 2024 NCAA season.

9 Application to 2024 Season

9.1 Overview

As shown in the previous section, the selected RPI reweightings (which emphasize WP and/or OOWP relative to OWP) increase predictive accuracy while decreasing bias with respect to conference strength in simulations. Thus, we are motivated to examine the results for the 2024 season, and run similar tests, to see if the patterns and relationships hold in the real world.

9.2 Procedure and Results

Similar methodology as the previous simulations was used, but this time incorporating real data. First, teams, matchups *and results* were loaded from the

2024 NCAA D1 season. Then, estimated conference strengths were generated using the following methodology:

- Team strengths were estimated with a reverse Bradley-Terry-Davidson model with a home advantage weight (using Bayesian inference with the same methodology as previous simulations).
- Then, estimated conference strengths were calculated as the mean of the estimated team strengths for each conference.

Then, for each of the RPI weightings used in "Bias Evaluation in Simulations," the following metrics were calculated:

- Rank Error relative to the reverse Bradley-Terry-Davidson rankings
- Correlation coefficient between *estimated* conference strength (from the *reverse* Bradley-Terry-Davidson model) and the above rank error
- Average predictive accuracy of the rankings

Rank error, correlation coefficient and predictive accuracy were also computed for all other possible RPI reweightings (following the reweighting rules specified in the "RPI Reweighting" section). Results were outputted for the top 5 performing weights.

The results can be seen in **Table 34** in the Appendix.

9.3 Analysis of Results

As in the previous simulations, conference rank error was positively correlated to conference strength with $p < 0.05$ for the default RPI weighting ($c_1 = 0.25$, $c_2 = 0.5$, $c_3 = 0.25$). However, although the selected RPI reweightings lowered this correlation, they also reduced predictive accuracy. In fact, the default RPI weighting also significantly outperformed the reverse Bradley-Terry-Davidson

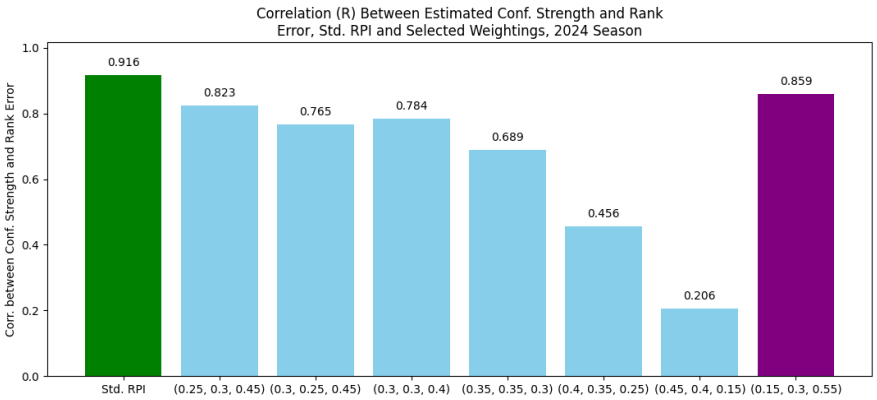

Figure 3: 2024 Season Conf. Strength Correlations

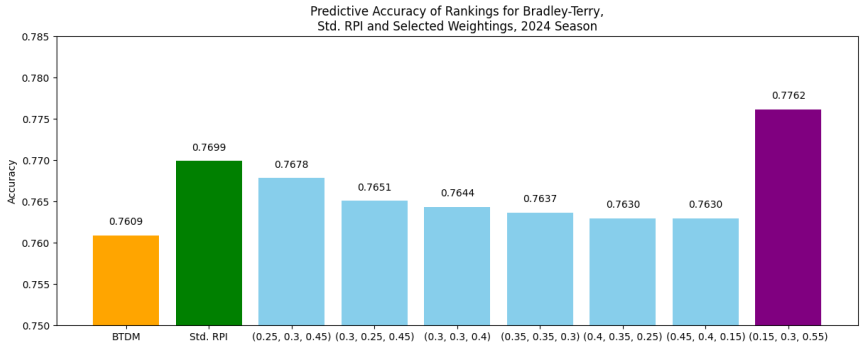

Figure 4: 2024 Season Accuracies

model from an accuracy standpoint. In addition, the highest-accuracy RPI reweighting for the 2024 NCAA D1 season was $c_1 = 0.15$, $c_2 = 0.3$, $c_3 = 0.55$. This weighting underweights performance ($c_1 < 0.25$). These results are shown in Figures 3-4.

These results suggest that, in the context of the 2024 season, the default RPI weighting favors teams from stronger conferences but simultaneously remains more predictive of match results when compared to a baseline reverse Bradley-Terry model. This contradicts the results of the previous simulations. An explanation is necessary.

The use of Bayesian inference with the Bradley-Terry-Davidson model is designed to determine the maximally likely team strengths based on match performances. However, it is outperformed from an accuracy perspective by the default RPI reweightings. There are two main explanations for this. Either

- Bradley-Terry-Davidson does not produce the optimal team strengths (and the RPI is a better model); or
- team strengths are not the only predictive factor for match results.

Both theories will be explored.

There are a host of reasons why the reverse Bradley-Terry-Davidson model would not produce the best approximation of team strengths. One reason would be that some of the assumptions of the model do not hold. For example, the assumption of transitivity between team strengths might not hold; if Team A is better than Team B in a head-to-head matchup, and Team B is better than Team C, that might not mean that Team A is better than Team C head-to-head. Or, the assumption that ν and ρ are constant across teams could be challenged. However, it must be noted that *the RPI metric makes similar assumptions.* It uses transitive rankings to compare teams and it weighs ties and home advantage equally for all teams. Thus, it is unlikely that this is the issue.

Concerns specific to the Bradley-Terry-Davidson model and Bayesian inference must also be addressed. For example, a poor choice of priors could lead to poor performance and imprecise or inaccurate results. Or perhaps a different pairwise ranking model with Bayesian inference could improve accuracy. Experimentation with different priors, models, or other factors are topics for further analysis. However, one must note that *it is not unreasonable to assume that the reverse Bradley-Terry-Davidson model should give good strength approximations for the 2024 season.* The model is an advanced machine-learning algorithm

which performed well throughout the simulations. It should outperform an arbitrary weighting system such as that used by the default Adjusted RPI metric. This would imply that the metric of accuracy for evaluating performance would be misleading. Let us now turns to this second hypothesis.

The usage of the predictive accuracy metric is motivated by the idea that results are primarily driven by team strengths. However, other factors contribute to match results that are not correlated to team strengths. For example, results are also driven by which team is home and which team is away; the team at home should benefit from some home-field advantage.

Because team and conference strengths in the synthetic data were chosen randomly and independently from the same distribution, we could assume that there was little correlation between underlying team or conference strength and home vs. away games. However, this assumption need not hold in the context of the 2024 season. Stronger conferences tend to have teams with larger budgets. These teams are often willing to pay smaller teams to come play them. Thus, it is not unreasonable to assume that teams from stronger conferences tend to play at home more due to their bigger budgets, even if the teams themselves aren't as strong!

This point will be illustrated with an example. Suppose Team B is a slightly better team than Team A. If Team B and Team A play all of the same games against the same opposition in the same locations, we should expect Team B to slightly outperform Team A over time. Now suppose Team A is from a larger conference, and has the resources to incentivize Team B to play them at home. Suppose that with home-field advantage, we can expect Team A to beat Team B more often than we can expect Team B to beat Team A. In this situation, a ranking system which ranks Team B above Team A should be rewarded because it accounts for home-field advantage. However, the metric of accuracy would

reward a ranking system that ranks Team A over Team B simply because Team A gets to play at home.

Thus, it can be argued that the use of accuracy in evaluating ranking systems is biased towards ranking systems that do not sufficiently adjust for home advantage when home matches are not evenly distributed. These ranking systems will be rewarded for overvaluing weaker teams that tend to play at home more often. These teams could tend to be weaker teams from stronger conferences.

Although the Adjusted RPI includes a bonus or penalty based on home and away form, the adjustments are minimal, arbitrary and only applied in certain scenarios. Thus it can be argued from a glance that RPI rankings do not weigh enough for home versus away matches.

To conclude, it may be the case that away teams for non-conference matchups tend to be from weaker conferences and disproportionately play away to teams from stronger conferences. In this scenario, overvaluing teams from stronger conferences would lead to higher predictive accuracy (because they tend to win more games) but would not be a reflection of their true strength (because they get the unfair benefit of being able to play more of their games at home). In this scenario, then, the RPI metric would be biased towards teams from stronger conferences, not because of an overweighting of context, but because of an underweighting of home advantage.

The way to test and adjust for this is to:

1. confirm that teams from stronger conferences tend to play at home more often (or to confirm that home matches are not evenly distributed); and

2. choose a metric beside predictive accuracy for evaluation of model efficacy which quantifies home-field advantage.

To confirm (1), the average home game percentage for each conference was calculated. Correlation analyses were performed to identify the relationship

between this metric and estimated conference strength as well as conference rank error for the default RPI rankings. The results are shown below.

Metric	**Correlation Coeff. (R)**
Home game % vs. Conference strength	**0.7713*** **
Home game % vs. RPI rank error	**0.86*** **

Table 7: Correlation coefficients involving team home game percentage

These results demonstrate a strong positive correlation between conference strength and percentage of games played at home. In fact, 77% of the variance in home game percentage is explained by variance in conference strength.

These are results that are inconsistent with the synthetic data. For that data, ground truth conference strengths and match locations were generated randomly and independently, meaning the expected correlation was zero. Thus one can quickly assume the correlation between home game percentage and conference strength to be roughly zero by the Law of Large Numbers.

Thus, the distribution of synthetic data in our simulations does not match the distribution of 2024 season data. Therefore, the usage of the accuracy metric is likely flawed in the context of empirical data. The distribution of home games is not independent of conference, meaning that teams from stronger conferences, regardless of true strength, can reap the advantages of playing more games at home.

The reason why teams from stronger conferences would want to play more home games should be elaborated. It would appear that teams have figured out the inadequate home advantage adjustment in Adjusted RPI and are looking to schedule more home games to boost their RPIs. This is an inefficiency which benefits the wealthier teams from stronger conferences more. In the open market of NCAA Men's Soccer scheduling, larger teams can use their financial resources to pay teams to play away from home. (This practice is already common in NCAA football and basketball, where it gets more public exposure [19].) In

this way, larger teams can use the benefits of home-field advantage to boost their win percentages and RPIs at the expense of stronger teams with smaller budgets. Supporting this hypothesis, Table 7 shows that 86% of the variance in home game percentage is explained by rank error in the RPI when compared to Bradley-Terry-Davidson estimated true rankings. In other words, teams which are overvalued by RPI (with respect to the Bradley-Terry-Davidson model) tend to play more games at home.

With this in mind, the results of this analysis of the 2024 season should not be held in contradiction to the results from the simulations. The variation in home match distribution is a colinearity influencing the accuracy metric in the 2024 season analysis but not in the simulation analyses.

To rephrase, the previous sections have shown that, in simulations with a roughly even distribution of home matches, the RPI metric is biased towards teams from stronger conferences by underweighting performance relative to context. This section only shows that, in empirical data, an unequal distribution of home-field advantage by conference complicates this analysis. Without a better metric for evaluating model efficacy, this paper should not use predictive accuracy to draw conclusions about the underweighting of performance relative to context from the 2024 season data.

There is motivation to explore a different metric for evaluation of model efficacy which can quantify home vs. away bias. However, it is outside of the scope of this undergraduate thesis, which seeks only to determine whether an underweighting of performance in RPI leads to conference-level bias. This remains a subject for further research.

We will now examine the highest-accuracy RPI reweighting $c_1 = 0.15$, $c_2 = 0.3$, $c_3 = 0.55$ for the 2024 season.

(Note that the reweighting $c_1 = 0.2$, $c_2 = 0.45$, $c_3 = 0.35$ had a matching

accuracy. The weighting $c_1 = 0.15$, $c_2 = 0.3$, $c_3 = 0.55$ was chosen for its marginally smaller correlation with conference strength. However, most of the findings in the following paragraphs will generally hold for $c_1 = 0.2$, $c_2 = 0.45$, $c_3 = 0.35$)

The weighting $c_1 = 0.15$, $c_2 = 0.3$, $c_3 = 0.55$ underweights performance relative to the default RPI weighting $c_1 = 0.25$. However, this reweighting also lowered the correlation between rank error and predicted conference strength, albeit only slightly.

This might be because the reweighting also underweights OWP relative to OOWP ($c_2 < c_3$). As indicated in the previous sections, an increase in c_3 with respect to c_2 would, in theory, more evenly distribute the variance between OWP and OOWP. This might improve calculations of context, which could increase accuracy. (It should be noted that the ratio between c_1 and c_2 remains the same for both the default RPI weighting and the highest-accuracy RPI weighting ($c_1/c_2 = 0.5$).)

There is another reason why this reweighting could be beneficial. Assume for example that WP is independently distributed by team and that WP has variance v. Assume that each team plays 16 games, for simplicity. Then, the variance of the default RPI metric would be

$$\text{Var} = 0.25 \cdot \text{Var(WP)} + 0.5 \cdot \text{Var(OWP)} + 0.25 \cdot \text{Var(OOWP)}$$

$$= 0.25 \cdot v + 0.5 \cdot (1/16 \cdot v) + 0.25 \cdot (1/256 \cdot v)$$

$$= 289/1024 \cdot v$$

And the variance of the highest-accuracy RPI reweighting $c_1 = 0.15$, $c_2 =$

0.3, $c_3 = 0.55$ would be

$$\text{Var(Default)} = 0.15 \cdot \text{Var(WP)} + 0.3 \cdot \text{Var(OWP)} + 0.55 \cdot \text{Var(OOWP)}$$

$$= 0.15 \cdot v + 0.3 \cdot (1/16 \cdot v) + 0.55 \cdot (1/256 \cdot v)$$

$$= 175/1024 \cdot v$$

Thus the variance in the reweighting is reduced. Thus, we should expect the RPI adjustment A for home-field advantage to have a stronger effect on the distribution. As mentioned above, the RPI formula includes a weak adjustment for home-field advantage. This adjustment will have a greater impact on the distribution if the distribution becomes narrower. A decrease in the general variance of the ranking metric should narrow the distribution, thus allowing for A to provide stronger adjustment for home-field advantage.

This hypothesis aligns with the home-field correlation findings because it implies that the RPI metric insufficiently accounts for the advantage of playing matches at home. It would appear that teams from stronger conferences have identified this and have used their financial resources to schedule more games at home. They know that home advantage will disproportionately help them from an RPI standpoint by artificially boosting their WP. The most accurate RPI weighting $c_1 = 0.15$, $c_2 = 0.3$, $c_3 = 0.55$, then, by underweighting WP and increasing OOWP, places less emphasis on a team's proportion of home games. Thus, although it benefits stronger-conference teams by increasing weight on context ($c_2 + c_3$), it also reduces the home advantage bias. This could explain an increase in accuracy (and the slight decrease in rank error correlation) for a reweighting which weights c_1 less. Further research is needed to test this theory.

The above intuitions are consistent with results and conclusions from the previous sections. Further analysis, especially on the relationship between con-

ference strength, home-field advantage and c_1 weightings, is needed in future research to confirm these intuitions.

9.4 Application to 2024 Championship Selection Process

It is motivated to examine how a shift in RPI weightings would have affected the selection process to the NCAA tournament. Thus, it is motivated to select one of the reweightings for further analysis. I selected the rankings from the RPI weighting $c_1 = 0.3$, $c_2 = 0.3$, $c_3 = 0.4$ as a generalizable case study. These rankings were compared to rankings from the default RPI weighting ($c_1 = 0.25$, $c_2 = 0.5$, $c_3 = 0.25$) to see how updated weightings would have potentially shifted team selection to the NCAA D1 Championship.

Table 8 displays the adjusted rankings for teams on the cusp of Championship selection. (As mentioned in the "Data" section, slight deviations from the official NCAA standard RPI rankings are the result of NCAA adjustments for non-Division I games.)

Table 8 shows that even a small shift in the weight on performance with respect to context (coupled with an emphasis on OOWP with respect to OWP) has a large impact for teams on the cusp of NCAA selection. Teams from smaller conferences such as UNC Greensboro (SoCon) and South Carolina (Sun Belt) get boosted into tournament contention. Multiple Big Ten teams drop places, including Indiana (-4), Michigan (-4), Maryland (-4) and Washington (-8); however, these teams stay in contention for an at large bid. Teams from Atlantic 10 conference, meanwhile, are subject to a curious shift; Duquesne bumps up 16 places while Saint Louis and Fordham fall 11 and nine places respectively. But perhaps the biggest shift is that California, a strong-conference team (ACC), is not even shown here; the Golden Bears fell 11 places to 55th in the updated rankings. No team with an RPI below 50 has been selected to

RPI Reweighting (0.3*WP + 0.3*OWP + 0.4*OOWP)			
Rank	Team	Conference	Rank Diff. From Std. RPI
30	Indiana (S)	Big Ten	-4
31	NC State (S)	ACC	-3
32	UNC Greensboro	SoCon	6
33	Monmouth	CAA	2
34	Seattle U (AQ)	WAC	12
35	Michigan (S)	Big Ten	-4
36	Maryland (S)	Big Ten	-4
37	UC Santa Barbara (S)	Big West	0
38	Washington (S)	Big Ten	-8
39	South Carolina	Sun Belt	12
40	Duquesne	Atlantic 10	16
41	Gardner-Webb (AQ)	Big South	13
42	Fordham (S)	Atlantic 10	-9
43	Bowling Green	MVC	16
44	Northwestern	Big Ten	-2
45	Saint Louis (S)	Atlantic 10	-11
46	Elon	CAA	1
47	Creighton	Big East	6
48	SIUE (AQ)	OVC	16
49	Drake	MVC	-6

Table 8: Rank Changes Under RPI Reweighting

the NCAA Championship in the last 10 years, meaning this drop would have effectively eliminated California from NCAA Championship contention. Thus, an overweighting of performance in this RPI reweighting benefits teams from smaller conferences in their quest for selection to the NCAA Championship.

Next and for comparison, the rankings from the highest-accuracy RPI weighting ($c_1 = 0.15$, $c_2 = 0.3$, $c_3 = 0.55$) will be examined to determine how they would have shifted team selection to the NCAA D1 Championship. Table 9 displays the adjusted rankings for teams on the cusp of Championship selection.

The modified rankings give three teams from traditionally strong conferences (California (ACC), Northwestern (Big Ten) and Wisconsin (Big Ten)) a significant boost in their contention for NCAA DI Championship selection. Meanwhile, two teams previously selected to the 2024 Championship from smaller

RPI Reweighting (0.15*WP + 0.3*OWP + 0.55*OOWP)			
Rank	Team	Conference	Rank Diff. From Std. RPI
30	Massachusetts (S)	Atlantic 10	-8
31	UCLA (S)	Big Ten	9
32	Oregon St. (S)	WCC	-7
33	Western Mich. (S)	MVC	-6
34	California	ACC	10
35	Northwestern	Big Ten	7
36	Saint Louis (S)	Atlantic 10	-2
37	Wisconsin	Big Ten	11
38	Monmouth	CAA	-3
39	Fordham (S)	Atlantic 10	-6
40	Notre Dame	ACC	9
41	Virginia Tech	ACC	-2
42	UNC Greensboro	SoCon	-4
43	High Point	Big South	-7
44	Creighton	Big East	9
45	UC Santa Barbara (S)	Big West	-8
46	UCF	Sun Belt	4
47	George Mason	Atlantic 10	-6
48	New Hampshire	America East	-3
49	Elon	CAA	-2

Table 9: Rank Changes Under RPI Reweighting

conferences (Fordham (Atlantic 10) and UC Santa Barbara (Big West)) drop six and 10 places respectively in the new rankings. This shows that the highest-accuracy RPI weighting favors teams from stronger conferences for NCAA Championship selection. Thus, an underweighting of performance in this RPI reweighting disadvantages teams from smaller conferences even though it may improve predictive accuracy in the context of the 2024 season.

10 Conclusion

This undergraduate senior thesis sought to explore the hypothesis that the Rating Percentage Index (RPI) exhibits systematic bias against teams from weaker conferences by overweighting context with respect to performance. After motivating the intuition behind this hypothesis, simulations were performed to test

it using a variety of RPI reweightings. Across all simulations, the default RPI weighting showed both a decrease in predictive accuracy and a strong preference for teams from stronger conferences when compared to a baseline Bradley-Terry model. In addition, across all simulations, increased weightings of performance led to an increases in accuracy and decreases in the correlation between conference strength and rank error (with respect to the baseline Bradley-Terry model). These results show a clear bias resulting from an underweighting of performance in the context of these simulations.

In addition, in the process of examining reweightings of performance, another potential inefficiency in the RPI weightings was revealed. Throughout the simulations and real data, the results consistently showed that reweightings of RPI context which emphasized OOWP and deemphasized OWP were more effective from both an accuracy and a conference bias standpoint. This finding shows that RPI (which more heavily weights OWP relative to OOWP) could be misleading if a team's opponents play weaker than average (or stronger than average) opposition. This potential inefficiency, and its relationship to conference strength, is a topic for further analysis.

After performing simulations, a similar procedure was used to analyze the empirical data from the 2024 NCAA D1 Men's Soccer season. This analysis surprisingly found that the most accurate RPI weightings underweighted performance with respect to the default RPI weighting. Although these findings appeared at first glance to contradict the simulation findings, they are likely the result of a third inefficiency in the RPI metric: an inadequate adjustment for home-field advantage. This inefficiency benefits teams from stronger conferences with greater financial resources. These teams disproportionately play at home, allowing them to enjoy the benefits of home advantage to boost their win percentage and with it their RPI. A clear distinction was drawn between the

simulations, in which home game percentage was not correlated to conference strength, and in the empirical data, in which home game percentage was highly correlated to conference strength. This distinction implies that the predictive accuracy metric, which does not account for home-field advantage, holds value in the simulations but is likely misleading in the context of the 2024 season data.

Because of this, the 2024 data should not be seen as conflicting with the results of the simulated data, although further analysis is needed to confirm and quantify the extent of the potential home-advantage bias. Instead, the data indicates that, although a higher weighting of winning percentage helps weaker conferences by devaluing strength of schedule, it also helps stronger conferences by emphasizing matches which are disproportionately played at home.

Thus, this undergraduate these has identified three separate inefficencies which challenge the integrity of the RPI baseline ranking system. This motivates potential modifications to the metric.

The NCAA could make a simple adjustment to the RPI metric to account for home advantage by utilizing an adjustment it has already made for other sports. It could weight away wins more than home wins in the calcuation for WP (without changing OWP or OOWP). This is already the case in college baseball, where home wins (and road losses) are worth 0.7 while away wins (and home losses) are worth 1.3 [20]. (Before switching to the NET metic, the NCAA also used a similar system in basketball, except home wins were worth 0.6 wins while away wins were worth 1.4 wins.) A similar adjustment for college soccer is reasonable and intuitive. It is also statistically justified; when excluding draws, the home team won 62.7% of matches in my dataset for the 2024 NCAA D1 Men's Soccer regular season, a number very close to the 60 percent rate at which home teams traditionally win in college baseball [21]. Measures which could rectify the other two inefficiencies have also be mentioned earlier in this

undergraduate thesis:

- performance (WP) could be weighted higher; and
- the weighting of OOWP could be increased relative to OWP.

Note that these bulleted suggestions are based on the simulated data, but reweighting for home advantage is based on empirical evidence from the 2024 season. Without a stronger adjustment for home vs. away results, richer teams from stronger conferences will always be able to game the system by playing more home matches to boost win percentage. In fact, an increased weighting of winning percentage may even emphasize this effect. Thus, I propose that the NCAA adopt baseball's home-adjusted RPI formula in the context of Men's Soccer. Once this is done, the NCAA should strongly consider increasing the weight on winning percentage to reduce conference bias. In addition, the NCAA should strongly consider a reweighting of OOWP relative to OWP to ensure that teams whose opponents play inferior opposition are not overvalued. Without these or similar adjustments, and without the adoption of a more robust alternative metric, the NCAA will continue to systematically select weaker teams to the D1 Championship simply by virtue of their conference strength and financial resources.

Thus, this undergraduate thesis has highlighted a systematic underweighting of winning percentage in simulations. Additionally, it has identified two other inefficencies in Adjusted RPI which are counter-intuitive and motivate modification. However, discrepancies between simulation and real-data results show that the relationship between the RPI weighting of performance and conference bias is complicated. This leaves room for further analysis on the modification to the RPI metric in NCAA DI Men's Soccer, especially in the context of data from prior seasons.

11 Conflict of Interest

The author is a NCAA student-athlete on the Men's Soccer team at Yale University. The author's relationship to the Yale Men's Soccer team or to the NCAA did not influence the design or results of this undergraduate thesis.

12 References

References

[1] NCAA Publications, *2024–2025 NCAA Division I Manual*, 2024. Available at: https://www.ncaapublications.com/p-4701-2024-2025-ncaa-division-i-manual.aspx

[2] NCAA, "NCAA General Administrative Guidelines Contents," 2024. Available at: https://ncaaorg.s3.amazonaws.com/championships/sports/soccer/d1/men/2024-25D1MSO_PreChampsManual.pdf

[3] J. Barsch, "Colorado Soccer SNUBBED by NCAA Tournament," *The Ralphie Report*, Nov. 5, 2018. Available at: https://www.ralphiereport.com/2018/11/5/18066212/colorado-soccer-snubbed-ncaa-womens-soccer-tournament

[4] "RPI Deep Dive in Final Week of Season," *BigSoccer Forum*, 2016. Available at: https://www.bigsoccer.com/threads/rpi-deep-dive-in-final-week-of-season.2133272/

[5] "RPI for Division I Women's Soccer - NCAA Tournament: Selection, Seeding, and Bracketing Criteria," 2024. Available at: https://sites.google.com/site/rpifordivisioniwomenssoccer/ncaa-selection-seeding-and-bracketing-criteria?authuser=0

[6] "College Basketball's NET Rankings, Explained," *NCAA.com*, Dec. 5, 2022. Available at: https://www.ncaa.com/news/basketball-men/article/2022-12-05/college-basketballs-net-rankings-explained

[7] "RPI for Division I Women's Soccer - RPI: Formula," 2025. Available at: https://sites.google.com/site/rpifordivisioniwomenssoccer/rpi-formula?authuser=0

[8] S. Sanders, "A Cheap Ticket to the Dance: Systematic Bias in College Basketball's Ratings Percentage Index," 2007. Available at: https://www.accessecon.com/pubs/EB/2007/Volume4/EB-07D80023A.pdf

[9] Wikipedia Contributors, "Rating Percentage Index," *Wikipedia*, Dec. 27, 2023. Available at: https://en.wikipedia.org/wiki/Rating_percentage_index

[10] "Seven ACC Teams Compete in NCAA Men's Soccer Round of 16 This Weekend," *TheACC.com*, Nov. 29, 2024. Available at: https://theacc.com/news/2024/11/29/seven-acc-teams-compete-in-ncaa-mens-soccer-round-of-16-this-weekend.aspx

[11] R. J. Paul and M. Wilson, "Political Correctness, Selection Bias, and the NCAA Basketball Tournament," *Journal of Sports Economics*, vol. 16, no. 2, pp. 201–213, Nov. 2012. doi: https://doi.org/10.1177/1527002512465413

[12] H. Jamil, "Analysis of Paired Comparison Data Using Bradley-Terry Models with Applications to Football Data," Sept. 6, 2021. Available at: https://haziqj.ml/publication/mastersthesis/

[13] R. A. Bradley and M. E. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, vol. 39, no. 3/4, p. 324, Dec. 1952. doi: https://doi.org/10.2307/2334029

[14] "Stan Reference Manual – Stan Docs," 2024. Available at: https://mc-stan.org/docs/reference-manual/index.html

[15] R. R. Davidson, "On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments," *Journal of the American Statistical Association*, vol. 65, no. 329, pp. 317–328, 1970. doi: https://doi.org/10.2307/2283595

[16] "NCAA College Men's Soccer Scores, Schedule," *NCAA.com*, 2018. Available at: https://www.ncaa.com/scoreboard/soccer-men/d1

[17] henrygd, "GitHub - henrygd/ncaa-api: Free API to Retrieve Live Scores, Stats, Standings, and Other College Sports Data from ncaa.com," Jan. 15, 2025. Available at: https://github.com/henrygd/ncaa-api

[18] conradlee123, "GitHub - conradlee123/amth491-cl-rpi-bias," GitHub, 2025. [Online]. Available: https://github.com/conradlee123/amth491-cl-rpi-bias

[19] T. Layberger, "Auburn, Michigan Paying Nearly $2 Million In Week 1 College Football Guarantee Games," *Forbes*, Aug. 29, 2023. [Online]. Available: https://www.forbes.com/sites/tomlayberger/2023/08/29/auburn-michigan-paying-nearly-2-million-in-week-1-college-football-guarantee-games/

[20] "Ground Rules: Understanding the new R.P.I," *Abca.org*, 2024. [Online]. Available: https://www.abca.org/magazine/magazine/2013-1-Winter/Ground_Rules_Understanding_the_new_RPI.aspx.

[21] "College Baseball Insider - Your Home for College Baseball," *Collegebaseballinsider.com*, 2025. [Online]. Available: https://www.collegebaseballinsider.com/09Articles/09RPI3.html

13 Appendix

RPI Reweighting	Average Accuracy
(0.25, 0.3, 0.45)	0.785428
(0.3, 0.3, 0.4)	0.785394
(0.3, 0.35, 0.35)	0.785364
(0.25, 0.25, 0.5)	0.785245
(0.25, 0.35, 0.4)	0.785200
(0.25, 0.5, 0.25)	0.779693

Table 10: Top-performing RPI reweightings by average accuracy, Round 1

RPI Reweighting	Average Accuracy
(0.25, 0.3, 0.45)	0.808670
(0.2, 0.25, 0.55)	0.808561
(0.25, 0.25, 0.5)	0.808510
(0.2, 0.2, 0.6)	0.808303
(0.3, 0.35, 0.35)	0.808275
(0.25, 0.5, 0.25)	0.803382

Table 11: Top-performing RPI reweightings by average accuracy, Round 2

RPI Reweighting	Average Accuracy
(0.45, 0.4, 0.15)	0.767450
(0.4, 0.35, 0.25)	0.767375
(0.5, 0.45, 0.05)	0.767345
(0.5, 0.4, 0.1)	0.767338
(0.35, 0.3, 0.35)	0.767328
(0.25, 0.5, 0.25)	0.760901

Table 12: Top-performing RPI reweightings by average accuracy, Round 3

RPI Reweighting	Average Accuracy
(0.3, 0.25, 0.45)	0.791365
(0.3, 0.3, 0.4)	0.791360
(0.25, 0.25, 0.5)	0.791347
(0.3, 0.35, 0.35)	0.791292
(0.25, 0.2, 0.55)	0.791267
(0.25, 0.5, 0.25)	0.786081

Table 13: Top-performing RPI reweightings by average accuracy, Round 4

RPI Reweighting	Average Accuracy
(0.45, 0.4, 0.15)	0.764676
(0.4, 0.35, 0.25)	0.764634
(0.45, 0.35, 0.2)	0.764616
(0.5, 0.4, 0.1)	0.764589
(0.5, 0.45, 0.05)	0.764528
(0.25, 0.5, 0.25)	0.758417

Table 14: Top-performing RPI reweightings by average accuracy, Round 5

RPI Reweighting	Average Accuracy
(0.3, 0.3, 0.4)	0.785131
(0.3, 0.35, 0.35)	0.785025
(0.35, 0.35, 0.3)	0.784973
(0.3, 0.25, 0.45)	0.784964
(0.35, 0.4, 0.25)	0.784881
(0.25, 0.5, 0.25)	0.780276

Table 15: Top-performing RPI reweightings by average accuracy, Round 6

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	**0.8281*** **	**0.5348** **	**0.8813*** **
(0.25, 0.3, 0.45)	**0.8731*** **	**0.7564*** **	**0.7297*** **
(0.3, 0.25, 0.45)	**0.7315*** **	**0.7880*** **	**0.3956**
(0.3, 0.3, 0.4)	**0.8088*** **	**0.7143*** **	**0.6396*** **
(0.35, 0.35, 0.3)	**0.6952*** **	**0.4985***	**0.6317*** **
(0.4, 0.35, 0.25)	**0.3254**	**0.1132**	**0.2490**
(0.45, 0.4, 0.15)	**0.0071**	**-0.2588**	**0.1050**

Table 16: Correlation between conference strength and rank error (w.r.t. reverse Bradley-Terry rankings) for selected RPI weightings, Round 1 (Note: * means $p < 0.05$, ** means $p < 0.01$ and *** means $p < 0.001$)

Model / Simulation Set	Set 1	Set 2	Set 3
Bradley-Terry-Davidson (BTDM)	0.769488	0.788873	0.782456
RPI	0.764957	0.783453	0.777395
(0.25, 0.3, 0.45)	0.768883	0.788253	0.781705
(0.3, 0.25, 0.45)	0.76834	0.787812	0.781294
(0.3, 0.3, 0.4)	0.768649	0.788111	0.781518
(0.35, 0.35, 0.3)	0.768425	0.787953	0.781412
(0.4, 0.35, 0.25)	0.768214	0.787661	0.781209
(0.45, 0.4, 0.15)	0.767938	0.787549	0.781132

Table 17: Predictive accuracy by model, evaluation Round 1

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	**-0.6086****	**-0.5853****	**-0.4922***
(0.25, 0.3, 0.45)	**-0.8798*****	**-0.8131*****	**-0.7872*****
(0.3, 0.25, 0.45)	**-0.9350*****	**-0.8416*****	**-0.8589*****
(0.3, 0.3, 0.4)	**-0.9247*****	**-0.8569*****	**-0.8314*****
(0.35, 0.35, 0.3)	**-0.9328*****	**-0.8817*****	**-0.8257*****
(0.4, 0.35, 0.25)	**-0.9451*****	**-0.8874*****	**-0.8420*****
(0.45, 0.4, 0.15)	**-0.9461*****	**-0.8989*****	**-0.8341*****

Table 18: Correlation between conference strength and rank error (w.r.t. ground truth rankings) for selected RPI weightings, Round 1

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	**0.5674****	**0.9531*****	**0.8364*****
(0.25, 0.3, 0.45)	**0.7426*****	**0.9652*****	**0.7947*****
(0.3, 0.25, 0.45)	**0.7638*****	**0.9384*****	**0.6371*****
(0.3, 0.3, 0.4)	**0.7060*****	**0.9432*****	**0.7349*****
(0.35, 0.35, 0.3)	**0.4729***	**0.8500*****	**0.6142****
(0.4, 0.35, 0.25)	-0.1716	0.2678	-0.0621
(0.45, 0.4, 0.15)	**-0.5953****	-0.3585	**-0.4299***

Table 19: Correlation between conference strength and rank error (w.r.t. reverse Bradley-Terry rankings) for selected RPI weightings, Round 2

Model / Simulation Set	Set 1	Set 2	Set 3
BTDM	0.801842	0.78567	0.805938
RPI	0.798632	0.78029	0.803818
(0.25, 0.3, 0.45)	0.802714	0.785997	0.807741
(0.3, 0.25, 0.45)	0.801705	0.785176	0.806322
(0.3, 0.3, 0.4)	0.802234	0.785643	0.806857
(0.35, 0.35, 0.3)	0.80157	0.785111	0.806112
(0.4, 0.35, 0.25)	0.80069	0.784415	0.804911
(0.45, 0.4, 0.15)	0.800261	0.783989	0.804119

Table 20: Predictive accuracy by model, evaluation Round 1

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	**-0.7288*** **	**-0.5837** **	**-0.5033***
(0.25, 0.3, 0.45)	**-0.7974*** **	**-0.7255*** **	**-0.6647*** **
(0.3, 0.25, 0.45)	**-0.8382*** **	**-0.8151*** **	**-0.7713*** **
(0.3, 0.3, 0.4)	**-0.8346*** **	**-0.7943*** **	**-0.7404*** **
(0.35, 0.35, 0.3)	**-0.8489*** **	**-0.8169*** **	**-0.7615*** **
(0.4, 0.35, 0.25)	**-0.8582*** **	**-0.8448*** **	**-0.7967*** **
(0.45, 0.4, 0.15)	**-0.8636*** **	**-0.8535*** **	**-0.8035*** **

Table 21: Correlation between conference strength and rank error (w.r.t. ground truth rankings) for selected RPI weightings, Round 2

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	**0.8249*** **	**0.7449*** **	**0.6043** **
(0.25, 0.3, 0.45)	**0.7723*** **	**0.7075*** **	**0.5637** **
(0.3, 0.25, 0.45)	**0.6285** **	**0.5796** **	**0.4435***
(0.3, 0.3, 0.4)	**0.6719*** **	**0.6137*** **	**0.4720***
(0.35, 0.35, 0.3)	**0.5659** **	**0.5223***	0.3944
(0.4, 0.35, 0.25)	0.3598	0.3668	0.2825
(0.45, 0.4, 0.15)	0.1972	0.2496	0.2059

Table 22: Correlation between conference strength and rank error (w.r.t. reverse Bradley-Terry rankings) for selected RPI weightings, Round 3

Model / Simulation Set	Set 1	Set 2	Set 3
BTDM	0.779724	0.772387	0.768248
RPI	0.772922	0.767021	0.763291
(0.25, 0.3, 0.45)	0.778414	0.771793	0.767883
(0.3, 0.25, 0.45)	0.779122	0.772547	0.76862
(0.3, 0.3, 0.4)	0.779215	0.7725	0.768615
(0.35, 0.35, 0.3)	0.779407	0.772838	0.768913
(0.4, 0.35, 0.25)	0.779345	0.772896	0.769013
(0.45, 0.4, 0.15)	0.779264	0.772903	0.768991

Table 23: Predictive accuracy by model, evaluation Round 3

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	-0.3031	-0.1099	-0.2372
(0.25, 0.3, 0.45)	**-0.5813** **	-0.399	-0.391
(0.3, 0.25, 0.45)	**-0.7877*** **	**-0.6774*** **	**-0.5903** **
(0.3, 0.3, 0.4)	**-0.7652*** **	**-0.6325** **	**-0.555** **
(0.35, 0.35, 0.3)	**-0.8282*** **	**-0.7181*** **	**-0.6323** **
(0.4, 0.35, 0.25)	**-0.8698*** **	**-0.7894*** **	**-0.7028*** **
(0.45, 0.4, 0.15)	**-0.8874*** **	**-0.8164*** **	**-0.7339*** **

Table 24: Correlation between conference strength and rank error (w.r.t. ground truth rankings) for selected RPI weightings, Round 3

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	**0.9110*** **	**0.9099*** **	**0.9166*** **
(0.25, 0.3, 0.45)	**0.8870*** **	**0.9041*** **	**0.8933*** **
(0.3, 0.25, 0.45)	**0.8049*** **	**0.8641*** **	**0.8256*** **
(0.3, 0.3, 0.4)	**0.8362*** **	**0.8769*** **	**0.8470*** **
(0.35, 0.35, 0.3)	**0.7685*** **	**0.8367*** **	**0.7924*** **
(0.4, 0.35, 0.25)	**0.5672** **	**0.7212*** **	**0.6611*** **
(0.45, 0.4, 0.15)	0.3555	**0.5656** **	**0.5246* **

Table 25: Correlation between conference strength and rank error (w.r.t. reverse Bradley-Terry rankings) for selected RPI weightings, Round 4

Model / Simulation Set	Set 1	Set 2	Set 3
BTDM	0.787558	0.794746	0.766667
RPI	0.792088	0.799409	0.771502
(0.25, 0.3, 0.45)	0.79165	0.79901	0.771586
(0.3, 0.25, 0.45)	0.791992	0.799285	0.771833
(0.3, 0.3, 0.4)	0.791623	0.79876	0.771459
(0.35, 0.35, 0.3)	0.790883	0.797669	0.770884
(0.4, 0.35, 0.25)	0.790333	0.797068	0.770289

Table 26: Predictive accuracy by model, evaluation Round 4

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	**-0.6204** **	**-0.627** **	**-0.5046* **
(0.25, 0.3, 0.45)	**-0.7805*** **	**-0.8296*** **	**-0.7487*** **
(0.3, 0.25, 0.45)	**-0.8694*** **	**-0.9194*** **	**-0.8828*** **
(0.3, 0.3, 0.4)	**-0.8536*** **	**-0.9078*** **	**-0.8628*** **
(0.35, 0.35, 0.3)	**-0.8756*** **	**-0.9265*** **	**-0.8946*** **
(0.4, 0.35, 0.25)	**-0.8962*** **	**-0.9411*** **	**-0.921*** **
(0.45, 0.4, 0.15)	**-0.9028*** **	**-0.9454*** **	**-0.929*** **

Table 27: Correlation between conference strength and rank error (w.r.t. ground truth rankings) for selected RPI weightings, Round 4

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	**0.5327* **	**0.7289*** **	**0.5084* **
(0.25, 0.3, 0.45)	**0.4890* **	**0.6804*** **	**0.4660* **
(0.3, 0.25, 0.45)	0.3707	**0.5960** **	0.3564
(0.3, 0.3, 0.4)	0.3935	**0.6194** **	0.3786
(0.35, 0.35, 0.3)	0.3100	**0.5720** **	0.3036
(0.4, 0.35, 0.25)	0.1974	**0.4914* **	0.1989
(0.45, 0.4, 0.15)	0.1142	**0.4367* **	0.1218

Table 28: Correlation between conference strength and rank error (w.r.t. reverse Bradley-Terry rankings) for selected RPI weightings, Round 5

Model / Simulation Set	Set 1	Set 2	Set 3
BTDM	0.753165	0.752854	0.757793
RPI	0.756481	0.757657	0.762766
(0.25, 0.3, 0.45)	0.757223	0.75844	0.763852
(0.3, 0.25, 0.45)	0.757274	0.75849	0.763851
(0.3, 0.3, 0.4)	0.757398	0.758795	0.764193
(0.35, 0.35, 0.3)	0.757521	0.758937	0.764463
(0.4, 0.35, 0.25)	0.757496	0.759	0.764447

Table 29: Predictive accuracy by model, evaluation Round 5

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	-0.1429	-0.2867	-0.0895
(0.25, 0.3, 0.45)	-0.249	-0.1023	-0.2234
(0.3, 0.25, 0.45)	**-0.4322***	**-0.451***	**-0.4483***
(0.3, 0.3, 0.4)	-0.4027	-0.1399	-0.4119
(0.35, 0.35, 0.3)	**-0.4916***	-0.2904	**-0.5219***
(0.4, 0.35, 0.25)	**-0.577****	**-0.451***	**-0.625****
(0.45, 0.4, 0.15)	**-0.6221****	**-0.5287***	**-0.6783*****

Table 30: Correlation between conference strength and rank error (w.r.t. ground truth rankings) for selected RPI weightings, Round 5

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	**0.6781*****	**0.7796*****	**0.7294*****
(0.25, 0.3, 0.45)	**0.5446****	**0.7654*****	**0.6486*****
(0.3, 0.25, 0.45)	**0.5446****	**0.6667*****	0.3555
(0.3, 0.3, 0.4)	**0.5792*****	**0.6831*****	0.4222
(0.35, 0.35, 0.3)	**0.4792***	**0.5752*****	0.1876
(0.4, 0.35, 0.25)	0.3042	0.3934	-0.1325
(0.45, 0.4, 0.15)	0.1638	0.2169	-0.3225

Table 31: Correlation between conference strength and rank error (w.r.t. reverse Bradley-Terry rankings) for selected RPI weightings, Round 6

Model / Simulation Set	Set 1	Set 2	Set 3
BTDM	0.770385	0.78055	0.768004
RPI	0.774349	0.78486	0.772452
(0.25, 0.3, 0.45)	0.774847	0.784948	0.773254
(0.3, 0.25, 0.45)	0.774868	0.784996	0.773162
(0.3, 0.3, 0.4)	0.77488	0.784534	0.773215
(0.35, 0.35, 0.3)	0.774537	0.783928	0.773038
(0.4, 0.35, 0.25)	0.774293	0.78323	0.772864

Table 32: Predictive accuracy by model, evaluation Round 6

RPI Reweighting	Set 1	Set 2	Set 3
(0.25, 0.5, 0.25)	-0.412	**-0.4887***	**-0.6188****
(0.25, 0.3, 0.45)	**-0.5581****	**-0.5907****	**-0.7177*****
(0.3, 0.25, 0.45)	**-0.7229*****	**-0.713*****	**-0.8097*****
(0.3, 0.3, 0.4)	**-0.6937*****	**-0.6908*****	**-0.793*****
(0.35, 0.35, 0.3)	**-0.7525*****	**-0.7363*****	**-0.8242*****
(0.4, 0.35, 0.25)	**-0.8078*****	**-0.7806*****	**-0.8553*****
(0.45, 0.4, 0.15)	**-0.8303*****	**-0.7998*****	**-0.8684*****

Table 33: Correlation between conference strength and rank error (w.r.t. ground truth rankings) for selected RPI weightings, Round 6

RPI Reweighting	Correlation	Accuracy
(0.25, 0.5, 0.25)	**0.9201*****	0.769924
(0.25, 0.3, 0.45)	**0.829*****	0.767845
(0.3, 0.25, 0.45)	**0.7732*****	0.765073
(0.3, 0.3, 0.4)	**0.7922*****	0.764380
(0.35, 0.35, 0.3)	**0.6977*****	0.763687
(0.4, 0.35, 0.25)	**0.4753***	0.762994
(0.45, 0.4, 0.15)	0.2324	0.762994
Highest-performing reweightings		
(0.15, 0.3, 0.55)	**0.8621*****	0.776161
(0.2, 0.45, 0.35)	**0.868*****	0.776161
(0.2, 0.5, 0.3)	**0.8624*****	0.774775
(0.2, 0.4, 0.4)	**0.87*****	0.774775
(0.15, 0.15, 0.7)	**0.8372*****	0.774775

Table 34: Rank error correlation (w.r.t reverse Bradley-Terry rankings) and accuracy for various RPI weightings, 2024 NCAA D1 data