

MARCH MADNESS

An Analysis of a Nonstandard Basketball Pool

The National Collegiate Athletic Association (NCAA) holds tournaments and competitions to determine national championships in college and university sports. Probably the most popular of these is the Division I men's basketball tournament, known to sports fans as "March Madness." In this highly publicized competition, 64 schools are selected to compete in a single elimination tournament. The schools are assigned a seed from 1 to 16, and placed into one of four regions with the lower-numbered seeds representing the teams thought to have the best chance of winning. Each region produces one winner who proceeds to the Final Four where a national champion is crowned. The starting field for the 2000 tournament is shown in the background on the next page. A glance at its structure suggests that the design and outcome of such tournaments are rich in mathematical and statistical content. We consider some of that content in light of a prediction contest, or pool.

AARON ARCHER is a graduate student in the School of Operations Research and Industrial Engineering at Cornell University in Ithaca, NY, where **RICK CLEARY** is Associate Dean for Undergraduate Programs. **ROBIN LOCK** is Professor of Mathematics at Saint Lawrence University in Canton, NY. **JOHN TRONO** is Associate Professor of Computer Science at Saint Michael's College in Colchester, VT.

Everybody In The Pool!

One of the reasons that March Madness is so popular is that people like to compete to see how well they can predict the outcome of the tournament. Many groups of friends and co-workers form pools where the object is to submit the best set of predictions for the tournament. A typical pool requires the entrant to fill in a winner in each game in each bracket, filling in the 63 blank lines shown in the background on the next page. A specified number of points is awarded for each correct prediction, usually with later rounds of the tournament weighted more heavily. The winner of the pool is the person with the most points at the end of the tournament. Such pools, and the tournament structure that underlies them, have been studied quite extensively by mathematicians and statisticians.

Filling in every possible game throughout the championship might discourage people from entering a pool if they are only casual fans. It certainly forces people to make decisions about games in which they might know very little about one or both teams. One could overcome this simply by taking the lower-seeded team in each game, but this would eliminate much of the fun of playing.

A pool could, naturally, be run just for fun or bragging rights. The authors of this article have come to understand that some people actually gamble on their ability to predict these results, each paying a required stake to enter

the pool with the winner taking the monetary prize. While the authors do not encourage readers to violate any local statutes on gambling, we will as a pedagogical tool use the idea that any pool we discuss carries with it an entry fee of one dollar.

An Alternate Way To Play

To encourage participation from people intimidated by the traditional "whole bracket" pool Robin Lock created an alternate pool that allows participants to pick only a few teams. This is done by establishing a price for teams at each seed level. The prices for the 2000 pool as set by Rick Cleary appear below as Figure 1. Players in this pool buy \$1 worth of teams. They receive a point each time one of the teams they select wins a game. The winner of the pool is the person who gets the most points.

Seed	Cost
1	.25
2	.21
3	.18
4	.15
5	.12
6	.10
7	.08
8	.06
9	.05
10	.04
11	.03
12	.02
13-16	.01

Figure 1. The prices for the 2000 pool

Like the traditional pool, this one raises a number of interesting questions about strategy and structure. This paper addresses two of the most basic, using some simple mathematical and statistical tools. First, can we find a strategy based on historical data that increases our chances of winning the pool? Second, looking back after the tournament, what would have been the optimal strategy? We will address these questions one at a time. There are many other directions of inquiry left to explore. We list some of these at the end of the article.

Strategy

Using the established prices shown in Figure 1, participants wish to choose a set of teams that gives them the best chance of winning the pool. Establishing the prices based on the tournament structure is in itself an interesting question that we leave for future study. To choose a strategy we employ the results of the tournaments played to date. It is doubtful if any *a priori* strategy could win this pool every year. As one might guess, there are an enormous number of different ways to spend the \$1 in this pool. Without picking any of the “penny” teams (seeds 13 through 16), there are over 3 billion distinct entries possible.

Many different rating schemes are available to rank the relative strengths of the teams in the tournament. Serious fans of college basketball like to use these, or develop their own rankings, to choose teams that might exceed expectations. The tournament almost always produces one or two teams who score many wins for a bargain price. For example Tulsa, a #7 seed in the 2000 NCAA tourney, was a good buy, collecting 3 wins for just eight cents. Because we want to keep the focus on mathematics and statistics, and not on basketball, the analysis that follows does not take into consideration the strengths and weaknesses of individual teams but relies solely upon the data collected for the last sixteen tournaments.

Beginning with the tournament played in 1985, 64 teams have been invited to compete for the NCAA championship each year. Tables I–III

Table I. Average Wins Per Year and Cost (Cents/Win) for Each Seed

Seed	Wins	Cost/Win
1	13.44	7.4
2	9.69	8.7
3	6.63	10.9
4	6.44	9.3
5	4.56	10.5
6	5.44	7.3
7	3.19	10.0
8	2.94	8.2
9	2.31	8.7
10	2.69	6.0
11	1.88	6.4
12	1.75	4.6
13	0.94	4.3
14	0.94	4.3
15	0.19	21.0
16	0.00	—

summarize the results of the sixteen tournaments played from 1985–2000. For each seed level Table I gives the average number of wins and cost per win for the four teams at that seed level. For example, the first line tells us that the four number 1 seeds have averaged a total of 13.44 wins per year. Since it would cost a dollar to buy all four, those wins cost 7.4 cents each. Table II shows the number of wins earned by individual teams at each seed level for the last 16 years. Since it takes six wins to become the national champion the last column is a summary of where the eventual winners were seeded. Note that nine of the sixteen champions began as top seeds in their region. Also note that the 16th-seeded teams have yet to win a game!

With the information in Tables I and II, plus a little historical data about our pool (it usually takes about 20 wins to be the champion) we formulate the following strategy: Take all teams seeded 6th (40 cents), 10th (16 cents) and 12th–14th (16 cents). By picking your favorite, or by choosing at random, take a single #1 seed. (If you are feeling aggressive, enter the pool four times and take each #1 seed on one of your entries!) This top-seeded team costs 25 cents. We have three cents left to buy a single 11th seed or three 15th seeded teams. We justify this strategy in the following paragraph using the

data in Table III. Table III shows the number of games won using this strategy each year. It gives the number of games won by each seed chosen by our strategy. There is a range of values for each year depending on which number 1 seed was chosen. The column “won” shows the winning score for the years 1992–2000 when this pool was run with about 100 entries per year. The “optimal” column shows the best possible total that could have been achieved once results are known, described later in this article.

Using the average cost per win from Table I, the #12–#14 seeds appear to be the best buys. As listed in Table III, those teams have won between 3 and 6 games each year, costing a total of only 16 cents for 12 teams...until 2000 when they did not register a single win! Since we are choosing teams based on their seed rather than their individual merit, we will take them all and count on their continued success until we can decide if this year’s results were a fluke. We also have historical data that supports our strategy of taking all the 12–14 seeds. From Table III we see that the winning total is usually very close to 20, so spending 5 cents per win is a good goal. The 12th–14th seeds are the only teams that have on average earned wins for less than five cents each so they seem the right place to start.

TABLE II. Distribution of the number of wins, for each seed level, for the 64 teams at that level over the past 16 years

Seed #	Wins						
	0	1	2	3	4	5	6
1	0	9	11	16	13	6	9
2	3	19	12	17	7	3	3
3	13	23	14	7	2	4	1
4	12	22	19	4	5	1	1
5	18	25	18	1	1	1	0
6	19	19	16	7	1	1	1
7	26	28	7	3	0	0	0
8	35	22	2	2	2	0	1
9	29	33	1	1	0	0	0
10	38	14	7	5	0	0	0
11	45	11	6	1	1	0	0
12	46	8	10	0	0	0	0
13	52	9	3	0	0	0	0
14	51	11	2	0	0	0	0
15	61	3	0	0	0	0	0
16	64	0	0	0	0	0	0

TABLE III. Strategy results

	14	13	12	10	6	1's	Total	Won	Optimal
1985	0	1	2	0	1	1,3,4,5	5,7,8,9	–	30
1986	3	0	2	1	5	1,3,4,5	12,14,15,16	–	34
1987	1	2	2	4	9	3,3,4,6	21,21,22,24	–	30
1988	1	2	0	1	10	2,3,4,5	16,17,18,19	–	29
1989	1	1	1	1	0	2,2,3,4	6,6,7,8	–	29
1990	1	0	3	3	6	1,2,3,6	14,15,16,19	–	33
1991	1	1	2	4	2	2,3,4,4	12,13,14,14	–	30
1992	1	1	2	2	10	1,3,3,6	17,19,19,22	20	30
1993	0	1	2	0	5	3,4,5,6	11,12,13,14	18	26
1994	0	0	3	3	4	1,3,3,6	11,13,13,16	20	28
1995	2	1	1	1	6	2,2,3,6	13,13,14,17	–	26
1996	0	1	3	2	5	1,2,4,6	12,13,15,17	21	29
1997	2	0	1	5	8	2,4,4,5	18,20,20,21	20	30
1998	1	2	1	4	3	1,3,3,4	12,14,14,15	19	31
1999	1	2	3	8	7	2,4,5,6	23,25,26,27	25	32
2000	0	0	0	4	7	1,1,2,6	12,12,13,17	22	34

Notes:

1. These are the results for spending only 97 cents. The remaining three cents could have resulted in some years in additional points for wins by #11 or #15 seeds.
2. Columns under seed totals are wins by all four teams at that seed level. There are four values in the top-seed column corresponding to the four choices for our top seed.
3. The numbers in the “won” column represent the best entry in a pool of approximately 100 players.
4. The numbers in the “optimal” column represent the best total possible once the tournament results were known.

CHAMPION We have 84 cents left to spend. What else can we learn from the historical data? In eleven of the sixteen tournaments for which we have computed optimal totals, a single top seed is present in the optimal answer. There are four times when two such teams appear, and only once does no top seed show up in the best solution. Spending a quarter on a top seed can be a pretty good purchase. As noted, nine of the past sixteen champs held a number 1 seed, and another five have reached the final game before losing. A team that costs a quarter and gets five or six wins fits our “five cents/win” criterion. The persistent ability of the top seeds to reach the final gives a player the chance to get many wins with a single team.

It is important to keep in mind that to win this pool one must have the best entry out of many. A system of selecting teams that does well on average is not good enough. We need a system that gives a non-trivial chance of getting the highest score. After including our single top seed (25 cents) and all

of the 10th seeds (the next best position on average—16 cents) we would, if strictly following Table I, choose the 11th seeds. However the 6th seeds have also done well historically, and with a much higher variance than the 11th seeds. It has been conjectured that the 6th seeds have a higher likelihood of reaching the final eight than the 4th or 5th seeds. (See the article by Scott Berry cited in the “For Further Reading” section.) For these reasons, we choose to spend the 40 cents needed to buy all of the 6th seeds. We are now at 97 cents spent. The remaining three cents can be spent on either a single 11th seed or three 15th seeds.

As can be seen from the bold numbers in Table III, with the right choice of top seed this strategy could have won or tied in our pool in three of the nine years in which it has been played. It might well have won in 1987 when our pool wasn’t established yet. While any system chosen by examining historical trends will naturally tend to perform less well in the future than it has in the past, this strategy has the strength that

knowledge of the teams is not necessary to play.

A Knapsack Problem: The Tournament in Retrospect

Determining the best strategy for selecting teams before the tournament is an open-ended, highly charged sports question that lends itself to empirical solutions based on historical data as presented in our strategy section. The mathematical underpinnings are murky, and the best strategy will evolve over time as more tournaments are played and more data becomes available. But there is a related problem that has a simple, elegant solution. After the tournament is over, how do we calculate what the best selection of teams *would have been*?

This question is an example of the knapsack problem, a famous and much-studied problem in computer science and operations research. Here is the standard formulation. You are given a list of n items. Each item i has a weight w_i and a value v_i . You have a knapsack that can carry any combination of items so long as their total weight does not exceed the capacity C . The capacity, values and weights are all positive integers. Your challenge is to select a subset S of the items to maximize the total value of the items selected without exceeding the capacity of the knapsack. That is, we want to maximize the sum of the v_i 's subject to the constraint that the sum of the w_i 's is less than or equal to C , where the sums are taken over the elements in the subset S .

For the NCAA tournament pool, the items are teams, the value v_i is the number of games team i won in the tournament, the weight w_i is the number of cents team i cost to buy. The knapsack capacity is $C = 100$, since we are allowed to spend up to a dollar. But how do we find the subset of teams that gives us the greatest total of wins?

From one point of view, this problem is trivial. There are only $n = 32$ teams that won at least one game. We could simply enumerate every possible subset, check to see which ones are *feasible* (satis-

Table IV. A piece of the dynamic programming table

region	team	seed	cost	wins	cents/win	0	1	2	3	4	5...	...32	33	34	35
WEST	Arizona	1	25	1	25	0	25*	—	—	—	—	—	—	—	—
	Wisconsin	8	6	4	1.5	0	25	—	—	6*	31*	—	—	—	—
	Texas	5	12	1	12	0	12*	37*	—	6	18*	—	—	—	—
	LSU	4	15	2	7.5	0	12	15*	27*	6	18	—	—	—	—
	Purdue	6	10	3	3.3	0	12	15	10*	6	18	—	—	—	—
	Oklahoma	3	18	1	18	0	12	15	10	6	18	—	—	—	—
	Gonzaga	10	4	2	2	0	12	4*	10	6	14*	—	—	—	—
	St. John's	2	21	1	21	0	12	4	10	6	14	—	—	—	—
MIDWEST	Michigan St.	1	25	6	4.2	0	12	4	10	6	14	—	—	—	—
	Utah	8	6	1	6	0	6*	4	10	6	12*	—	—	—	—
	Kentucky	5	12	1	12	0	6	4	10	6	12	—	—	—	—
	Syracuse	4	15	2	7.5	0	6	4	10	6	12	—	—	—	—
	UCLA	6	10	2	5	0	6	4	10	6	12	—	—	—	—
	Maryland	3	18	1	18	0	6	4	10	6	12	—	—	—	—
	Auburn	7	8	1	8	0	6	4	10	6	12	—	—	—	—
	Iowa St.	2	21	3	7	0	6	4	10	6	12	226*	—	—	—
EAST	Duke	1	25	2	12.5	0	6	4	10	6	12	205*	226*	251*	—
	Kansas	8	6	1	6	0	6	4	10	6	12	193*	211*	232*	257*
	Florida	5	12	5	2.4	0	6	4	10	6	12	138*	150*	162*	175*
	Illinois	4	15	1	15	0	6	4	10	6	12	138	150	162	175
	Pepperdine	1	3	1	3	0	3*	4	7*	6	9*	133*	141*	153*	165*
	Oklahoma St.	3	18	3	6	0	3	4	7	6	9	129*	136*	144*	151*
	Seton Hall	10	4	2	2	0	3	4	7	6	9	119*	125*	133*	140*
	Temple	2	21	1	21	0	3	4	7	6	9	119	125	133	140
SOUTH	Stanford	1	25	1	25	0	3	4	7	6	9	119	125	133	140
	NorthCarolina	8	6	4	1.5	0	3	4	7	6	9	98*	104*	110*	118*
	Connecticut	5	12	1	12	0	3	4	7	6	9	98	104	110	118
	Tennessee	4	15	2	7.5	0	3	4	7	6	9	98	104	110	118
	Miami (Fla)	6	10	2	5	0	3	4	7	6	9	96*	102*	108*	114*
	Ohio St.	3	18	1	18	0	3	4	7	6	9	96	102	108	114
	Tulsa	7	8	3	2.7	0	3	4	7	6	9	88*	94*	98*	104*
	Cincinnati	2	21	1	21	0	3	4	7	6	9	88	94	98	104

fy the capacity constraint), then take the one with the highest total value. But there are 2^{32} (about 4.3×10^9) subsets to check, so this method is computationally impractical.

Fortunately, there is a more clever approach using a technique called *dynamic programming*. Define $g(i, v)$ to be the minimum cost of a subset of the first i teams in our table (the order in which we put the teams in the table doesn't matter) that has value exactly v (taken to be infinity if no such subset exists). The boundary values of g (that is, when $i = 1$ or $v = 0$) are easy to compute. Take a look at Table IV, a piece of our dynamic programming table, to see the following rules in action.

1. $g(i, 0) = 0$ for each i since every team in our table won at least one game.

2. $g(1, v) = 25$ for $v = 1$, since Arizona, the first team in our table, won exactly one game and cost 25 cents. The rest of the values on the line, shown by dashes, should be considered infinitely large. (There is no amount of money that would get us two or more wins when Arizona is the only team available.) Note that if Wisconsin was the first team in the table, we'd give $g(1, 4) = 6$ with dashes everywhere else on the first row, since Wisconsin won exactly four games and cost six cents.

What is not immediately obvious is that each row can be easily computed from the row above it and, at the end, the computation can be unraveled to find the optimal subset. Now that we have filled in the first row of the table,

consider this recursion: For $i > 1$, $g(i, v) = \min(g(i-1, v), g(i-1, v-w_i) + w_i)$.

This statement compares the cheapest way to get exactly v wins with the first $i-1$ teams ($g(i-1, v)$) to the cheapest way to get exactly v wins that includes team i . Look in Table IV at the entry $g(5, 3) = 10$. The recursion above becomes

$$g(5, 3) = \min(g(4, 3), g(4, 0) + 10) = \min(27, 10) = 10.$$

In words, this value says that the cheapest way to get exactly three wins using the first five teams is to just take Purdue, at a cost of 10 cents. Using only the first four teams, we had to pay 27 cents to get three wins, 15 cents for two wins from LSU and 12 cents for a win by Texas. We put an asterisk next to an

entry in the table whenever the team on that row is included among the teams that make up that total. Retracing these asterisks through the table from the optimal total allows us to name the teams that make up the best entry.

There are only 63 games in the tournament, since each of the 64 teams except the champion loses exactly once. Thus our table will have 64 columns and we are interested in values in the table that are less than or equal to 100, since we have 100 cents to spend. Thus the optimal total will be the number above the rightmost column that contains a value of 100 or less on the last row. For the 2000 tournament, that value turned out to be 34 wins for 98 cents (see Table IV). Note that there is no asterisk at $g(32, 34) = 98$ because Cincinnati, the team listed in the last row, is not included in the best entry. While space considerations prevent us from showing the entire 2000 table, we will for the record point out that the best entry for 2000 included Tulsa, Miami (Florida), North Carolina, Seton Hall, Pepperdine, Florida, UCLA, Michigan State, Gonzaga, Purdue and Wisconsin.

Want to work on this problem?

If you're looking for an undergraduate student project consider these questions. We've left a few items that any reader could work on! Here are some we encourage readers to try, and feel free to contact us for more information.

1. How should we set prices for the pool? Fixing the prices for the top seeds at \$.25 per team provides a simple way to enter for the conservative participant: buy the four #1 seeds, and hope that there are few upsets. Is pricing the 16th-seeded teams at a penny apiece another natural choice? How should we choose the prices in the middle?
2. We propose a strategy based on empirical data from recent years. Can you propose a different strategy that beats ours regularly?
3. How would your strategy change if you were allowed to buy the same team multiple times? For instance, suppose your favorite team enters the tournament as a sixth seed and you were allowed to spend your dollar by buying that team 10 times.

4. How would pricing and strategy change under other reward systems? For instance, suppose that every team costs 25 cents so you could only buy four teams, but upset wins were rewarded with more points?
5. How about a simulation? You could use the tables we provide to create a simulation for this pool, or the usual "fill in the whole bracket" pool. See if our strategy does as well as one that you might suggest. ■

For Further Reading

- Berry, S. M., "A Statistician Reads the Sports Pages—My Triple Crown," *Chance*, (2000) 13:3.
- Breiter, D. J. and Carlin, B. P., "How to Play the Office Pools if You Must," (1997) *Chance*, 10:1.
- Schwenk, A. J., "What is the Correct Way to Seed a Knockout Tournament," *The American Mathematical Monthly*, (2000) 107:2.
- Stern, H.S. and Mock, B., "College Basketball Upsets: Will a 16-Seed Ever Beat a 1-Seed?," *Chance*, (1998) 11:1.



Oklahoma State University Assistantships and Fellowships



The Department of Mathematics at Oklahoma State University invites qualified applicants to join its Graduate Program. We have highly regarded research groups in Algebraic Geometry, Analysis, Mathematics Education, Number Theory, Representation Theory and Topology and are widely recognized for our supportive environment for graduate students. The Department has 34 faculty, on average 10 postdoctoral fellows, and 45 graduate students and is located in the pleasant, safe and affordable city of Stillwater.

Programs: MS in Pure and Applied Mathematics.
PhD in Mathematics and Mathematics Education.

Financial Aid: Teaching Assistantships, tuition waivers and fellowships, which increase stipends for outstanding applicants. Advanced students may obtain Research Assistantships.

We have many weekly seminars, an outstanding visitor and colloquium program and offer early exposure to research, internships and interdisciplinary experiences. The Masters in Applied Mathematics is an excellent program for those seeking careers in business or industry. The PhD with Specialization in Mathematics Education is an excellent program for preparation of college teachers. Application forms and more information may be obtained from the web pages or by writing to:

Director of Graduate Studies
Oklahoma State University
Department of Mathematics
Stillwater, OK 74078-1058

Web Page: <http://www.math.okstate.edu/grad/maa.html>
E-mail: graddir@math.okstate.edu
Telephone: (405) 744-5688
FAX: (405) 744-8275