

A Comparison of Potential Playoff Systems for NCAA I-A Football

David H. Annis* Samuel S. Wu†

Abstract

Properties of various knockout tournament designs are discussed and theoretical results presented. Potential playoff schemes for Division I-A football are examined via simulation studies. Several metrics are used to assess the relative merits of playoff scenarios, which differ in number, selection and seeding of playoff teams. Most suggest that college football would benefit from a limited playoff system. Interestingly, for the class of playoff systems examined, the number of teams influences the performance far more than does the seeding procedure.

1 Introduction

College football remains the only Division I-A varsity sport which does not crown its national champion in a post-season tournament. Instead, throughout much of its history, the “national champion” (or champions, in some cases) has been determined by media and coaches’ polls. With the advent of the Bowl Championship Series (BCS) in 1998, a combination of polls and model-based ranking methods has been used to determine two entrants in a championship game. Although some feel that this system represents an improvement over the traditional bowl system, it is not without controversy. A thorough explanation of the BCS system and its shortcomings is given by Stern (2004) and discussed extensively by Billingsley (2004), Colley (2004), Harville (2004), Massey (2004) and Mease (2004).

It is clear that the NCAA’s constant “tweaking” of the BCS system has not been able to determine a champion reliably or equitably. All of these controversies could have been avoided were there a playoff system in place. In light of this, we propose a number of different playoff schemes and investigate their performance via extensive simulation studies. Carlin and Stern (1999) utilize different methods than those given herein to evaluate potential single-elimination college football tournaments.

The presentation is organized as follows. Section 2 discusses issues related to tournament design, specifically the number of teams admitted and the match-ups involving those teams. Section 3 explains the simulation methods used to evaluate potential playoff formats, the results of which are given in section 4. Finally, in section 5, we conclude with an interpretation of results and a discussion of potential improvements to the NCAA’s current system.

2 Tournament Design

We restrict attention to *knockout* tournaments, in which the loser of each game is eliminated from future competition. The sole unbeaten team remaining at the end of the tournament is the winner. When designing a tournament, one must balance the importance of a strong regular season performance (qualifying for the tournament) with that of an end-of-season winning streak in the tournament.

*David H. Annis is Assistant Professor, Operations Research Department, Naval Postgraduate School, Monterey, CA 93943 (Email: annis@nps.edu).

†Samuel S. Wu is Assistant Professor of Biostatistics, College of Medicine, University of Florida, Gainesville, FL 32610 and Biostatistician at the Malcom Randall VA Medical Center, Gainesville, FL 32608.

2.1 Number of Entrants

Large playoff fields devalue the regular season and can produce championship teams that most would agree were not the “best.” For example, consider the season in which the NCAA expanded the basketball tournament to 64 teams, 1985. Villanova (25-10) won the national championship game over defending champ and top-ranked Georgetown (35-3), after losing both of their head-to-head regular-season meetings. In fact, since the NCAA expanded its basketball tournament to 64 or more teams, only 10 of 21 “national” champions earned their conference’s championship outright.

On the other hand, small playoff fields make it difficult for even top-notch teams to qualify and therefore risk excluding the top teams. For instance, in that same 1985 season, Georgetown would not have qualified for a postseason tournament which included only regular-season conference champions — despite finishing with the #1 national ranking and a better overall record, Georgetown (14-2) finished a game behind St. John’s (15-1) in the Big East conference. A good tournament must balance the probability that the top team qualifies with the probability that it wins once the tournament begins.

2.2 Three Seeding Methods

In most tournaments, opponent pairings are not determined at random. Rather, the participants are *seeded* or assigned to a particular position in the tournament draw based on their perceived merit. Virtually all standard seeding methods accomplish two simultaneous objectives: (1) better teams are rewarded with easier draws than worse teams; and (2) top teams will not face each other until late in the tournament.

For tournaments of a given size, a standard, fixed method for seeding a tournament can be defined recursively as follows: *assuming no upsets occur*, when 2^k ($k \in \mathbb{N}$) competitors remain, opponents’ seeds satisfy $i + j = 2^k + 1$. Figure 1 gives an example for an eight-team ($k = 3$) tournament; we will denote this particular seeding by $[[18][45]] [[27][36]]$, where $[ij]$ denotes a head-to-head game between teams i and j . Thus, the top half of the bracket, $[[18][45]]$, contains match-ups between the first and eighth seeds as well as the fourth and fifth seeds, with the winners to play each other in the second round. This traditional method is intuitively appealing and is probably most widely used in tournaments, e.g., in the NCAA basketball tournament. However, the standard seeding is not monotone, i.e., a better, higher-seeded team may have a lower probability of winning the championship than does a weaker, lower-seeded team.

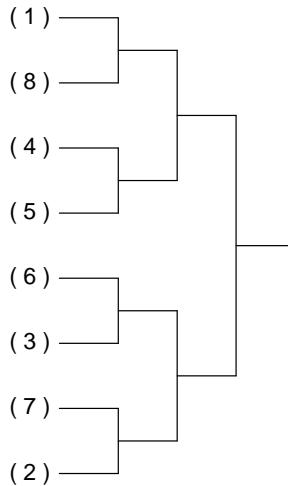


Figure 1: An eight-team standard tournament draw.

Let $p_{i,j}$ denote the probability that team i defeats team j in a game. The probability matrix, $P = [p_{i,j}]$, is said to satisfy *strong stochastic transitivity* (SST), as in David (1963), if for each triplet $\{i, j, k\}$, if $p_{i,j} \geq 1/2$ and $p_{j,k} \geq 1/2$, then $p_{i,k} \geq \max\{p_{i,j}, p_{j,k}\}$. Equivalently, for every i and j , $p_{i,k} > p_{j,k}$ for some k implies

$p_{i,k} \geq p_{j,k}$ for every k (Hwang, 1982). SST ensures that for some ordering of teams, P is doubly monotonic, and consequently there is a unique, unambiguous ranking of teams. Hwang (1982) and Schwenk (2000) both present SST preference schemes for which team 2 has a higher probability of winning the tournament than does team 1 under the standard seeding, despite team 1’s superior ability and ostensibly more favorable draw. Each proposes a tournament structure which remedies this deficiency of the standard tournament draw. Schwenk (2000) discusses two axioms which “good” tournaments should satisfy:

SR **Sincerity Rewarded:** A higher-seeded team should never be penalized by being given a schedule more difficult than that of any lower seed.

DC **Delayed Confrontation:** Two teams rated among the top 2^j shall not meet until the field has been reduced to 2^j or fewer teams.

Figure 2(a) gives an example of a four-team playoff which violates both conditions. Sincerity is not rewarded since either of the top two seeds would benefit from being seeded third. In addition, confrontation is not delayed since the two best teams face each other immediately rather than in a potential championship game. Both problems are remedied by the draw given by Figure 2(b).

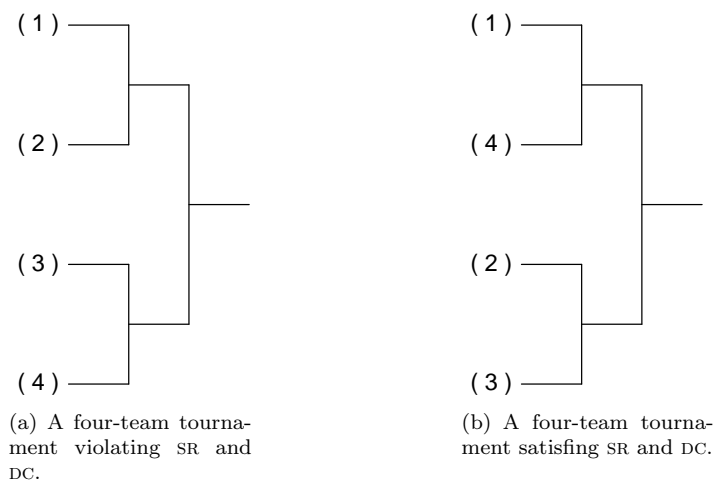


Figure 2: Four-team tournaments illustrating SR and DC.

In addition to Schwenk’s axioms, we propose explicit conditions on the tournament’s balance.

AB **Absolute Balance:** Every team must face the same number of opponents in order to win (i.e., there are no byes).

RB **Relative Balance:** For every pair of teams in the draw, neither should need to face more than one additional opponent than the other in order to win (i.e., no team should receive more than one bye).

Many tournaments are absolutely balanced, however, some, such as the National Football League (NFL) playoffs are merely relatively balanced. In the case of the NFL, six teams from each conference qualify for the playoffs with the top two receiving byes into the second round. The following “backward elimination” tournament shown in Figure 3 illustrates why some degree of balance is desirable when seeding tournaments. (Variants of this design are used in some bowling tournaments where the phrase “climbing the ladder” refers to a low-seeded competitor defeating many higher seeds in succession to win. As expected, such an occurrence is rare.)

Although the example satisfies Schwenk’s (2000) DC and SR axioms, the lack of balance makes it undesirable. (We note, however, that Schwenk’s desire to minimize favoritism would preclude such a draw.) The large number of games involving only lower-ranked teams would certainly detract from its commercial appeal. Furthermore, some may complain that the top seeds receive excessive favoritism at the expense of

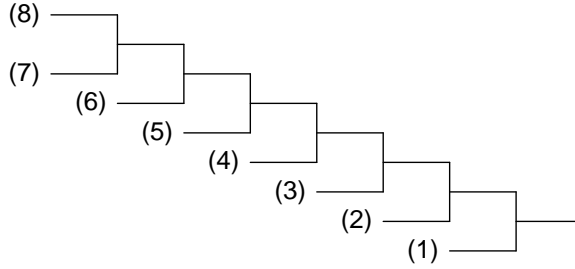


Figure 3: Tournaments satisfying delayed confrontation and sincerity rewarded may be far from balanced.

lower ones. Henceforth, we propose tournament designs that satisfy the relative balance axiom and can, therefore, accommodate tournament fields of any size, such as the NFL's. Although we require only relative balance, ensuing discussions assume absolute balance. They immediately extend to RB tournaments by including the necessary number of teams — each having zero ability, and thus being guaranteed to lose — needed to bring the number of entrants to 2^k for some positive integer k . (In practice these games are simply byes.)

To seed $n = 2^k$ teams, Schwenk (2000) defines k cohorts with cohort C_i ($1 \leq i \leq k$) consisting of seeds $\{2^{i-1} + 1, \dots, 2^i\}$ and randomly places teams within each cohort. His procedure can be thought of as randomly assigning the seeds in each cohort to the teams of which it is comprised and subsequently following the standard, fixed draw. He calls this *cohort randomized seeding*, and proves that it produces the tournament which is most equitable (as compared to a truly random tournament) subject to the DC and SR axioms. Figure 4 gives an eight-team illustration. The cohorts are $C_1 = \{1, 2\}$, $C_2 = \{3, 4\}$ and $C_3 = \{5, 6, 7, 8\}$. Teams 3 and 4 will be randomly placed in the two C_2 slots; and teams 5 through 8 in those reserved for C_3 . Due to the complete randomization in the other cohorts, equivalent draws result in deterministically placing teams 1 and 2 in either of the C_1 positions. Figure 4(a) shows the *fixed* positions occupied by the cohorts, and Figure 4(b) gives a possible realization of the *random* draw within cohorts.

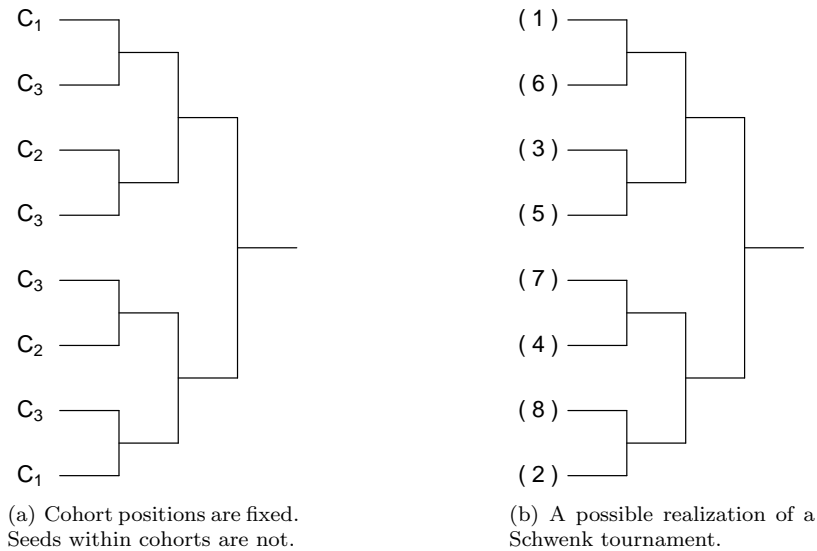


Figure 4: Schwenk's cohort randomized seeding minimizes favoritism.

Hwang (1982), on the other hand, eschews the idea of favoritism and focuses on monotonicity so that the probability of a given team winning the tournament exceeds the maximum of the probabilities of those teams seeded below it. He does this by proposing that teams be reseeded after each round of competition such that the highest remaining seed faces the lowest, the second-highest faces the second-lowest, and so

forth. For example, suppose in an eight-team draw, the first round winners are $\{1,4,6,7\}$. Under a fixed tournament seeding (see Figure 1), the best remaining teams (1 and 4) would face in the semi-finals as would the worst two. This method ensures not only that one of the two best remaining teams is eliminated, but also that one of the two worst advances to the finals. Hwang’s method results in more appealing semi-final matches, $[[17][46]]$, as evidenced in Figure 5.

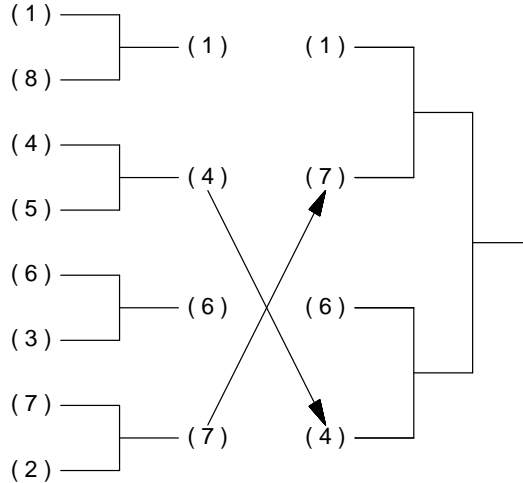


Figure 5: Hwang reseeding ensures monotonicity and favors high seeds.

While Schwenk’s procedure minimizes favoritism accorded the top seeds, Hwang’s tends to maximize it. For small tournaments, this can be quantified explicitly. We state without proof the following properties of Hwang’s seeding method:

1. For an AB four-team playoff with any SST preference scheme, Hwang seeding maximizes the probability that the best team wins the tournament.
2. For an AB eight-team playoff satisfying SR and DC with fixed first-round pairings, Hwang seeding maximizes the probability that the best team wins the tournament.

For larger tournaments, Hwang’s method tends to maximize favoritism for the top seed, however counter-examples can be constructed for which it does not.

Each of the seeding procedures discussed has its merits. The fixed draw is ubiquitous, widely-accepted and easy to implement. Schwenk’s (2000) method minimizes favoritism, while Hwang’s (1982) maximizes it in some instances. The choice of seeding may depend on how much favoritism for the top seeds is desired. Hwang reseeding provides a large degree of favoritism and Schwenk’s cohort seeding a minimal degree, with fixed seeding often somewhere in between.

3 Simulation Method

Simulation studies were used to test the effectiveness of each of the proposed tournament constructions. In all cases, games were considered to be independent Bernoulli trials with probabilities depending on the teams involved and the venue. The team-specific strength parameters and the effect of home-field advantage were estimated jointly using a Bayesian modification of the Bradley-Terry model (described below). After determining the posterior distribution of the model parameters based on 2004 data, future seasons were simulated from the posterior predictive distribution (see, e.g., Gelman et al., 1995).

3.1 Parametric Strength Model

Bradley and Terry (1952) impose additional structure on the SST probability matrix by positing that each team has an inherent ability, $\xi_i > 0$, and that given those abilities, the probability that team i defeats team j is

$$p_{i,j} = \frac{\xi_i}{\xi_i + \xi_j}. \quad (1)$$

The canonical parameterization is on the log-odds scale, for which

$$\log(\text{odds})_{i,j} = \log\left(\frac{p_{i,j}}{p_{j,i}}\right) = \log(\xi_i) - \log(\xi_j) = \alpha_i - \alpha_j,$$

where $\alpha_i = \log(\xi_i)$. Note that since the parameters are unique up to an additive constant, a constraint is required to define a unique solution. We choose to let $\sum_{i=1}^n \alpha_i = 0$, or equivalently $\prod_{i=1}^n \xi_i = 1$. In addition, home-field advantage (δ) can be included as an offset in the generalized linear model. The probability that i defeats j at home is, therefore,

$$p_{i,j} = \frac{e^\delta \xi_i}{e^\delta \xi_i + \xi_j} \quad (2)$$

$$\log(\text{odds})_{i,j} = \alpha_i - \alpha_j + \delta.$$

Agresti (1990, chap 10) illustrates this model for baseball games. Occasionally, I-A teams play against teams from lower divisions: I-AA, II, etc. Rather than discard these games (which would benefit those division I-A teams that lost to lower-division opponents), we choose to include an extra composite “non I-A” team.

Although straightforward, the Bradley-Terry model is often inappropriate for college football data, as the maximum likelihood estimate of abilities for unbeaten (winless) teams are infinite (zero). This results in predicting an undefeated team will never lose a future game, while a winless team will never win one. Davidson and Solomon (1973) take a Bayesian approach to the Bradley-Terry model and introduce a class of conjugate priors on the team strengths. In addition, we place an independent uniform prior on the home-field advantage parameter, δ . This results in the prior given by

$$\pi(\xi, \delta) \propto \prod_{i < j} \left(\frac{\xi_i}{\xi_i + \xi_j}\right)^{w_{ij}^0} \left(\frac{\xi_j}{\xi_i + \xi_j}\right)^{w_{ji}^0}. \quad (3)$$

The prior distribution may be interpreted as outcomes of hypothetical games between teams, where w_{ij}^0 represents the number of prior wins for i against j . Degenerate solutions (i.e., infinite or zero-valued MLEs) can be avoided by requiring $\sum_j w_{ij}^0 > 0$ and $\sum_j w_{ji}^0 > 0$ for all i . For our simulations, we let $w_{ij}^0 = 1/[2(n-1)]$ for all $i \neq j$.

Since the choice of prior (3) contains no information about the home-field advantage parameter, it can be viewed as a neutral-site, round-robin with total weight equal to one game per team. While preferring no team over another, this prior is informative in the sense that it shrinks all teams toward equality and, therefore, incorporates the prior belief that no team is infinitely better (or worse) than others.

3.2 Simulation Studies

Unfortunately, teams’ rankings and their probabilities of defeating potential opponents are unknown. Therefore, to assess various playoff systems, we use a number of simulation studies with the true parameters generated from the normal approximation to the posterior distribution (see Gelman et al., 1995, chap 4) given data from the 2004 college football season.

First, we obtained data from all games in the 2004 season which involved at least one Division I-A team. (Rather than ignore games in which a Division I-A played a team from a lower division, we chose to treat

all lower division opponents as a single non I-A team.) Then, we derived the normal approximation of the posterior distribution of $\theta = (\alpha_1, \alpha_2, \dots, \alpha_{n-1}, \delta)^T$. The Laplace approximation (Tierney and Kadane, 1986) was used and the resulting distribution is denoted by $N(\hat{\theta}, \hat{\Sigma})$, where $\hat{\theta}$ is the posterior mode (the parameter value which maximizes the posterior density) and $\hat{\Sigma}$ is the covariance matrix.

Game results were simulated using an approximation to the posterior predictive distribution. This was achieved in two stages. First the season-specific vector of team strengths and the home-field advantage were sampled from $N(\hat{\theta}, \hat{\Sigma})$. Once the parameters were drawn for a season, game results were simulated as independent Bernoulli trials with probability of the home team winning given by (2); for neutral site games probabilities followed (1). This two-stage process was repeated for 10,000 simulated seasons. We call this simulation the *Standard* setup, as it mimics the observed 2004 season.

One might wonder if there were features unique to the 2004 season which could influence the analysis. Therefore, to mitigate the season-specific features of these simulations, different variations of the approach were investigated.

- *Competitive*: The ability of the regular season and subsequent playoff to determine the correct #1 team depends heavily on how much better the #1 team is than the others. When one team dominates all others, any method will often correctly identify the champion. However, when the difference between teams' abilities is slight, the choice of procedure becomes more important. In light of this, a second simulation was conducted with vector of parameters generated from $N(\hat{\theta}/2, \hat{\Sigma}/4)$. Obviously, this is equivalent to generating θ from $N(\hat{\theta}, \hat{\Sigma})$ and subsequently multiplying all parameter values by 1/2. This transformation shrinks the log odds toward 0, making games more competitive. It performs a square root transformation on the team strengths, shrinking all ξ_i parameters toward 1 due to the constraint that $\prod_{i=1}^n \xi_i = 1$. Therefore, an alternative interpretation is that this reduces differences between teams and simultaneously lessens the advantage of playing at home. Once again, after generating the parameter values, results of an entire season were simulated using the Bradley-Terry model. This simulation also consists of 10,000 seasons.
- *Permuted Strengths*: To ensure that the observed behavior was due to the proposed tournament designs and nothing particular to the 2004 schedule, we permuted the team IDs after generating parameters from $N(\hat{\theta}, \hat{\Sigma})$. Thus, although the schedules were nominally identical (e.g., Florida still plays Georgia) the strength parameters could have originally corresponded to other teams. The effect of this permutation is to vary the distribution of strengths over all teams while still retaining conference-type scheduling. For instance, it is possible, under this scheme, that perhaps the best 4 teams in college football would be in the same conference. An advantage of this scheme over a completely random schedule is that it preserves the conference structure of college football that makes comparing teams from different leagues difficult. Two variants of this approach were used.
 - I. Teams were grouped into BCS conference and non-BCS conference teams, and strengths were permuted within groups.
 - II. The three teams with the largest estimated strengths (USC, Oklahoma and Auburn) and the three smallest (Ball St., Western Michigan and Central Florida) were exchanged. Subsequently, the resulting parameters were permuted within groups (as in I). This ensured that, on average, the three best teams were non-BCS teams and the three worst were BCS teams. (Note, however, that for any given vector of parameters, the three best teams need not be those with the largest estimated abilities after the 2004 season.)

It is not uncommon for two (or more) teams to tie for their regular season conference or, in the case of split-conferences, division title. In such cases, the following tie-breaking system was employed to determine which team would be named conference or division champion. This designation becomes important for determining participants in conference championship games as well as for automatic selection of teams in the BCS. Our sequential tie-breaking procedure is similar to the (different) procedures employed by many conferences.

1. Conference win/loss record.

2. Division win/loss record among tied teams (if applicable).
3. Head-to-head win/loss record among remaining tied teams.
4. Fewest overall losses among remaining tied teams. This was chosen, as most conferences (ACC, Big East, Big-XII, SEC) use the BCS rankings at some point in their tie-breaking procedures, and the BCS rankings are heavily dependent on polls, which favor teams with fewer losses.
5. Best computer ranking using a posterior estimate from the Bradley-Terry model and the simulated season's results.

4 Simulation Results

4.1 Comparison of playoff sizes

Four tournament sizes (one, two, four and eight teams) were investigated. The one-team tournament is trivial and crowns the national champion based on estimating the best team after the regular season. The two-team playoff is essentially what the NCAA currently uses in Division I-A football. The BCS is designed to match the two best teams *at the end of the regular season* in the BCS championship game. Many have suggested a “plus one” system, whereby the two best teams *after the bowl games* would play for the national championship. This scheme is analogous to a four-team playoff, and would presumably reconcile situations encountered in 2003 and 2004 where there was little separation between the top three teams. Though many pundits have touted this idea over the last two seasons, this system is far from perfect. In years where the national champion was undisputed and the only remaining unbeaten, such as Oklahoma in 2000 and Miami (FL) in 2001, requiring a unanimous top-ranked team to play one more game increases the chance of an upset and an inferior team being dubbed #1. Finally, an eight team playoff was examined. Carlin and Stern (1999) examine playoffs including up to 16 teams. Their results suggest that an eight-team playoff is nearly as likely to contain the best team as a 16-team playoff, and far more likely than a four-team one. Furthermore, given the reluctance of the NCAA to implement any playoff system, playoffs containing more than eight teams were considered impractical, and deemed beyond the scope of this study.

All tournament simulations assumed that post-season games were played on neutral sites as are the current bowl games. Five tournament fields were examined: a trivial, one-team playoff (Top 1), a two-team playoff (BCS 2), a four team-playoff (Top 4) and two eight-team playoffs (Top 8 and 6 Plus 2). The “Top m ” designation indicates that the m highest-ranking teams were selected for the draw (e.g., “Top 8” is an eight-team field consisting of the eight highest-ranked teams at the regular season’s end). The “6 Plus 2” field, which mirrors how the NCAA chooses participants for the four BCS bowls, consists of the six BCS conference champions and the two remaining highest ranking (at-large) teams. To emphasize the BCS’s current practice of selecting two teams for the title game, “BCS 2,” instead of “Top 2,” was used as an identifier. For a fixed tournament field, three methods for each playoff bracket were considered: fixed (F), Schwenk (S) and Hwang (H). For one- and two-team playoffs, there is no difference in the methods. For a four-team playoff, fixed and Hwang reseeding are identical. However, for the eight-team playoff, each yields a different result.

Each simulated season produces a different playoff field (even if the teams are the same, their true strengths differ across simulated seasons) and different seedings. Given the probability matrix, which is a function of θ , exact probabilities of various metrics given the tournament size and seedings were computed.

In subsequent tables, the best and worst playoff designs are in **bold** and *slanted* text, respectively. In addition, both are underlined for quick reference. Table 1 shows that the eight-team playoffs had the highest chance of including the true #1 team, which should not come as a surprise, with the “Top 8” tournament slightly outperforming the “6 Plus 2” configuration. The drop-off between the eight- and four-team playoffs, while substantial, is much less than the drops to two or one team. Results of the competitive simulation matched those of the standard simulation relatively, although the absolute performance of all methods was worse (as would be expected). Finally, the results of standard simulations matched the results of the regular simulation qualitatively, indicating that schedule likely does not influence the findings substantially.

Table 1: Probability of true #1 team being included in a playoff. Standard errors are given parenthetically.

	Standard	Competitive	Permuted I	Permuted II
Top 1	<u>0.363</u> (0.005)	<u>0.318</u> (0.005)	<u>0.314</u> (0.005)	<u>0.252</u> (0.004)
BCS 2	0.588 (0.005)	0.499 (0.005)	0.520 (0.005)	0.440 (0.005)
Top 4	0.809 (0.004)	0.704 (0.005)	0.769 (0.004)	0.702 (0.005)
Top 8	<u>0.954</u> (0.002)	<u>0.874</u> (0.003)	<u>0.940</u> (0.002)	<u>0.921</u> (0.003)
6 Plus 2	0.906 (0.003)	0.825 (0.004)	0.898 (0.003)	0.675 (0.005)

Table 2 shows that the eight-team playoffs also maximized the probability of the best team winning the championship; however, the difference between the eight- and four-team playoffs was slight. This can be explained by noting that it is generally more difficult for the best team (as well as for *any* team) to win a larger playoff than a smaller one. Although large playoff fields reduce the chance of the best team being left out, they require more consecutive wins once a team has qualified.

Table 2: Unconditional probability of true #1 winning the playoff. Standard errors are given parenthetically.

	Standard	Competitive	Permuted I	Permuted II
Top 1	<u>0.363</u> (0.005)	<u>0.318</u> (0.005)	<u>0.314</u> (0.005)	<u>0.252</u> (0.004)
BCS 2	0.491 (0.004)	0.377 (0.004)	0.438 (0.004)	0.376 (0.004)
Top 4(F)	0.569 (0.003)	0.410 (0.003)	0.544 (0.003)	0.505 (0.004)
Top 4(S)	0.568 (0.003)	0.409 (0.003)	0.543 (0.003)	0.504 (0.004)
Top 8(F)	0.594 (0.002)	0.416 (0.002)	0.586 (0.002)	0.579 (0.002)
Top 8(S)	0.593 (0.002)	0.413 (0.002)	0.585 (0.002)	0.578 (0.002)
Top 8(H)	<u>0.596</u> (0.002)	0.416 (0.002)	<u>0.587</u> (0.002)	<u>0.580</u> (0.002)
6 Plus 2(F)	0.591 (0.003)	0.414 (0.002)	0.581 (0.003)	0.482 (0.004)
6 Plus 2(S)	0.587 (0.003)	0.410 (0.002)	0.579 (0.003)	0.480 (0.004)
6 Plus 2(H)	0.593 (0.003)	<u>0.417</u> (0.002)	0.582 (0.003)	0.483 (0.004)

Since the current system is designed to create the best championship game, a number of “championship game metrics” were used to test viability of playoff designs. They were: the *worst* true rank of a team in the championship game, the *best* true rank of teams excluded from the championship game and the probability of the true #1 and #2 teams meeting in the championship game. Because a championship game requires two competitors, the one-team playoff was excluded from these comparisons. Table 3 gives the average true rank of the worst team to qualify for the title game (higher ranks are worse than lower ones). Regardless of simulation method, the two-team playoff is the poorest choice. For all metrics and simulations methods, all four- and eight-team playoffs produced more desirable results (smaller true rank of championship qualifiers,

larger true rank of teams excluded, and larger probability of #1 vs. #2) than did the two-team playoff. Therefore, if the NCAA strives for the “best” championship game, one could argue that *any* reasonable system is better than their current one.

Table 3: Average true rank of the worst team to qualify for the final. Standard errors are given parenthetically.

	Standard	Competitive	Permuted I	Permuted II
BCS 2	<u>6.386</u> (0.057)	<u>10.666</u> (0.115)	<u>6.747</u> (0.057)	<u>7.515</u> (0.059)
Top 4(F)	4.771 (0.028)	8.692 (0.063)	5.082 (0.030)	5.470 (0.031)
Top 4(S)	4.830 (0.024)	8.719 (0.056)	5.098 (0.026)	5.496 (0.027)
Top 8(F)	4.300 (0.018)	8.120 (0.04)	4.449 (0.019)	4.599 (0.019)
Top 8(S)	4.349 (0.014)	8.221 (0.035)	4.488 (0.014)	4.649 (0.015)
Top 8(H)	<u>4.189</u> (0.015)	<u>7.947</u> (0.037)	<u>4.354</u> (0.016)	<u>4.536</u> (0.017)
6 Plus 2(F)	4.538 (0.021)	9.011 (0.044)	4.715 (0.022)	5.880 (0.027)
6 Plus 2(S)	4.647 (0.017)	9.208 (0.039)	4.780 (0.018)	5.974 (0.023)
6 Plus 2(H)	4.362 (0.018)	8.556 (0.041)	4.540 (0.019)	5.628 (0.024)

4.2 Comparison of seeding methods

Although the theoretical properties of seeding methods require that the true ordering of teams be known, in reality, rankings must be estimated. Seeding will not necessarily achieve perfect agreement with underlying ability. Despite this complication, Table 2 shows that the Hwang method still maximizes the probability of the true #1 team (not necessarily the top seed) winning the tournament. Conversely, cohort seeding results in somewhat reduced probabilities of the top team winning. The standard fixed seeding produces results somewhere between these extremes. It is interesting to note that, empirically, even when true ranks are unknown (and seeding, therefore, is imperfect), the Hwang method tends to maximize favoritism, while the Schwenk method tends to minimize it. Finally, the choice of seeding methods had far less impact on the playoff performance than did the tournament size. This could be due, in part, because seeds are estimated (based on team performance in a short regular season) rather than given. Another potential explanation is that only reasonable seeding methods were tested. The probability of the top-seed winning can be drastically increased by use of unfair seeding, such as [[18][67]] [[23][45]], or drastically decreased, e.g., if seeding follows [[12][34]] [[56][78]].

5 Discussion

Much of the literature focusing on knockout tournaments focuses on properties of the tournament design given known probability matrices (see, e.g., David, 1963; Hartigan, 1968; Chung and Hwang, 1978). Searls (1963) and Appleton (1995) compare winning probabilities for various tournament schedules. Unlike their circumstance, we have focused on a more difficult situation in which the strengths of the participants are unknown. However, empirical evidence from simulation studies suggests that Hwang’s (1982) reseeding method serves to maximize favoritism *for the best team* (even if that team was not seeded first) while Schwenk’s (2000) serves to minimize it, with the standard fixed seeding method somewhere in between.

In all cases, the probability of correctly choosing a champion (or final two teams, as might be desired for television purposes) would be improved by enhancing the rating procedure before the playoff — certainly, a perfect rating system would remove all doubt as to which team is the best. With or without playoff, i.e., even with a selection committee approach, a good rating system would help tremendously. College basketball's reliance on the quirky Ratings Percentage Index (RPI) illustrates that the demand for objective, interpretable rating procedures sometimes exceeds the supply.

Our results indicate that the ability to identify correctly the best team over the course of the college football regular season leaves much to be desired, as evidenced by the low probability of a one-team playoff producing the correct champion. This is likely due to the large number of games played between teams of very disparate abilities, as such contests don't provide much additional insight into either team's ability. This view is furthered by noticing that, despite requiring the eventual champion to win three consecutive games against top competition, the eight-team tournaments showed the highest probability of correctly identifying the best team. A compromise between the two extremes is the "plus one" system advocated by many in the popular media. Such a system would approximate a four-team tournament, which yields a probability of correctly determining a champion almost equal to that of the eight-team fields, but requires substantially fewer changes to the *status quo*.

References

- Agresti, A. (1990). *Categorical Data Analysis* (First ed.). New York: John Wiley & Sons.
- Appleton, D. R. (1995). May the best man win? *The Statistician* 44(4), 529–538.
- Billingsley, R. (2004). Discussion – statistics and the college football championship. *The American Statistician* 58, 190.
- Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39(3/4), 324–345.
- Carlin, B. P. and H. S. Stern (1999). Designing a college football playoff system. *Chance* 12(3), 21–26.
- Chung, F. R. K. and F. K. Hwang (1978). Do stronger players win more knockout tournaments? *Journal of the American Statistical Association* 73(363), 593–596.
- Colley, W. (2004). Discussion – statistics and the college football championship. *The American Statistician* 58, 191–192.
- David, H. A. (1963). *The Method of Paired Comparisons*. New York: Hafner.
- Davidson, R. R. and D. L. Solomon (1973). A bayesian approach to paired comparison experimentation. *Biometrika* 60(3), 477–487.
- Gelman, A., B. P. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*. New York: Chapman and Hall.
- Hartigan, J. A. (1968). Inference from a knockout tournament. *The Annals of Mathematical Statistics* 39(2), 583–592.
- Harville, D. A. (2004). Discussion – statistics and the college football championship. *The American Statistician* 58, 187–189.
- Hwang, F. K. (1982). New concepts in seeding knockout tournaments. *American Mathematical Monthly* 89(4), 235–239.
- Massey, K. (2004). Discussion – statistics and the college football championship. *The American Statistician* 58, 185–187.

- Mease, D. (2004). Discussion – statistics and the college football championship. *The American Statistician* 58, 192–194.
- Schwenk, A. J. (2000). What is the correct way to seed a knockout tournament? *The American Mathematical Monthly* 107(2), 140–150.
- Searls, D. T. (1963). On the probability of winning with different tournament procedures. *Journal of the American Statistical Association* 58(304), 1064–1081.
- Stern, H. S. (2004). Statistics and the college football championship. *The American Statistician* 58, 179–185.
- Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* 81(393), 82–86.