Horseshoe pitchers' hot hands

GARY SMITH

Pomona College, Claremont, California

Gilovich, Vallone, and Tversky's (1985) analysis of basketball data indicates that a player's chances of making a shot are not affected by the results of earlier shots. However, their basketball data do not control for several confounding influences. An analysis of horseshoe pitching, which does not have these defects, indicates that players do have modest hot and cold spells.

Gilovich, Vallone, and Tversky (1985) present evidence against the popular belief that basketball players get hot hands. Their survey of 100 basketball fans showed an overwhelming belief that a player has a better chance of making a basket after having made shots than after having missed shots. Five of seven Philadelphia 76ers basketball players also believed this.

However, their analysis of the performance of individual Philadelphia 76ers during the 1980–1981 season found that the probability of making a shot was usually somewhat *lower* after having made shots than after having missed shots, although the observed differences were not statistically persuasive. Similarly, an analysis of the number of runs of hits and misses by individual players across all games and within individual games generally showed slightly more runs than would be expected if shots were independent (the opposite of the hot hands phenomenon), although again the results were seldom statistically persuasive.

One weakness of their analysis is that it ignores the time interval between shots. A player's two successive shots might be taken several minutes apart, before and after the halftime intermission, or even in different games. The results of such widely separated shots is not what fans mean by a hot hand. Another problem is that a player's shot selection may be affected by his recent hits or misses. A player who has made several shots in a row may be tempted to try more difficult shots, thereby reducing his probability of success. A player who has been missing shots may pass up attempts he would normally take and shoot easier shots. Also, the score or other game considerations may affect team strategies. For example, the opposing team may guard hot players more closely and cold players more loosely. A team that is behind may take more low-percentage shots by shooting more quickly and attempting more 3-point shots. A person may play less aggressively if he is in foul trouble and more aggressively if his defender is in foul trouble.

The average starter takes from 10 to 20 shots a game, 5–10 in each half. If the hot hand is a relatively modest phenomenon, a statistical test based on 5–10 shots will have very little power and may be hidden in data sets that combine shots of varying difficulty separated by long periods of time. When data are aggregated from different games, the confounding influences include the home court advantage, the number of days between games, and the amount of travel to get to a game.

Gilovich et al. (1985) also examined free-throw accuracy for nine Boston Celtics players during the 1980–1981 and 1981–1982 seasons. When shooting pairs of free throws, four players were more likely to make the second shot after making the first shot, and five were more likely to make the second shot after missing the first shot; none was statistically significant. This is arguably their most persuasive evidence; however, the aggregation of infrequent shots taken in different games does not really encompass the usual notion of hot streaks.

Finally, they conducted a controlled experiment in which 26 Cornell players took 100 shots, moving along arcs drawn from the basket. Overall, the players were slightly more likely to make a basket after having made 1–3 shots, but the differences were statistically persuasive for only 1 player.

Gilden and Wilson (1995) presented some experimental evidence of nonconstant success probabilities in golf putting and dart throwing. However, their results might be criticized for their artificiality, since their experiments involved volunteers making 300 repetitions and being paid \$5 plus (in three of the four studies) five cents for each hit. With so many trials and such small stakes, it is conceivable that there are substantial fluctuations in the attentiveness of poorly motivated volunteers. If success rates are higher when they are focused on their assignment and lower when they are bored, hits and misses will cluster in the data and will exhibit the statistical patterns associated with streakiness—for example, higher hit rates after hits than after misses and fewer runs of hits and misses than would be expected with a constant hit rate.

The hot hands question is whether highly skilled and motivated athletes sometimes encounter hot and cold

Correspondence concerning this article should be addressed to G. Smith, Department of Economics, Pomona College, Claremont, CA 91711 (e-mail: gsmith@pomona.edu).

spells that cannot be easily explained as chance fluctuations about a constant success rate. Serious athletes competing for meaningful stakes are likely to remain focused on their task and to provide meaningful data for testing whether hot hands are real or an illusion.

HORSESHOE PITCHING DATA

One sport that has very few confounding influences is horseshoe pitching. The rules described here have been used at recent world championships and represent the most commonly followed norms. A horseshoe court has two stakes, placed 40 ft. apart, with each stake centered in a 6-foot-square pitcher's box that contains pitching platforms on each side of a 3-foot-wide pit. The players can pitch their shoes from the front of either pitching platform, reducing the distance to the opposite stake to 37 ft. Extended platforms are used in women's competition to reduce the pitching distance to 27 ft.

In each inning, a player pitches two shoes, and then the other player pitches two shoes, with the score tallied after all four shoes have been pitched. A shoe that encircles the stake is a ringer, worth 3 points; a nonringer that is within 6 in. of the stake is a *shoe in count*, worth 1 point. In conventional cancellation scoring, only one contestant can score in each inning. Ringers thrown by both players cancel each other (*dead ringers*), as do shoes in count that are equidistant from the stake. For instance, if both players throw double ringers, these are all dead ringers, and no points are scored. If one player throws double ringers and the opponent throws one ringer and a shoe in count, the live ringer scores 3 points.

A shoe is flipped like a coin toss to determine who pitches first in the first inning. Thereafter, the player who scores pitches first in the next inning. If neither player scores, the player who pitched second pitches first in the next inning. The first player to score 40 or more points is the winner. At the World Championships, 16 players qualify for the men's and women's championship matches, and each player pitches against each of the 15 other players. The final standings are determined by the players' overall won-lost records.

In top competition, players typically throw 60%–80% ringers and games last 20–30 innings. One of the greatest games of all time occurred at the 1965 World Championships, when Ray Martin threw 89.7% ringers and lost a 2.5-h, 97-inning marathon to Glen "Red" Henton, who pitched 90.2% ringers.

Walter Ray Williams, Jr., six times world champion, generously provided detailed data from the 2000 and 2001 World Championships for men and women (Williams, 2002). These score sheets record each player's live and dead ringers, shoes in count, and misses in each inning but do not distinguish between a player's first and second pitches.

ANALYSIS

In comparison with basketball data, horseshoe pitching has many valuable properties for testing streakiness in performance. There are no confounding influences from team play, defenses, or strategy based on the score. World class pitchers always try to throw ringers, and every pitch is from the same distance and separated by only brief intervals of time.

Conditional Probabilities

The key issue is whether the number of ringers a player pitches in an inning is independent of the number of ringers in the previous inning. Each game was analyzed separately in that I did not consider whether the first inning of a game was influenced by the last inning of the previous game. Because double misses are unusual for world class pitchers, each player's inning was characterized as a double ringer or not a double ringer. The top male and female players throw doubles roughly half the time, making a nice analogy to coin flips.

Table 1 shows the overall frequencies with which the players pitched doubles after doubles or nondoubles in the preceding one or two innings. These overall frequencies might have been affected by the fact that the best pitchers were more likely to be pitching after throwing doubles. This effect should be small, since these were all championship caliber pitchers; nonetheless, Table 1 also shows simple unweighted averages of the individual player frequencies.

Men and women were both somewhat more likely to throw a double after a double than after a nondouble and were also more likely to throw a double after two doubles in the preceding two innings than after two nondoubles. (These hot hand patterns imply analogous cold hands; if a hit is more likely after a hit than after a miss, then a miss is more likely after a miss than after a hit.)

When Fisher's exact test is used, of those pitchers who were more likely to throw a double after a double, 8 had one-sided p values less than .025; of those pitchers who

Table 1
Frequency of Doubles Following One or Two Doubles or Nondoubles, With Unweighted Averages

Players	Two Nondoubles		One Nondouble		One Double		Two Doubles	
	Freq.	Aver.	Freq.	Aver.	Freq.	Aver.	Freq.	Aver.
Men, 2000	.467	.477	.475	.480	.531	.514	.550	.521
Women, 2000	.451	.478	.492	.501	.562	.544	.570	.544
Men, 2001	.469	.502	.489	.505	.587	.561	.604	.552
Women, 2001	.479	.504	.509	.516	.593	.573	.599	.564
Total	.466	.490	.491	.501	.569	.548	.582	.545

were more likely to throw a double after two doubles, 8 had one-sided p values less than .025. (None of the pitchers who were less likely to throw a double after a double or after two doubles had p values below .025). If each of the 64 pitchers had a .025 probability of a p value below .025, the binomial distribution shows that the probability that as many as 8 would have p values below .025 is .0002.

Another way to look at the data is that 25 of the 32 male pitchers and 26 of the 32 females were more likely to throw a double after a double than after a nondouble. If each pitcher had a .5 probability of being more successful after a double, the binomial distribution shows that there is only a .0000009 probability that as many as 51 of 64 players would be more successful after a double. In addition, 23 of the 32 men and 25 of 31 women were more likely to throw a double after two doubles than after two nondoubles (1 woman was equally likely in 2000). If each pitcher had a .5 probability of being more successful after two doubles, there is only a .00002 probability that as many as 48 of 63 players would be more successful after two doubles.

Runs

Independence can also be tested by the length of the longest run and by the number of runs in a game. Bate-man (1948) showed that, in the case of first-order dependence, a test based on the number of runs in a sequence is more powerful than a test based on the length of the longest run.

After each player's throws in each inning are categorized as either a double or a nondouble, the exact probability that the number of runs would be as small as actually observed (indicating the presence of hot and cold streaks) can be calculated (Stevens, 1939). However, in addition to the low power with small sample sizes, the possible p values are not continuous. Consider, for example, a game with 10 double and 10 nondouble innings. If double and nondouble innings are independent, the probability of six or fewer runs is .0185, and the probability of seven or fewer runs is .0513. The probability of a p value less than .025 is .0185, not .025.

Another complication is that the number of doubles and nondoubles may be such that the number of runs cannot possibly be statistically persuasive. For example, if a player throws all doubles, this seems to be evidence of a hot streak, but the statistical fact that such data will always yield exactly one run means that a runs test, which looks for the clumping of doubles and nondoubles, cannot provide statistically persuasive evidence against the null hypothesis that doubles and nondoubles are independent.

There were no perfect games at these World Championships, but there were several close enough to make a runs test useless. In one game in 2000, Alan Francis pitched 13 doubles in 14 innings. Under the null hypothesis, there is a $2/14 = .143$ probability of two runs and a $12/14 = .857$ probability of three runs. There is no chance of rejecting the null hypothesis at the 5% level.

One way to avoid these problems was to calculate the expected value of the number of runs in each game under the null hypothesis of independence and then see whether the actual number of runs was higher or lower than this expected value. I then tabulated the total number of games in which the actual number of runs was above and below the expected value and used the binomial distribution to test the null hypothesis that the actual number of runs was equally likely to be above or below its expected value.

Table 2 summarizes the results. Too few runs (evidence of streakiness in performance) were consistently more likely than too many runs. Overall, there is only a .000009 probability of such an imbalance between the number of games with fewer runs than expected and the number of games with more runs than expected.

Ringer Percentages by Game

Another kind of evidence of streakiness would be if a pitcher's performance fluctuated from game to game more than would be expected if horseshoe pitching were a Bernoulli process with constant success probability. Table 3 shows an example. Mary Ann Peninger threw 74.4% ringers in her 15 games at the 2000 championships. If her chances of throwing a ringer were the same in each game, the expected value of the number of ringers in each game would be .744 multiplied by the number of pitches in that game. In practice, her actual ringer frequencies in these 15 games varied from .574 to .827. The chi-square statistic gauges whether the variations between the actual and the expected values shown in Table 3 are improbably large. For these particular data, the p value is .139, not low enough to reject the null hypothesis at the 5% level.

Similar calculations for all the pitchers showed that 5 of 32 men and 8 of 32 women had p values less than .05. If the null hypothesis were true, so that each pitcher has a .05 probability of a p value less than .05, the probability that as many as 13 of 64 pitchers (20.3%) would have p values less than .05 is .00001.

FIRST AND SECOND PITCHERS

Although horseshoe data are in many ways ideal, one possible confounding influence is that there may be a physical or psychological difference between pitching first and second. For example, shoes may be more likely to bounce off the stake if they land on another shoe than if they land in a bare pit. The high ringer percentages in

Table 2
Number of Games with Fewer or More Runs Than Would Be Expected With Independence

Players	Fewer Runs	More Runs	p Value
Men, 2000	129	107	.0857
Women, 2000	136	103	.0191
Men, 2001	137	98	.0065
Women, 2001	138	99	.0067
Total	540	407	.000009

Table 3
Mary Ann Peninger's Performance by Game in 2000,
With Expected Values (Exp.)

Game	Accuracy	Ringers		Nonringers		Total
		Performance	Exp.	Performance	Exp.	
1	.673	35	38.7	17	13.3	52
2	.574	31	40.2	23	13.8	54
3	.733	66	67.0	24	23.0	90
4	.771	37	35.7	11	12.3	48
5	.821	64	58.1	14	19.9	78
6	.750	54	53.6	18	18.4	72
7	.775	79	75.9	23	26.1	102
8	.763	58	56.6	18	19.4	76
9	.692	36	38.7	16	13.3	52
10	.683	41	40.2	19	15.3	60
11	.827	43	38.7	9	13.3	52
12	.704	38	40.2	16	13.8	54
13	.827	43	38.7	9	13.3	52
14	.720	36	37.2	14	12.8	50
15	.784	58	55.1	16	18.9	74
Total	.744	719		147		966

championship tournaments indicate that this does not happen frequently. Still, it might give a slight advantage to the player pitching first.

Since the player who scores in any inning pitches first in the next inning, it is conceivable that the hot and cold streaks documented above simply reflect the advantages of pitching first. On the other hand, the pitching order rotates when neither player scores, which happened in 29% of the innings.

Conditional Probabilities

One way to control for pitching order is to separate the conditional probabilities for those pitching first from the conditional probabilities for those pitching second. Unfortunately, the reduced sample sizes make it more difficult to reject the null hypothesis if the differences in conditional probabilities are relatively small. On average, individual players pitched first in roughly 160 innings after throwing a double in the preceding inning and in roughly 50 innings after throwing a nondouble in the preceding inning, with the numbers reversed for players pitchers second. With these sample sizes, Figure 1 shows, for various values of the differences in ringer probabilities, $p_1 - p_2$, the probability of a sufficiently large observed difference in ringer frequencies to reject the null hypothesis $p_1 = p_2$. This power function shows that for modest differences in ringer probability, the sample sizes are too small to reject the null hypothesis consistently. For example, if the probability of a double after a double is $p_1 = .55$ and the probability of a double after a nondouble is $p_2 = .45$, there is only a .234 probability that the observed performance differences will be sufficiently large to reject the null hypothesis. If $p_1 - p_2 = .05$, the probability of rejecting the null hypothesis is only .090.

Nonetheless, there is statistically persuasive evidence here of nonconstant success probabilities. Table 4 shows that the chances of throwing a double are somewhat lower

when pitching second but that, for both first and second pitchers, the chances of throwing a double are generally higher after throwing a double in the preceding one or two innings. A comparison with Table 1 indicates that controlling for pitching order makes the differences in observed conditional probabilities somewhat smaller but does not eliminate them.

If, whether a player pitches first or second, the probability of throwing a double does not depend on previous results, each player should be equally likely to be more successful after previous doubles than after previous nondoubles. I tabulated the number of players who were

Probability of rejecting $p_1 = p_2$

Figure 1. The probability of a sufficiently large difference in success frequencies to reject the null hypothesis that the success probabilities p_1 and p_2 are equal.

Table 4
Frequency of Doubles Following One or Two Doubles or Nondoubles, With Unweighted Averages

Players	Two Nondoubles		One Nondouble		One Double		Two Doubles	
	Freq.	Aver.	Freq.	Aver.	Freq.	Aver.	Freq.	Aver.
Men, 2000								
First	.495	.512	.518	.524	.536	.522	.558	.529
Second	.458	.466	.461	.466	.522	.500	.539	.515
Women, 2000								
First	.483	.522	.524	.534	.572	.553	.583	.552
Second	.441	.468	.483	.491	.538	.522	.548	.533
Men, 2001								
First	.482	.512	.548	.561	.598	.576	.625	.572
Second	.465	.501	.472	.489	.565	.531	.573	.522
Women, 2001								
First	.519	.534	.560	.562	.608	.591	.625	.593
Second	.468	.497	.495	.503	.553	.520	.561	.519
Total								
First	.494	.520	.536	.545	.579	.561	.600	.562
Second	.458	.483	.478	.487	.545	.518	.556	.522

more successful after doubles and the number who were more successful after nondoubles and used the binomial distribution to test this null hypothesis. In 81 of 128 cases (63%), players were more likely to throw a double after a double than after a nondouble ($p = .0017$), and in 85 of 127 cases (67%), they were more likely to throw a double after two doubles than after two nondoubles ($p = .00009$).

Correlations

If individual ringer probabilities vary from game to game, a player who is hot should have elevated success probabilities whether he is throwing first or second; a player who is cold should have deflated success probabilities. If so, looking across games, we might find a positive correlation between a player's success rates when throwing first and when throwing second. To examine this issue, correlation coefficients were calculated for each player, using data for every game in which the player had at least 10 innings of first pitches and 10 innings of second pitches.

With 64 players, we would expect to find some positive and negative correlations. The question is whether the number and magnitudes are greater than would be expected by chance. Of the 64 players, 42 (66%) had positive correlation coefficients. If positive and negative correlation coefficients were equally likely, the binomial distribution tells us that the probability of so many positive correlations is .0084. Eight of 64 players had positive correlation coefficients with p values less than .025. If the null hypothesis is true, the probability of so many statistically significant correlation coefficients is .0002.

Ringer Percentages by Game

We can also use a chi-square statistic to test the null hypothesis that a player's pitching-first and pitching-second ringer percentages do not vary from game to game. Table 5 shows the data used for Mary Ann Peninger in 2000. The calculations compare the actual and the ex-

pected values, as in Table 3, but the details are not shown here. The expected values are calculated by assuming that she has a .761 probability of throwing a ringer when pitching first and a .732 probability of throwing a ringer when pitching second. The p value is .0013, demonstrating that it is extremely unlikely that the observed game-to-game variations in her accuracy are simply random fluctuations about constant success probabilities. Overall, 9 of 64 players have p values less than .05, and 5 of these have p values less than .01. If success probabilities do not vary by game, the probabilities of so many low p values are .0044 and .0005, respectively.

DISCUSSION

Gilovich et al. (1985) argued that basketball performances are misperceived by fans and players to contain remarkable hot and cold spells. In addition to contradicting the perceptions of fans, their results seemingly contradict substantial evidence (e.g., Bandura, 1977; Taylor,

Table 5
Mary Ann Peninger's Ringers by Game in 2000, When Pitching First and When Pitching Second

Game	Accuracy When First	Accuracy When Second
1	.864	.571
2	.688	.500
3	.786	.688
4	.615	1.000
5	.900	.778
6	.722	.794
7	.750	.808
8	.867	.682
9	.633	.800
10	.833	.589
11	.767	.900
12	.727	.667
13	.800	.850
14	.607	.850
15	.778	.778
Total	.761	.732

1979) from a wide variety of sports that has indicated that athletic performance is enhanced by a person's self-efficacy, the personal assessment of one's ability to perform a specific task. In an arm wrestling study (Nelson & Furst, 1972), the weaker subjects won 10 of 12 matches when both contestants incorrectly believed the weaker person to be stronger than his opponent; when the contestants correctly identified the weaker contestant, the stronger subjects won all 12 matches. Other studies have shown that positive self-talk improves the performance of basketball players (Kendall, Hrycaiko, Martin, & Kendall, 1990), skiers (Rushall, Hall, Roux, Sasseville, & Rushall, 1988), and swimmers (Rushall & Shewchuk, 1989).

It is plausible that the positive reinforcement provided by an athlete's success increases self-efficacy and, thereby, enhances performance. But perhaps this enhancement is relatively small for professional basketball players, in relation to such confounding factors as shot selection, lengthy spells between shots, and strategic adjustments. Championship horseshoe data are cleaner in that every pitch is from the same distance and made at regular, brief intervals with intense concentration and little or no strategy. Variations in player performances within games and between games at the 2000 and 2001 World Championships indicate that success probabilities are not completely independent of previous outcomes.

REFERENCES

- BANDURA, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, **84**, 191-215.
- BATEMAN, G. (1948). On the power function of the longest run as a test for randomness in a sequence of alternatives. *Biometrika*, **35**, 97-112.
- GILDEN, D. L., & WILSON, S. G. (1995). Streaks in skilled performance. *Psychonomic Bulletin & Review*, **2**, 260-265.
- GILOVICH, T., VALLONE, R., & TVERSKY, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, **17**, 295-314.
- KENDALL, G., HRYCAIKO, D., MARTIN, G. L., & KENDALL, T. (1990). The effects of an imagery rehearsal, relaxation, and self-talk package on basketball game performance. *Journal of Sport & Exercise Psychology*, **12**, 157-166.
- NELSON, L. R., & FURST, M. L. (1972). An objective study of the effects of expectation on competitive performance. *Journal of Psychology*, **81**, 69-72.
- RUSHALL, B. S., HALL, M., ROUX, L., SASSEVILLE, J., & RUSHALL, A. C. (1988). Effects of three types of thought content instructions on skiing performance. *Sport Psychologist*, **2**, 283-297.
- RUSHALL, B. S., & SHEWCHUK, M. L. (1989). Effects of thought content instructions on swimming performance. *Journal of Sports Medicine & Physical Fitness*, **29**, 326-334.
- STEVENS, W. L. (1939). Distribution of groups in a sequence of alternatives. *Annals of Eugenics*, **9**, 10-17.
- TAYLOR, D. E. M. (1979). Human endurance: Mind or muscle? *British Journal of Sports Medicine*, **12**, 179-184.
- WILLIAMS, W. R., JR. (2002). [Score sheets from 2000 and 2001 world championships.] Unpublished raw data.

(Manuscript received October 5, 2001;
revision accepted for publication May 1, 2002.)

Call for Papers Electronic Archiving: Norms, Stimuli, and Data A Special Issue of *Behavior Research Methods, Instruments, & Computers*

In August 2004 *BRMIC* will focus on a new component of the Psychonomic Society web site: an archival repository for a variety of materials that are of general utility to researchers in the many fields of experimental psychology. These materials are most usefully made available in electronic (rather than printed) form.

Such materials may include:

- Norms for verbal and pictorial stimuli that may be useful to other researchers.
- Databases or data archives for model testing and evaluation.
- Technical supplements related to data analyses, equipment calibration, etc.
- Program source code for statistical analysis, stimulus generation, and similar applications.
- Visual and auditory stimuli for use by other researchers.
- Other tabular or graphic information that is too extensive for paper publication.

The archival web site will be formally inaugurated upon publication of the August 2004 special issue of *BRMIC*. Information about implementation of the archive will be available at <http://www.psychonomic.org/brmic/>.

Submissions are invited both for articles that include files for the archival web site, and for articles that define and illuminate the challenges and issues associated with such data archiving.

Submissions, or questions about them, should be sent to Jonathan Vaughan, Editor, at brmic@hamilton.edu. See <http://www.psychonomic.org/BRMIC/manuscript.htm> for instructions about manuscript submission. The cover letter or e-mail should conspicuously note that the contribution is "For the Special Issue on Archiving."

The deadline for submissions is January 1, 2004.