Ian G. Ludden*, Arash Khatibi, Douglas M. King and Sheldon H. Jacobson

Models for generating NCAA men's basketball tournament bracket pools

https://doi.org/10.1515/jqas-2019-0022

Abstract: Each year, the NCAA Division I Men's Basketball Tournament attracts popular attention, including *bracket challenges* where fans seek to pick the winners of the tournament's games. However, the quantity and unpredictable nature of games suggest a single bracket will likely select some winning teams incorrectly even if created with insightful and sophisticated methods. Hence, rather than focusing on creating a single bracket to perform well, a challenge participant may wish to create a pool of brackets that likely contains at least one high-scoring bracket. This paper proposes a power model to estimate tournament outcome probabilities based on past tournament data. Bracket pools are generated for the 2013–2019 tournaments using six generators, five using the power model and one using the Bradley-Terry model. The generated brackets are assessed by the ESPN scoring system and compared to those produced by a traditional pick favorite approach as well as the highest scoring brackets in the ESPN Tournament Challenge for each year.

Keywords: bracket generation; Bradley-Terry; March madness; model selection; power model; sports forecasting

1 Introduction

In the NCAA Division I Men's Basketball Tournament, teams compete in a 68-team single-elimination bracket to determine a national champion. In the years since 1985 (referred to as the *modern era*), the tournament has included at least 64 teams and six main rounds; Table 12 in Appendix A.1 lists the official and colloquial names of all six rounds. The tournament attracts popular attention throughout the United States; before the tournament begins, many fans participate in *bracket challenges* where they pick the winner of each of the tournament's 63 games (the *First Four* play-in games are not included in any such challenges). The term *bracket*, originally used to describe the structure of a sports tournament, is also used colloquially to refer to a "filled-out" bracket, that is, a complete set of game predictions for a tournament. As the tournament progresses, all of the submitted brackets are scored based on their correct picks, with correct picks in later rounds typically earning more points than correct picks in earlier rounds. The unpredictable nature of game outcomes and the large number of games suggest that a single bracket, even if it has been created with insightful and sophisticated methods and data, is likely to incorrectly select some winning teams. Hence, rather than selecting a single bracket that seeks to correctly pick all game winners, a challenge participant may wish to create a pool of brackets such that at least one bracket in the pool is likely to score well.

One appealing approach for creating a bracket is to select the favorite (based on seed) to win in each game. The participating teams are divided into four regions, each comprising sixteen teams seeded between 1 and 16, with lower numbers signifying more highly-ranked teams. Hence, a *pick favorite* approach deterministically specifies all picks in the four regions (i.e. 60 of the 63 games), resulting in the four 1-seeds advancing to the Final Four. Eight different pick favorite brackets can be created, depending on which of the equally-seeded regional champions are selected to win in the three remaining games of the tournament. While these eight brackets, collectively referred to as the *pick favorite pool*, appear to be the most likely outcomes, they represent only a very small portion of the more than nine quintillion $\left(9 \times 10^{18}\right)$ possible brackets. Moreover, over the last five tournaments, there have been on average 13 upsets in the first two rounds of play, whose winners a pick favorite bracket will not select correctly. Creating high-quality brackets depends on the ability to correctly identify upsets.

While upsets have reliably occurred in the tournament, and hence, must be present in brackets that are likely to score well, these upsets must be selected intelligently. One possible approach is to generate

***Corresponding author: Ian G. Ludden,** University of Illinois, Computer Science, Urbana, IL, USA, e-mail: iludden2@illinois.edu
Arash Khatibi and Douglas M. King: University of Illinois, Industrial and Enterprise Systems Engineering, Urbana, IL, USA, e-mail: arashkhatibi7@gmail.com; dmking@illinois.edu
Sheldon H. Jacobson: University of Illinois, Computer Science, Urbana, IL, USA, e-mail: shj@illinois.edu

brackets according to the probability mass function (PMF) of tournament outcomes. From this perspective, the outcome of each tournament is one of 2^{63} potential outcomes of a discrete random variable, and the PMF prescribes the probability of each of these outcomes for a given year. However, the PMF is unknown and must be estimated from data, which may include past tournament outcomes, team performance in regular season games, or other factors. Many approaches have been proposed to estimate the PMF, which are useful if the objective is to generate brackets to maximize expected winnings in a bracket pool (Kenter 2016). One approach considers factors such as the victory margin (using a team's performance history) and the game's venue to predict the result of each game. For example, Lopez and Mathews (2015) proposes a prediction model by combining the point spreads set by Las Vegas sports-books with possession-based team efficiency metrics. Gupta (2015) proposes a rating system and a four-predictor probability model to generate brackets for the 2009–2014 tournaments. Ruiz and Perez-Cruz (2015) modifies a classical model for soccer to model both the specific behavior of each conference and different strategies of teams. Yuan et al. (2015) proposes methods to forecast the results of the tournament and discusses the difficulties in using publicly available data while presenting novel ideas for post-processing statistical forecasts. Kaplan and Garstka (2001) uses Sagarin ratings (i.e. expected scoring rates on a team-by-team basis) to estimate the win probability of each team. West (2006, 2008) presents an ordinal logistic regression method to predict how many games each team will win in the tournament and compares the results to tournament simulations using the Bradley-Terry model (Bradley and Terry 1952). Koenker and Bassett Jr. (2010) defines offensive and defensive strength parameters and a home adjustment factor to compute the win probability of each team. Kvam and Sokol (2006) uses victory margins of teams in regular-season games from 1999 to 2005 to predict game winners based on a Markov chain/logistic regression model. While these models study the performance of each team more precisely, their parameters depend on teams' specific information such as victory margin and game's venue. Therefore, they cannot be easily applied to other tournaments.

An alternative approach uses seeds to estimate the win probability of each team in each round (see for example Boulier and Stekler 1999; Jacobson et al. 2011; Khatibi, King, and Jacobson 2015). These models consider all teams of the same seed number as equivalent, ignoring factors such as the game's venue and the victory margin. Assuming the same performance for all teams of the same seed number may increase the estimation error but simplifies the computational effort, since model parameters depend only on a team's seed number. This simplification gives seed-based models flexibility for use in future tournaments and other sports competitions with the same structure.

This paper proposes a probability model that implicitly estimates the PMF of bracket outcomes with the end goal of generating strong bracket pools. The power model estimates the win probability of each team in a game as a power function of the teams' seed numbers. This approach generalizes the Bradley-Terry model, adding round-dependent parameters and using a different estimation technique (Bradley and Terry 1952). The power model is used in a bracket generation algorithm with five different parameter settings to define five randomized bracket generators. A Bradley-Terry-based generator is also considered as a baseline. These six generators are evaluated by using each to create pools of brackets for the 2013–2019 tournaments and scoring each bracket with the ESPN scoring system. The bracket pools are then evaluated against the pick favorite pool and the highest-scoring brackets in the ESPN Tournament Challenge for each year. The six generators are then compared by the multiple comparisons with the best (MCB) method proposed by Hsu (1984).

The ESPN scoring system awards 10 points for each correctly picked winner in Round 1. There are 32 games played in total across four regions in Round 1, which results in a maximum of 320 points for this round. The value of each correct pick in a subsequent round is twice that of the previous round; hence, correctly picking the winner of the final game earns 320 points. For each round, the maximum score is the same (320 points). The maximum total score, which is achieved by correctly picking the winner in all 63 games of the tournament, is 1920. While points are awarded for correctly picking winners of individual games, correct picks in later rounds are critical, both because they individually earn more points and because they implicitly earn points for correct picks made in earlier rounds. For example, correctly picking the champion team secures 630 points since at least one winner is correctly picked in each of the six rounds. Therefore, a good generator must minimize the number of incorrect picks in later rounds. However, as the incorrect picks of the earlier rounds propagate through the tournament, the best generated bracket must also correctly pick many of the winners in the beginning rounds.

The power model has been implemented at (http://bracketodds.cs.illinois.edu) as a web-based bracket generator, which has attracted significant public and media attention. This web site was launched as a tool to make the

public familiar with the mathematical modeling of sports events and data analytics.

The paper is organized as follows. Section 2 introduces the power model, proposes six bracket generators, and establishes how these generators will be evaluated. Section 3 reports the power and Bradley-Terry model parameters for predicting tournament outcomes in 2013–2020 and the results of bracket generation experiments. Section 4 provides concluding remarks and potential directions for future work.

2 The power model

This section introduces the power model for estimating win probabilities of all possible seed match-ups and presents a general procedure for generating brackets with these estimates.

2.1 Estimating win probabilities

In a game between two different seed numbers, the win probability typically favors the team with the smaller seed number, i.e. the stronger seed. The simplest function to model a relation between the teams' seed numbers with a larger weight on better seeds is a linear function: In a game between two teams with seed numbers s_1 and s_2, the linear function $s_2/(s_1 + s_2)$ can be used to estimate the win probability of seed s_1. However, the linear estimator has two drawbacks. First, while the results of the modern era tournaments show that the performance of a seed number depends on the round (discussed in more detail later in this section), the outcome of the linear estimator is independent of the round number. Second, the linear estimator assumes the relative strength of a team is proportional to the reciprocal of its seed number, which fails to capture more nuanced patterns in team strength.

To avoid these drawbacks of the linear estimator, the power model estimates the win probability of each seed in a game as a power function of the seeds' ratio. Throughout the following definitions, let s_1 and s_2 be seeds with $s_1 < s_2$. Let $p_{j,(s_1,s_2)}$ denote the probability that s_1 will defeat s_2 in Round j. The odds ratio for seed s_1 is modeled as

$$\frac{p_{j,(s_1,s_2)}}{1-p_{j,(s_1,s_2)}} = \left(\frac{s_2}{s_1}\right)^{\alpha_{j,(s_1,s_2)}}, \tag{1}$$

where $\alpha_{j,(s_1,s_2)}$, referred to as the α-value, is a function of the round number (j) and the seeds (s_1 and s_2). An equivalent definition using the logit function is

$$\begin{aligned}\operatorname{logit}\left(p_{j,(s_1,s_2)}\right) &= \log\left(\frac{p_{j,(s_1,s_2)}}{1-p_{j,(s_1,s_2)}}\right)\\ &= \alpha_{j,(s_1,s_2)} \cdot \left(\log(1/s_1) - \log(1/s_2)\right).\end{aligned} \tag{2}$$

For comparison, the Bradley-Terry model is

$$\operatorname{logit}\left(p_{j,(s_1,s_2)}\right) = \frac{p_{j,(s_1,s_2)}}{1-p_{j,(s_1,s_2)}} = \beta_{s_1} - \beta_{s_2}, \tag{3}$$

where β_{s_1} and β_{s_2} are score parameters for their respective seeds (Bradley and Terry 1952). The power model therefore generalizes the Bradley-Terry model by starting with score parameters $\beta_s = \log(1/s)$ for $s \in \{1, 2, \ldots, 16\}$ and adding round-dependent logit scaling factors $\alpha_{j,(s_1,s_2)}$. This round-dependent scaling, which allows probability estimates for the same seed match-up to vary between rounds, is a key advantage of the power model over the Bradley-Terry model. Data from past tournaments suggest the relative performance of seeds varies across the rounds, warranting this increased granularity of parameters. While the sample sizes for individual match-ups are too small to provide statistically significant evidence of this variation, the specific outcomes present some counterintuitive scenarios. For example, 5-seeds have surprisingly won all four match-ups against 2-seeds in the Elite Eight, but such a small sample size precludes strong statistical conclusions. Therefore, the overall win rate of the stronger seed is compared across rounds. In the modern era (1985–2019), the stronger seed has won in 200 of 280 games (71.4%) in the Sweet Sixteen and 77 of 140 games (55%) in the Elite Eight. The standard z-test for difference in proportions is used to test the null hypothesis that the proportion of games won by the stronger seed is the same in both rounds against a one-sided alternative that the proportion is greater in the Sweet Sixteen. The z statistic obtained is 3.344, and the corresponding p-value of 0.00041 is significant at the 0.05 level, suggesting that the null hypothesis should be rejected. The one-sided test therefore provides evidence that stronger seeds are less successful in the Elite Eight than in the Sweet Sixteen. It must be mentioned that the teams competing in the Elite Eight necessarily won their Sweet Sixteen games, so these samples are not entirely independent. As further justification of round-dependent parameters, an earlier study finds evidence that 1-, 2-, and 3-seeds do not have significantly different historical win percentages in the later rounds (Jacobson and King 2009).

To estimate the power model parameters for year y, tournament games from 1985 through the year $y - 1$ are used as the training data set. These data are taken from

the official NCAA 2019 Men's Final Four Records Book (NCAA 2019). The specific data extracted are $n_{j,(s_1,s_2)}$, the number of times an s_1-seed has defeated an s_2-seed in Round j in the training data set, for all $j \in \{1, 2, \ldots, 6\}$ and $s_1, s_2 \in \{1, 2, \ldots, 16\}$. Let $n_{j,(s_1,\bullet)} = \sum_{s_2} n_{j,(s_1,s_2)}$ denote the number of times an s_1-seed has won in Round j over any other seed. Furthermore, let $\bar{p}_{j,(s_1,s_2)} = n_{j,(s_1,s_2)}/(n_{j,(s_1,s_2)} + n_{j,(s_2,s_1)})$ be the proportion of times an s_1-seed has defeated an s_2-seed in Round j. The total number of games played in Round j is represented as $n_j = \sum_{s_1} n_{j,(s_1,\bullet)}$. To obtain an α-value estimate $\hat{\alpha}_{j,(s_1,s_2)}$, one can then substitute the observed proportion $\bar{p}_{j,(s_1,s_2)}$ for the unknown probability $p_{j,(s_1,s_2)}$, assuming s_1 and s_2 have met in Round j before. Rearranging (1) then yields

$$\hat{\alpha}_{j,(s_1,s_2)} \equiv \frac{\operatorname{logit}\left(\bar{p}_{j,(s_1,s_2)}\right)}{\log(1/s_1) - \log(1/s_2)}. \tag{4}$$

Finally, $p_{j,(s_1,s_2)}$ can be estimated as

$$\hat{p}_{j,(s_1,s_2)} \equiv \frac{s_2^{\hat{\alpha}_{j,(s_1,s_2)}}}{s_1^{\hat{\alpha}_{j,(s_1,s_2)}} + s_2^{\hat{\alpha}_{j,(s_1,s_2)}}}. \tag{5}$$

The α-value summarizes the performance history of seeds s_1 and s_2 in Round j. By (5), a positive α-value indicates a larger winning probability for the stronger seed while a negative value indicates a larger winning probability for the weaker seed. Note that setting $\hat{\alpha}_{j,(s_1,s_2)} = 0$ is equivalent to estimating a 0.5 winning probability for each seed (i.e. picking the winner randomly), while setting $\hat{\alpha}_{j,(s_1,s_2)} = +\infty$ results in the Pick Favorite (PF) approach, which always picks the stronger seed as the winner. One drawback of this method is that it ignores the number of games between two seeds (e.g. there is no difference between two pairs of seeds whose pairwise win records are 40–60 and 4–6). As more tournaments occur, providing more data, this limitation will have less and less of an impact.

The estimated probability $\hat{p}_{j,(s_1,s_2)}$ matches the observed proportion $\bar{p}_{j,(s_1,s_2)}$, which follows directly from (4) and (5). Matching the estimated probability to the observed proportion in this manner is of course only one of many valid approaches to choosing parameters.

Small training data sets may cause extreme parameter values. If $\bar{p}_{j,(s_1,s_2)}$ is 1 or 0, indicating one of the seeds has always won in past meetings, then $\hat{\alpha}_{j,(s_1,s_2)} = \pm\infty$. This scenario is particularly likely in later rounds since they have so few games and more possible match-ups. Every possible match-up in Round 1 always occurs four times per year, yielding 140 games each since 1985. However, there are $\binom{16}{2} = 120$ possible match-ups in Round 5 between distinct seeds, and the 68 Round 5 games in the modern era have only included 20 of these. Furthermore, 11 of these 20 have occurred only once, guaranteeing $\bar{p}_{j,(s_1,s_2)} \in \{0, 1\}$. This precludes the seed which has always lost from being chosen as the winner even though that seed can conceivably win a game in a future tournament. Establishing an upper bound K on the magnitude of the α-value estimates prevents this problem, so this modification will be used in applying the power model to the NCAA tournament.

Furthermore, some match-ups, particularly in later rounds, may have no prior data, i.e. $n_{j,(s_1,s_2)} = n_{j,(s_2,s_1)} = 0$, so $\hat{\alpha}_{j,(s_1,s_2)}$ is undefined. One workaround is computing a weighted average $\hat{\alpha}_j$ of the α-value estimates for all observed match-ups in a round:

$$\hat{\alpha}_j = \frac{1}{n_j} \cdot \sum_{(s_1,s_2)\in S} \left(n_{j,(s_1,s_2)} + n_{j,(s_2,s_1)}\right) \cdot \hat{\alpha}_{j,(s_1,s_2)}, \tag{6}$$

where $S = \{(s_1, s_2) : 1 \le s_1 < s_2 \le 16\}$ is the set of all possible pairings of stronger and weaker seeds. Games between identical seeds, which may occur in Rounds 5 and 6, are not included in the computation of $\hat{\alpha}_j$.

The power model can then use $\hat{\alpha}_j$ as the estimated α-value for all match-ups in that round. Effectively, $\hat{\alpha}_j$ summarizes the performance of stronger seeds versus weaker seeds in Round j, capturing the overall likelihood of upsets in the round. In Rounds 2–6, many possible match-ups have few or no occurrences in the NCAA tournament history since 1985, so this modification will be applied to those rounds.

2.2 Generating brackets

Bracket generation requires selecting the winner of all 63 games. Given win probability estimates such as those provided by the power model, the general procedure presented in Algorithm 1 generates a random bracket. Given estimated win probabilities (from, e.g. the power or Bradley-Terry models) $P = \{\hat{p}_{j,(s_1,s_2)}\}_{j\in[6],(s_1,s_2)\in S}$, the procedure generates an entire bracket round by round by choosing s_1 (resp., s_2) as the winner with probability $\hat{p}_{j,(s_1,s_2)}$ (resp., $1 - \hat{p}_{j,(s_1,s_2)}$). The input r allows for the seeds reaching the r-th round to be sampled by some other model, implemented as SAMPLE; this may be desirable if one wishes to approximately match the historical distribution of seeds reaching later rounds. The provided SAMPLE function may sample seeds independently or dependently.

When $r = 1$, the SAMPLE parameter is unnecessary since the seeds in Round 1 are already fixed. However, setting $r = 1$ maximizes the potential for erroneous picks in

Algorithm 1: Generate an NCAA tournament bracket

Function: GenerateBracket(P, r, Sample)
Input : Estimated win probabilities $P = \{\hat{p}_{j,(s_1,s_2)}\}_{j\in[6],(s_1,s_2)\in S}$; round index $r \in \{1, 2, \ldots, 6\}$; sampling function Sample
Output: A bracket, i.e. predicted winners of all 63 tournament games

Sample which teams reach Round r using Sample
Fix game outcomes in prior rounds, as needed, to ensure sampled teams reach Round r
for *j = 1 to 6* **do**
  foreach *game g in Round j* **do**
    if *outcome of g is not yet fixed* **then**
      $s_1 \leftarrow$seed of first team (t_1) predicted to play in g
      $s_2 \leftarrow$seed of second team (t_2) predicted to play in g
      rand $\leftarrow$ uniform random real number between 0 and 1
      if rand $< \hat{p}_{j,(s_1,s_2)}$ **then**
        Pick t_1 as the winner of g
      end
      else
        Pick t_2 as the winner of g
      end
    end
  end
end
return bracket (predicted winners of all 63 games)

early rounds to propagate through the bracket, so $r = 4, 5$, and 6 are also considered in this study. Jacobson et al. (2011) finds that the truncated geometric distribution is a reasonable fit for the observed seed distribution in Rounds 4–6. The truncated geometric distribution for a random variable X with parameters q and $i_{\max}$ is defined by the probability mass function $\Pr[X = i] = kq(1-q)^{i-1}$ for $i = 1, 2, \ldots, i_{\max}$, where $k = 1/(1-(1-q)^{i_{\max}})$ is a coefficient included to ensure the probabilities sum to one.

Table 1 illustrates how the truncated geometric distribution is fit to the observed Final Four seed distribution from 1985 through 2019. The geometric parameter q is estimated as the reciprocal of the average seed number, or

$$\hat{q} = \frac{1}{\sum_{i=1}^{16} i \cdot n_{4,(i,\bullet)}}.$$

(Recall that $n_{4,(i,\bullet)}$ denotes the total number of times seed i has *won* in Round 4, and hence the total number of times seed i has *reached* the Final Four.) The geometric PMF for $\hat{q}$ is then truncated at 16, yielding the truncated geometric PMF. The expected counts are computed by multiplying the total number of observations by the truncated geometric PMF. The goodness-of-fit test statistic is $\chi^2 = 18.321$, which corresponds to a p-value of 0.246 at 15 degrees of freedom. The p-value suggests the null hypothesis that the Final Four seeds follow a truncated geometric

Table 1: Truncated geometric distribution fit to observed final four seed distribution for 1985–2019.

Seed	Observed count	Truncated geometric PMF	Expected count
1	57	0.3548	49.665
2	29	0.2290	32.063
3	17	0.1479	20.699
4	13	0.0954	13.362
5	7	0.0616	8.626
6	3	0.0398	5.569
7	3	0.0257	3.595
8	5	0.0166	2.321
9	1	0.0107	1.498
10	1	0.0069	0.967
11	4	0.0045	0.624
12	0	0.0029	0.403
13	0	0.0019	0.260
14	0	0.0012	0.168
15	0	0.0008	0.108
16	0	0.0005	0.070
Total	140	1.000	140

distribution should not be rejected, and hence the truncated geometric distribution may be used to sample the Final Four seeds. However, the 11-seeds have reached the Final Four far more often than the truncated geometric distribution would expect. Because of this and other overrepresented seeds for the truncated geometric fits, the bracket generation experiments in this paper use a two-stage sampling procedure which incorporates the truncated geometric distribution; see Appendix A.2 for a detailed description.

The SampleF4$_A$ function, one of two sampling functions used with $r = 5$, generates four independent and identically distributed (IID) samples from the truncated geometric distribution as the Final Four seeds. Each team in the Final Four comes from a different region, so independence is a reasonable assumption and repeat seed numbers are possible. The SampleF4$_B$ function models seeds 1–6 and 7–12 separately (and never samples seeds 13–16). For seeds 1–6, a truncated geometric distribution is fit directly, i.e. without the two-stage sampling procedure described in Appendix A.2. For seeds 7–12, a two-value distribution is used in which seeds 7, 8, and 11 are sampled with probability 2/9, and the remaining seeds are sampled with probability 1/9. To choose which set of seeds (1–6 or 7–12) to sample from, SampleF4$_B$ simply randomly chooses based on the historical proportions. For example, from 1985 through 2019, 126 of 140 Final Four teams have been seeded 1–6, so the truncated geometric distribution for seeds 1–6 is used with probability 0.9 and the two-value distribution for seeds 7–12 is used with probability 0.1. The four seeds returned by SampleF4$_B$ are then IID samples from this two-part distribution.

The SampleNCG function to be used with $r = 6$ is defined similarly, generating two IID samples from a truncated geometric distribution fit to the historical distribution of seeds reaching Round 6. However, SampleNCG must also select the region of both sampled seeds, since two regions (each with all 16 seeds) feed into each of the Round 6 slots. The region of each sampled seed is chosen by simulating a fair coin toss, implicitly assuming identical seeds as equally strong.

The IID sampling approach in SampleF4$_A$, SampleF4$_B$, and SampleNCG does not work for SampleE8 ($r = 4$), since each team is either from the *top half* of its region (seeds 1, 16, 8, 9, 5, 12, 4, 13) or the *bottom half* (seeds 6, 11, 3, 14, 7, 10, 2, 15). Therefore, two separate truncated geometric distributions are fit, one for each half-region. The SampleE8 function then generates four IID samples from each truncated geometric distribution, which assumes independence between the (half-)regions.

Table 2: Bracket generator variations for computational experiments.

Name	*P* Model	*r*	Sample
R64	Power	1	–
E8	Power	4	SampleE8
F4$_A$	Power	5	SampleF4$_A$
F4$_B$	Power	5	SampleF4$_B$
NCG	Power	6	SampleNCG
B-T	Bradley-Terry	1	–

Table 2 lists the six bracket generator variations considered in this paper, each of which uses Algorithm 1 with a particular set of parameters.

2.3 The implicit PMF

Since the power model estimates win probabilities for all possible match-ups, it implicitly defines a PMF on the sample space of all 2^{63} possible brackets. Let $\mathbf{b} = (b_1, b_2, b_3, b_4, b_5, b_6)$ be a representation of a bracket as its picks made in each of the six rounds. The estimated probability of a given bracket $\mathbf{b}$ is given by

$$\begin{aligned}\Pr(\mathbf{b}) = \Pr(b_6 \mid b_5) \cdot \Pr(b_5 \mid b_4) \cdot \Pr(b_4 \mid b_3) \\ \cdot \Pr(b_3 \mid b_2) \cdot \Pr(b_2 \mid b_1) \cdot \Pr(b_1),\end{aligned} \tag{7}$$

where $\Pr(b_i \mid b_{i-1})$ is the product of the estimated probabilities for each game in Round i (by an assumption of independence). The naïve bracket generation procedure suggested by (7) is to sort all 2^{63} brackets by decreasing probability and, given a desired pool size M, return the M most likely brackets. This is computationally infeasible, though, since 2^{63} is approximately 9 quintillion (9×10^{18}).

Exploiting the (assumed) independence of the four regions, an alternative formulation of (7) is

$$\Pr(\mathbf{b}) = \Pr(b_6 \mid b_5) \cdot \Pr(b_5 \mid b_4) \cdot \prod_{i=1}^{4} \Pr(\text{region } i), \tag{8}$$

where $\Pr(\text{region } i)$ is defined as the product of the (conditional) probabilities of the outcomes in each round, analogous to (7). This formulation suggests a different procedure: sort all 2^{15} possible region outcomes by decreasing probability, generate four regions by independent random samples from the top T most likely regions (weighted according to their estimated probabilities), and apply the power model probabilities to randomly choose winners in the Final Four and National Championship Game. The cutoff T may be as large as $2^{15} = 32,768$, but since sampling from such a large discrete distribution

is computationally expensive, a smaller choice such as $T = 100$ is more reasonable. Let MLR100 denote this most-likely-regions bracket generator with $T = 100$. The experimental performance of MLR100 will be compared to that of the R64 generator in Section 3.3.4.

2.4 Generator evaluation

The six generators listed in Table 2 are compared by generating several bracket pools with each for the 2013 through 2019 tournaments. Each bracket pool consists of N brackets generated independently with the same generator, where N is chosen to be much smaller than N_{ESPN} for any given year. Since brackets are generated independently, a bracket pool may contain duplicates. Furthermore, the score distribution for the basic power model will be compared against the score distribution for a sample from the 2016 ESPN Tournament Challenge reported by Wright and Wiens (2016).

The max score and ESPN count metrics measure the strength of each bracket pool. The *max score* metric for a bracket pool, denoted $s_{\max}$, is simply the highest score achieved in the pool. A participant in a bracket challenge wins if one of their brackets produces the highest score among all brackets submitted, so $s_{\max}$ is a reasonable measure of the quality of a bracket pool.

While attaining the unique highest score in a bracket challenge is desirable, a participant may also be content with placing on the leaderboard, especially in a large bracket challenge such as the ESPN Tournament Challenge. The ESPN Leaderboard consists of the top 100 submissions, as ranked by the ESPN scoring system, among all brackets submitted to the challenge that year. For a given year, let l_{ESPN} and u_{ESPN} be the minimum and maximum scores, respectively, in the ESPN Leaderboard, and let N_{ESPN} be the total number of submissions. The *ESPN count* metric, denoted n_{ESPN}, is the number of brackets in the pool which would have reached the ESPN Leaderboard in that year, i.e. achieved a score greater than or equal to l_{ESPN}. The ESPN count provides some indication of how a generated bracket pool compares to the entire ESPN bracket pool. Each bracket pool will also be compared to the pick favorite pool. Let p_{PF} denote the proportion of brackets in the pool which achieve a score greater than or equal to s_{PF}, the maximum score from the pick favorite pool.

Bracket pools can also be evaluated against the pick favorite pool. Let s_{PF} denote the maximum score from the pick favorite pool for a given year. The best of the pick favorite pool is the bracket which correctly picks the winners in Rounds 5 and 6, but these picks only matter if some 1-seeds advance to the Final Four. If few or no 1-seeds reach the Final Four, then the pick favorite pool scores quite poorly. In particular, if neither team in the National Championship Game is a 1-seed, then all pick favorite brackets have the same score and are guaranteed to earn zero of 640 points in the last two rounds.

Table 3 presents benchmarks from the pick favorite pool and the ESPN Leaderboard for the 2013 through 2019 tournaments. Naturally, s_{PF} tends to be lower in years with more upsets. For instance, in 2015, when three 1-seeds reached the Final Four and two of these reached the National Championship Game, $s_{\text{PF}} = 1530$, whereas in 2014, when upsets led to a National Championship Game between a 7-seed and an 8-seed, $s_{\text{PF}} = 680$. The ESPN Leaderboard score range is more consistent, though it also tends to be lower in years with more upsets. In the 2018 tournament, for example, the unprecedented upset of a 16-seed over a 1-seed and the appearance of an 11-seed in the Final Four seem to have shifted the ESPN Leaderboard score range down.

Once the max score and ESPN count metrics have been computed for each bracket pool, the six generators will be compared with respect to each metric for each year using Hsu's multiple comparisons with the best (MCB) method (Hsu 1984). The MCB method compares different *treatments* (in this case, bracket generators) by generating R replications with each treatment and computing the sample means of a specified metric. The method then constructs confidence intervals (CIs) for the difference between the population mean of the metric for each treatment and the best mean of the other treatments. From these CIs, one may draw conclusions about the relative performance of the different treatments. For example, if the CI for a treatment has a (non-inclusive) lower bound of zero, then one may conclude (with the chosen confidence level) that that treatment is the best. Similarly, an upper bound of zero provides evidence that that treatment is not the best. Bechhofer, Santner, and Goldsman (1995, Section 4.4) explains the MCB method in greater detail.

Table 3: Benchmarks from pick favorite pool and ESPN Leaderboard (ESPN.com 2013; Adler 2014; Draper 2016; Ota 2017, 2018, 2019).

Year	s_{PF}	l_{ESPN}	u_{ESPN}	N_{ESPN}
2013	1120	1590	1660	8.15 million
2014	680	1520	1730	11 million
2015	1530	1760	1830	11.6 million
2016	870	1630	1730	13 million
2017	1460	1650	1760	18.8 million
2018	1130	1550	1670	17.3 million
2019	1240	1730	1850	17.2 million

One shortcoming of the ESPN count metric is that it is difficult to interpret; since $N \ll N_{\text{ESPN}}$, directly comparing n_{ESPN} to the ESPN Leaderboard size of 100 is not informative. To allow direct comparison to the ESPN bracket pool, a scaled metric $\hat{n}_{\text{ESPN}} = n_{\text{ESPN}} \cdot (N_{\text{ESPN}}/N)$ is also computed. The scaled count $\hat{n}_{\text{ESPN}}$ estimates how many brackets would meet the ESPN Leaderboard score threshold (l_{ESPN}) if the same generator were used to produce a pool of N_{ESPN} brackets, assuming the proportion would be the same. Therefore, $\hat{n}_{\text{ESPN}}$ allows for direct comparison, because if $\hat{n}_{\text{ESPN}} \approx 100$, then the bracket pool performs about as well as the ESPN bracket pool. A value of $\hat{n}_{\text{ESPN}}$ greater than 100 suggests the generator is better (for that year) than the aggregate expertise of the fans creating the ESPN pool.

3 Results

This section first presents the power model and Bradley-Terry parameters estimated from historical data for predicting each of the last seven tournaments (2013–2019) and the next tournament (2020). Results are then reported for bracket generation experiments with both the six generators from Table 2 and the MLR100 generator defined in Section 2.3.

3.1 Power model parameters

The only parameters of the power model are the α-values. The Round 1 α-value estimates are presented in Table 4. Each column shows the α-values for predicting one of the 2013–2019 tournaments. For example, the 2016 column uses the results of tournaments from 1985 to 2015 as the data set to estimate α-values for generating brackets for 2016. An upper bound of $K = 2$ is imposed on the magnitude of the α-value estimates, affecting $\hat{\alpha}_{1,(1,16)}$ for 2013–2018 since no 16-seed had won before the 2018 tournament. The α-value estimate of 1.77 for 2019, which incorporates the 16-seed win in 2018, suggests $K = 2$ was an appropriate choice.

Changes in the α-value estimates over time yield insights into recent tournament trends. For example, the increasing values of $\hat{\alpha}_{1,(8,9)}$ from 2013 to 2016 indicate the 8-seeds performed better in 2013–2015, which flipped the sign in 2016. Since $\hat{\alpha}_{1,(8,9)}$ remains close to zero, the outcome is effectively a coin toss. Furthermore, the decreasing values of $\hat{\alpha}_{1,(6,11)}$ show an improving performance of 11-seeds in recent tournaments.

Table 5 presents the weighted averages $\hat{\alpha}_j$ for Rounds 2–6. Note that $\hat{\alpha}_j$ is computed after the ± 2 cutoff is applied to each match-up's α-value estimate as needed. The α-value estimates for Round 4 are close to zero, indicating the estimated winning probability of each seed is close to 0.5 and the performance of the teams is (nearly) independent of the seed number. For example, in three of the four games of Round 4 in the 2013 tournament, the weaker seed defeated the stronger one. The large proportion of upsets in Round 4 repeated in 2014, causing the drop in $\hat{\alpha}_4$ from 2013 to 2015. In every year, $\hat{\alpha}_4 < \hat{\alpha}_5 < \hat{\alpha}_6$, which represents the better performance of stronger seeds in later rounds.

Table 4: Round 1 α-value estimates for the 2013–2020 tournaments.

Parameter	2013	2014	2015	2016	2017	2018	2019	2020
$\hat{\alpha}_{1,(1,16)}$	2	2	2	2	2	2	1.77	1.78
$\hat{\alpha}_{1,(2,15)}$	1.45	1.36	1.38	1.40	1.34	1.36	1.38	1.39
$\hat{\alpha}_{1,(3,14)}$	1.16	1.14	1.13	1.07	1.06	1.08	1.10	1.13
$\hat{\alpha}_{1,(4,13)}$	1.10	1.10	1.13	1.17	1.16	1.19	1.15	1.14
$\hat{\alpha}_{1,(5,12)}$	0.76	0.69	0.62	0.68	0.66	0.68	0.73	0.67
$\hat{\alpha}_{1,(6,11)}$	1.10	1.12	1.08	1.04	0.95	0.87	0.84	0.87
$\hat{\alpha}_{1,(7,10)}$	1.12	1.18	1.23	1.29	1.25	1.30	1.34	1.22
$\hat{\alpha}_{1,(8,9)}$	−0.61	−0.59	−0.28	0.27	0	0.26	0	−0.49

Table 5: Weighted averages of the α-value estimates in Rounds 2–6 for the 2013–2020 tournaments tournaments.

Parameter	2013	2014	2015	2016	2017	2018	2019	2020
$\hat{\alpha}_2$	1.10	1.03	1.03	1.02	1.01	1.05	1.03	1.09
$\hat{\alpha}_3$	0.91	0.90	0.86	0.88	0.90	0.91	0.86	0.85
$\hat{\alpha}_4$	0.36	0.23	0.19	0.22	0.14	0.12	0.15	0.11
$\hat{\alpha}_5$	0.67	0.70	0.58	0.61	0.64	0.67	0.73	0.62
$\hat{\alpha}_6$	1.41	1.42	1.44	1.44	1.17	1.17	1.20	1.23

In most cases, incorporating the results of a single tournament does not have a large impact on the estimated α-values. However, if the outcome of the games between two seeds in a tournament deviates substantially from the prior frequency of observed events, the α-value estimates may change noticeably when that tournament is added to the training data set. For example, every 8-seed defeated every 9-seed in Round 1 of the 2015 tournament, flipping the sign of $\hat{\alpha}_{1,(8,9)}$ for 2016.

Although the 2020 tournament has not occurred as of the time of writing, Tables 4 and 5 include entries for 2020 to show the α-values that will be used with the power model to predict outcomes in the 2020 tournament.

3.2 Bradley-Terry model parameters

The sixteen seed-strength parameters for the Bradley-Terry model are estimated by a standard iterative maximum likelihood estimation algorithm (Hunter 2004). Table 14 in Appendix A.3 presents the Bradley-Terry parameter estimates for predicting the 2013 through 2020 tournaments. Figure 1 places the Bradley-Terry win probabilities for predicting games in Round 1 of the 2020 tournament alongside the corresponding power model probabilities. The estimates are similar for most match-ups, but the Bradley-Terry model has noticeably lower estimates of $p_{1,(6,11)}$ and $p_{1,(7,10)}$ than the power model, meaning the Bradley-Terry model will predict more upsets in these match-ups. The models also disagree on the (8, 9) match-up, with the power (resp., Bradley-Terry) model slightly favoring 9-seeds (resp., 8-seeds), but both probability estimates are within 0.025 of a fair coin toss.

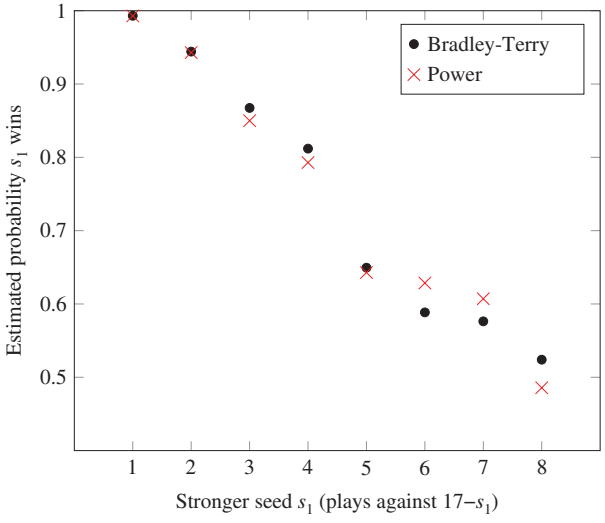

Figure 1: Estimates of $p_{1,(\cdot,\cdot)}$ for power and Bradley-Terry models for 2020 predictions.

3.3 Bracket generation experiments

This subsection uses the six generators in Table 2 to independently generate $R = 25$ batches of $N = 50{,}000$ bracket pools for each of the seven most recent tournaments (2013–2019). The methods described in Section 2.4 are then used to measure the strength of the bracket pools and assess the performance of the six generators.

3.3.1 MCB with max score and ESPN count metrics

The generators are first compared using the MCB method with the max score and ESPN count metrics with the standard 95% confidence level. Table 6 provides the average max scores $\bar{s}_{\max}$ for each set of bracket pools. The highest average in each year is in bold. Italicized entries indicate the MCB method concluded the corresponding generator was not the best for that year (i.e. its CI was strictly negative). The MCB method did not produce any strictly positive CIs, so no generator is conclusively the best (with respect to max score) for any given year. Furthermore, every generator except R64 attained the highest average max score in at least 1 year (2 years each for B-T and E8), suggesting the generators differ in meaningful ways. For 2014, the tournament in which a 7-seed defeat an 8-seed in the NCG, the MCB method eliminates all but the E8 and NCG generators, suggesting these two may be particularly well-suited for years with big upsets.

Table 7 presents the average ESPN counts $\bar{n}_{\text{ESPN}}$ for each set of bracket pools. The highest average ESPN count

Table 6: Max score sample averages and MCB results.

Year	B-T	R64	E8	$F4_A$	$F4_B$	NCG
2013	*1553.6*	*1592.0*	1598.8	**1617.2**	*1592.4*	*1592.8*
2014	*1380.0*	*1381.2*	**1465.6**	*1383.2*	*1391.6*	1464.4
2015	**1678.0**	1667.6	1663.6	1674.4	1672.8	1670.4
2016	**1640.0**	1626.0	1638.0	1626.4	1632.0	1630.4
2017	1684.4	1677.6	**1696.8**	1678.4	1687.6	*1670.0*
2018	1568.8	1566.4	1569.2	1578.8	**1583.6**	1570.0
2019	*1699.2*	1724.4	1719.6	1727.6	1724.8	**1736.4**

Table 7: ESPN count sample averages and MCB results.

Year	B-T	R64	E8	$F4_A$	$F4_B$	NCG
2013	*0.20*	0.84	1.16	**1.32**	0.88	0.84
2014	*0.00*	*0.04*	**0.40**	*0.00*	*0.16*	0.32
2015	**0.04**	0.00	0.00	0.00	0.00	0.00
2016	0.96	0.64	**1.12**	0.64	0.60	0.64
2017	1.92	1.64	**2.36**	2.12	**2.36**	*1.40*
2018	1.60	1.36	1.40	1.44	**1.84**	1.28
2019	*0.20*	0.60	0.48	0.80	0.76	**0.84**

for each year is in bold. Italicized entries indicate the MCB method concluded the corresponding generator was not the best for that year (i.e. its CI was strictly negative). The MCB method did not produce any strictly positive CIs, so no generator is conclusively the best (with respect to ESPN count) for any given year. Furthermore, every generator except R64 attained the highest average ESPN count in at least 1 year, suggesting (as mentioned for the max score metric) that the generators have different strengths and weaknesses. In particular, the MCB method results for 2014 mirror those for the max score metric, with only the E8 and NCG generators still in the running for best. Hence, the E8 and NCG generators seem to handle big upsets better than the other four generators.

3.3.2 Comparison with pick favorite approach

The generators are also evaluated against the pick favorite approach by computing $\bar{p}_{\text{PF}}$, the average proportion of brackets scoring at least as well as the best of the pick favorite pool, as defined in Section 2.4. Table 8 presents these pick favorite proportions for each generator and year. The largest $\bar{p}_{\text{PF}}$ for each year is in bold. In most years, the six generators achieve similar $\bar{p}_{\text{PF}}$ values, with the notable exception of B-T having a much lower $\bar{p}_{\text{PF}}$ than the other generators for all years but 2015. In most years, the NCG generator attains the highest $\bar{p}_{\text{PF}}$, though never by a large margin.

For 2016, $\bar{p}_{\text{PF}} \approx 0.14$ for every generator, suggesting that roughly one in seven of the brackets generated by each generator scores at least as well as the best of the pick favorite pool. Since the pick favorite pool contains eight brackets, this suggests all six generators are better than the pick favorite approach for predicting the 2016 tournament. For 2015, however, each generator has $\bar{p}_{\text{PF}} \approx 0.001$, so roughly one in 1000 of the brackets generated by each generator beats the pick favorite pool. In the remaining years, the proportions are roughly one in 24 (2013), one in 12 (2014), one in 264 (2017), one in 30 (2018), and one in 26 (2019). These proportions suggest the six generators considered should only be used to generate large pools; if a bracket challenge limits the number of submissions to a small value such as ten, then the pick favorite approach may be a better choice than any one of these six generators. However, if a large number of submissions is allowed, then these generators should be considered.

Further supporting these generators over the pick favorite approach, the average max scores in Table 6 are more consistent over the years than the s_{PF} scores in Table 3. This suggests these generators produce bracket distributions with sufficient variance to capture diverse sets of tournament outcomes, provided the pools are sufficiently large, whereas the pick favorite approach only performs well for a small set of tournament outcomes.

3.3.3 Additional comparisons with ESPN pool

The generated bracket pools are also compared to the ESPN pool using $\tilde{\hat{n}}_{\text{ESPN}}$, the sample average of the scaled ESPN counts $\hat{n}_{\text{ESPN}}$, given in Table 9. For 2016–2018, the scaled counts for all six generators exceed 100, suggesting the generators outperform the ESPN pool participants in these years. For 2014 and 2015, however, many of the scaled counts are zero and none exceed 100. In 2013 and 2019, the B-T generator has low scaled counts, but the five generators based on the power model achieve scaled counts over 100. These results suggest that seed numbers alone often provide sufficient information to produce strong brackets.

While the primary goal of the generators is to produce bracket pools with high max scores, examining the score distribution for a large sample provides an additional perspective on a generator's performance. Figure 2 is a reproduction of the score distribution for a 112,000-bracket sample from the ESPN Tournament Challenge in 2016 (Wright and Wiens 2016). Note that Wright and Wiens (2016) uses the name "pick-the-seeds" for the pick favorite approach and divides scores by 10. Figure 3 presents the score distribution for 1,250,000 brackets (25 replications of 50,000) generated by the R64 generator for 2016,

Table 8: Pick favorite proportion sample averages.

Year	B-T	R64	E8	$F4_A$	$F4_B$	NCG
2013	0.029	0.043	0.041	**0.047**	0.044	0.045
2014	0.070	0.078	0.085	0.080	0.083	**0.090**
2015	**0.0013**	0.00087	0.00075	0.00085	0.00092	0.00092
2016	0.129	0.145	0.134	0.141	**0.151**	0.146
2017	0.0033	0.0038	0.0036	0.0038	0.0041	**0.0041**
2018	0.032	0.034	0.028	0.033	0.035	**0.036**
2019	0.031	0.040	0.033	0.041	0.042	**0.044**

Table 9: Scaled average ESPN counts $\tilde{\hat{n}}_{\text{ESPN}}$ per generator per year.

Year	B-T	R64	E8	$F4_A$	$F4_B$	NCG
2013	32.6	136.92	189.08	215.16	143.44	136.92
2014	0	8.8	88	0	35.2	70.4
2015	9.28	0	0	0	0	0
2016	249.6	166.4	291.2	166.4	156	166.4
2017	721.92	616.64	887.36	797.12	887.36	526.4
2018	553.6	470.56	484.4	498.24	636.64	442.88
2019	68.8	206.4	165.12	275.2	261.44	288.96

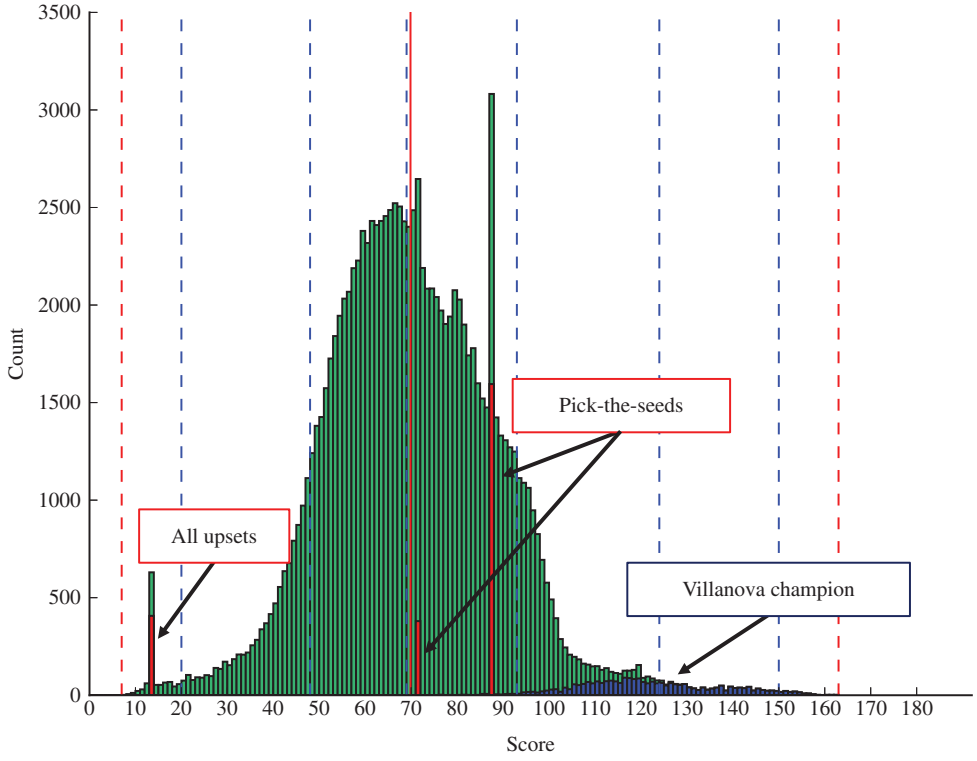

Figure 2: ESPN tournament challenge sample score distribution for 2016; reprinted with permission from Wright and Wiens (2016).

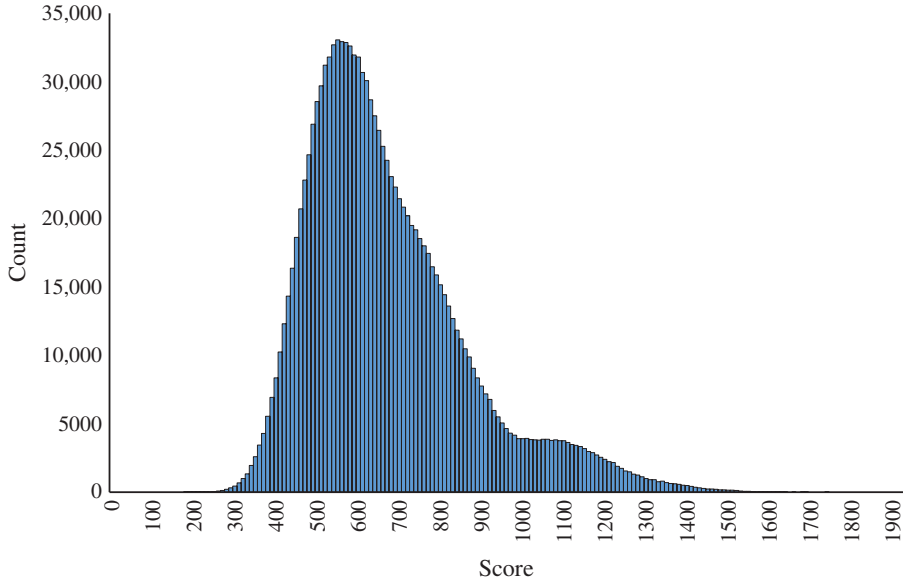

Figure 3: R64 generator sample score distribution for 2016.

approximately ten times the size of the sample from which Figure 2 was created. The R64 generator's distribution has a lower mode than the ESPN distribution, but its upper tail (above a score of 1000) contains a greater proportion of brackets, suggesting the R64 generator is more likely to generate high-scoring brackets than the general public.

3.3.4 MLR100 vs. R64 at different pool sizes

Recall the MLR100 generator defined in Section 2.3, which independently samples four region outcomes from the 100 most likely region outcomes, as estimated by the power model, and applies the power model probabilities to randomly choose winners in the last three games. Using the

power model probability estimates for predicting the 2019 tournament, the 100 most likely region outcomes all have a 1-, 2-, or 3-seed winning the region. Furthermore, the 1-seed always reaches the Elite Eight. Hence the MLR100 generator is not very robust; it will never allow a seed larger than 3 to advance to the last few rounds. In the last seven tournaments, 10 of 28 teams reaching the Final Four and 3 of 14 teams in the National Championship Game were seeded 4 or worse, so the inability of MLR100 to predict such seeds is a severe limitation.

Table 10 reports the average max scores across $R = 10$ replications of generating brackets for the 2013–2019 tournaments with the MLR100 and R64 generators. Pool sizes range from 10 to 50,000. The results vary greatly across years. In 2014, for example, when a 7-seed defeated an 8-seed in the NCG, the MLR100 generator scores very poorly. (Recall that MLR100 never sends a 4+-seed to the F4.) However, in 2019, MLR100 generates a higher average max score than R64 for all pool sizes but 5000 and 50,000. Generally, MLR100 performs better for small pools, and R64 performs better for large pools. In most years, the advantage changes around a pool size of 5000. These findings align with the intuition that generator diversity should scale with pool size to allow for appropriate exploration of the search space.

Table 10: Average max scores of MLR100 (top) and R64 (bottom) generators at different pool sizes.

Pool size	2013	2014	2015	2016	2017	2018	2019
10	1141	718	1310	1109	1272	1154	1305
	1018	733	1119	1004	1194	962	1186
25	1164	738	1440	1253	1394	1253	1387
	1187	731	1259	1136	1357	1145	1291
50	1152	756	1440	1406	1476	1221	1514
	1209	774	1358	1240	1364	1255	1347
100	1173	773	1520	1443	1501	1393	1551
	1249	813	1452	1290	1431	1266	1424
250	1188	789	1535	1501	1501	1389	1578
	1353	1010	1520	1407	1502	1347	1504
500	1209	793	1543	1532	1550	1449	1601
	1355	1033	1542	1456	1520	1398	1542
1000	1210	798	1562	1534	1548	1468	1642
	1439	1168	1559	1472	1525	1448	1589
5000	1234	806	1588	1582	1578	1481	1674
	1498	1264	1600	1565	1631	1562	1696
10,000	1241	813	1599	1598	1586	1499	1685
	1521	1274	1644	1598	1622	1535	1671
50,000	1250	824	1614	1609	1599	1517	1708
	1583	1413	1695	1635	1694	1601	1744

4 Conclusions

The NCAA Division I Men's Basketball Tournament is an annual single-elimination tournament which is the focus of significant media and fan interest in the United States. Predicting the games' results is a popular activity which has attracted the attention of academic researchers in recent years. This paper reviews the proposed models for predicting the results of the NCAA basketball tournament and introduces the power model, which estimates the teams' winning probabilities as a power function of the teams' seed numbers. The power model generalizes the Bradley-Terry model and has round-dependent parameters, an important feature given the observed tendency for relative seed strength to vary across rounds.

Brackets are generated by applying these estimates to choose game outcomes round by round. The power model is both simple and intuitive: the α-value for each match-up in each round summarizes the performance of the two seeds in the same round in all modern-era tournaments. The web-based implementation of the power model, introduced as a continuing platform to present research on a popular topic to a general audience, has gathered widespread media and public attention.

Five bracket generators using the power model exploit prior research findings on seed distributions to first select the seeds in later rounds before applying the power model to the remaining games. These five generators and a Bradley-Terry-based generator are applied to generate pools of brackets which are scored using the ESPN scoring system and assessed against the pick favorite approach and ESPN Tournament Challenge Leaderboard.

The generators are compared using Hsu's MCB method applied to the max score and ESPN count metrics. The MCB method is inconclusive for both metrics in most years, but it does provide evidence that some generators are not the best in some years.

Since at least one of the E8 and $F4_A$ generators attains the highest or second-highest average max score in each year, there is weak evidence supporting these generators over the other four. The E8 and $F4_A$ models may perform better because fixing the Final Four or Elite Eight seeds splits the bracket, perhaps reducing the likelihood of error propagation through the rounds, especially in contrast to the B-T and R64 generators. However, none of the six generators consistently outperformed the ESPN pool.

The six generators for the main experiments are assessed by generating pools of 50,000 brackets. For small pools, however, the MLR100 generator shows promise, outperforming R64 in most of the last seven tournaments

for sample sizes between 10 and 1000. It seems the MLR100 generator offers a balance between the naïve pick favorite approach and the high-variance generators derived from Algorithm 1. Tuning the parameter T for the most-likely regions approach could improve performance for specific pool sizes.

Table 11 provides a qualitative summary of the bracket generators discussed in this paper. Future work could seek to characterize the tournament conditions in which certain generators perform better or worse than others. If a set of generators is found such that at least one generator performs well for each year, then a hybrid pool containing some brackets from each generator in the set may achieve more consistent success for future tournaments.

The power model is naturally limited by the small size of data sets for match-ups, especially in later rounds. To address this problem, future work may consider using the results of a seed in one match-up to estimate its winning probability in other match-ups. For instance, a weighted average of the α-values for the stronger seed in the round may be used rather than the weighted average of all α-values for the round. Moreover, estimating the α-values by a weighted function of the tournament results might improve the models, since more recent tournaments may better represent the relative strength of seeds.

While generators based on the power model do not always produce extremely strong brackets, they perform surprisingly well given that the power model reduces each team to its seed number and uses only this limited data set. Also, the flexibility of the power model makes it a better choice than the Bradley-Terry model, which is a special case. McCrea and Hirt (2009) notes that probability matching, which the power model effectively does in its α-value estimation procedure, does not always lead to the most effective results. It is reasonable to suppose there exist optimal power model parameters which differ from those produced by both the probability-matching procedure and the Bradley-Terry model. Further work could seek alternative estimation procedures in pursuit of these optimal parameters.

Table 11: Summary of generators.

Generator	Variance	Ideal pool size	Performance on metrics	
			Max score	ESPN count
B-T	High	Large (>1000)	Varies by year	Varies by year
R64	High	Large (>1000)	Poor	Fair
E8	Medium	Large (>1000)	Good	Good
$F4_A$	Medium	Large (>1000)	Good	Good
$F4_B$	Medium	Large (>1000)	Fair	Good
NCG	Medium	Large (>1000)	Fair	Good
MLR100	Low	Small (<1000)	Varies by year	Not measured

Acknowledgement: This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE–1746047. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The fourth author has been supported in part by the Air Force Office of Scientific Research under Grant No. FA9550-19-1-0106. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Government, or the Air Force Office of Scientific Research.

A Appendices

A.1 Tournament round names

Table 12 lists the six main rounds of the tournament with their various names.

A.2 Description of two-stage sampling procedure

Each of the sampling functions SampleE8, SampleF4_A, SampleF4_B, and SampleNCG incorporates a truncated geometric distribution based on the results in Jacobson et al. (2011). When directly fitting a truncated geometric

Table 12: The six main rounds of the NCAA division I men's basketball tournament (NCAA 2019).

Round	No. of teams	Official NCAA name	Colloquial name	Abbrev.
1	64	First round	Round of 64	R64
2	32	Second round	Round of 32	R32
3	16	Regional semifinals	Sweet Sixteen	S16
4	8	Regional finals	Elite Eight	E8
5	4	National semifinals	Final Four	F4
6	2	Natl. championship	–	NCG

Table 13: Adjusted seeds for two-stage sampling procedures.

Sampling function	Adjusted seed
SampleE8 (top)	1
SampleE8 (bottom)	11
SampleF4_A	11
SampleF4_B	None
SampleNCG	8

Table 14: Parameters for Bradley-Terry model.

Year	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
2013	0.2228	0.1422	0.0960	0.0802	0.0723	0.0739	0.0568	0.0455
2014	0.2253	0.1362	0.0911	0.0890	0.0750	0.0686	0.0523	0.0488
2015	0.2231	0.1375	0.0889	0.0873	0.0705	0.0687	0.0558	0.0530
2016	0.2279	0.1369	0.0875	0.0857	0.0700	0.0669	0.0590	0.0544
2017	0.2233	0.1445	0.0879	0.0818	0.0670	0.0675	0.0624	0.0514
2018	0.2222	0.1443	0.0906	0.0806	0.0635	0.0681	0.0661	0.0513
2019	0.2237	0.1412	0.0931	0.0766	0.0616	0.0694	0.0657	0.0519
2020	0.2221	0.1417	0.0965	0.0769	0.0637	0.0686	0.0639	0.0496
Year	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
2013	0.0401	0.0458	0.0446	0.0386	0.0196	0.0145	0.0070	0.0000
2014	0.0446	0.0419	0.0409	0.0418	0.0220	0.0140	0.0085	0.0000
2015	0.0461	0.0432	0.0424	0.0404	0.0208	0.0139	0.0084	0.0000
2016	0.0451	0.0437	0.0418	0.0386	0.0198	0.0145	0.0081	0.0000
2017	0.0436	0.0474	0.0432	0.0370	0.0190	0.0148	0.0093	0.0000
2018	0.0425	0.0484	0.0457	0.0348	0.0180	0.0148	0.0090	0.0000
2019	0.0452	0.0473	0.0489	0.0326	0.0177	0.0148	0.0086	0.0016
2020	0.0450	0.0470	0.0480	0.0344	0.0178	0.0148	0.0084	0.0015

distribution to the historical seed distribution, it is often the case that one seed appears far more often historically than a truncated geometric fit would predict. Hence before each truncated geometric fit, the frequency of the most overrepresented seed s is reduced from f_s to f_s'. Table 13 lists which seeds are modified for each of the sampling functions. The truncated geometric fit then underpredicts the appearance of seed s, so a preliminary stage is added to the sampling procedure in which seed s is chosen with some constant probability p_s and the truncated geometric distribution is sampled otherwise (with probability $1 - p_s$). The value of p_s is chosen to match the total probability of sampling s to the historical proportion. Observe that p_s is fully determined by f_s'. Since fitting the truncated geometric distribution is a deterministic process, the entire two-stage sampling procedure is fully determined by f_s'. To obtain the best fit to the historical distribution, f_s' is chosen to minimize the χ^2 goodness-of-fit test statistic for the sampling distribution against the historical distribution.

Defining the sampling functions in this manner offers several advantages:

- Replicability – the parameters are entirely determined by the historical distribution.
- Flexibility – the sampling functions for future years are easily constructed.
- Goodness-of-fit – adjusting for the most overrepresented seed yields a better fit than the truncated geometric distribution alone.

A.3 Bradley-Terry model parameters

Table 14 presents the Bradley-Terry seed-strength parameter estimates for predicting game outcomes in the 2013 through 2020 tournaments. Recall from Section 2.1 that β_i is the strength parameter for seed i.

References

Adler, K. 2014. *All-Time Record High 11.01 Million Brackets Submitted in ESPN.com's Men's Tournament Challenge Game.* https://espnmediazone.com/us/press-releases/2014/03/all-time-record-high-11-01-million-brackets-submitted-in-espn-coms-mens-tournament-challenge-game/, accessed on 20 June 2018.

Bechhofer, R. E., T. J. Santner, and D. M. Goldsman. 1995. *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*. 605 Third Avenue, New York, NY 10158-0012: John Wiley & Sons, Inc.

Boulier, B. L. and H. O. Stekler. 1999. "Are Sports Seedings Good Predictors?: An Evaluation." *International Journal of Forecasting* 15:83–91.

Bradley, R. A. and M. E. Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. the Method of Paired Comparisons." *Biometrika* 39:324–345.

Draper, M. 2016. *Tournament Challenge: Final Pick Results and Bracket Numbers*. http://www.espn.com/blog/collegebasketballnation/post/_/id/112638/tournament-challenge-final-pick-results-and-bracket-numbers, accessed on 20 June 2018.

ESPN.com. 2013. *Tournament Challenge: 1 Perfect Bracket.* http://www.espn.com/blog/collegebasketballnation/post/_/id/81033, accessed on 20 June 2018.

Gupta, A. A. 2015. "A New Approach to Bracket Prediction in the NCAA Men's Basketball Tournament Based on a Dual-Proportion Likelihood." *Journal of Quantitative Analysis in Sports* 11:53–67.

Hsu, J. C. 1984. "Constrained Simultaneous Confidence Intervals for Multiple Comparisons with the Best." *The Annals of Statistics* 12:1136–1144.

Hunter, D. R. 2004. "MM Algorithms for Generalized Bradley-Terry Models." *The Annals of Statistics* 32:384–406.

Jacobson, S. H. and D. M. King. 2009. "Seeding in the NCAA Men's Basketball Tournament: When is a Higher Seed Better?" *Journal of Gambling Business and Economics* 3:63–87.

Jacobson, S. H., A. G. Nikolaev, D. M. King, and A. J. Lee. 2011. "Seed Distributions for the NCAA Men's Basketball Tournament." *Omega* 39:719–724.

Kaplan, E. H. and S. J. Garstka. 2001. "March Madness and the Office Pool." *Management Science* 47:369–382.

Kenter, F. H. J. 2016. "How to Win Your Betting Pool with Jensen's Inequality and the Law of Large Numbers." *The American Mathematical Monthly* 123:592–596.

Khatibi, A., D. M. King, and S. H. Jacobson. 2015. "Modeling the Winning Seed Distribution of the NCAA Division I Men's Basketball Tournament." *Omega* 50:141–148.

Koenker, R. and G. W. Bassett Jr. 2010. "March Madness, Quantile Regression Bracketology, and the Hayek Hypothesis." *Journal of Business & Economic Statistics* 28:26–35.

Kvam, P. and J. S. Sokol. 2006. "A Logistic Regression/Markov Chain Model for NCAA Basketball." *Naval Research Logistics (NRL)* 53:788–803.

Lopez, M. J. and G. J. Mathews. 2015. "Building an NCAA Men's Basketball Predictive Model and Quantifying its Success." *Journal of Quantitative Analysis in Sports* 11:5–12.

McCrea, S. M. and E. R. Hirt. 2009. "Match Madness: Probability Matching in Prediction of the NCAA Basketball Tournament." *Journal of Applied Social Psychology* 39:2809–2839.

NCAA. 2019. 2019 *Men's Final Four Records Book.* http://fs.ncaa.org/Docs/stats/m_final4/2019/MFFBook.pdf, accessed on 06 August 2019.

Ota, K. 2017. *ESPN Tournament Challenge Explodes to Record 18.8 Million Brackets*. https://espnmediazone.com/us/press-releases/2017/03/espn-tournament-challenge-explodes-record-18-8-million-brackets/, accessed on 20 June 2018.

Ota, K. 2018. 21*st ESPN Tournament Challenge Collects 17.3 Million Brackets.* https://espnmediazone.com/us/press-releases/2018/03/21st-espn-tournament-challenge-collects-17-3-million-brackets/, accessed on 20 June 2018.

Ota, K. 2019. *Not so Perfect: Only 0.25% of 17.2 Million Brackets Remain Perfect in ESPN's Tournament Challenge Game.* https://espnpressroom.com/us/press-releases/2019/03/not-so-perfect-only-0-25-of-17-2-million-brackets-remain-perfect-in-espns-tournament-challenge-game/, accessed on 07 June 2019.

Ruiz, F. J. R. and F. Perez-Cruz. 2015. "A Generative Model for Predicting Outcomes in College Basketball." *Journal of Quantitative Analysis in Sports* 11:39–52.

West, B. T. 2006. "A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament." *Journal of Quantitative Analysis in Sports* 2:3.

West, B. T. 2008. "A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament: Updated Results from 2007." *Journal of Quantitative Analysis in Sports* 4:8.

Wright, M. and J. Wiens. 2016. "Method to their March Madness: Insight from Mining a Novel Large-Scale Dataset of Pool Brackets." *KDD Workshop on Large-Scale Sports Analytics*. https://pdfs.semanticscholar.org/2ebf/f57d36cc583c90e52fe32f0b96ca9a23118f.pdf.

Yuan, L. H., A. Liu, A. Yeh, A. Kaufman, A. Reece, P. Bull, A. Franks, S. Wang, D. Illushin, and L. Bornn 2015. "A Mixture-of-Modelers Approach to Forecasting NCAA Tournament Outcomes." *Journal of Quantitative Analysis in Sports* 11: 13–27.