Queen's Economics Department Working Paper No. 1366

Correcting for bias in hot hand analysis: Analyzing performance streaks in youth golf

Christopher Cotton
Queen's University

Frank McIntyre
Rutgers University

Joseph Price
Brigham Young University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

9-2016

CORRECTING FOR BIAS IN HOT HAND ANALYSIS: ANALYZING PERFORMANCE STREAKS IN YOUTH GOLF

CHRISTOPHER S. COTTON, FRANK MCINTYRE, AND JOSEPH PRICE

Abstract. This paper illustrates the problems that arise with traditional tests for the hot hand and proposes instead using a consistent dynamic panel data estimator, which corrects for these problems and is easy to implement. Applying this estimator to a large dataset of amateur, youth golfers, we find no evidence of either hot or cold hand effects. When we restrict attention to the most-amateur of the golfers in our data, we do see weak evidence of a small hot hand. Thus casual athletes may experience small hot hands, but the effect does not persist among more serious athletes. This may give insight into why the belief in the hot hand in professional sports exists, even when the evidence suggests otherwise.

Date: **Current version: September 2016**.

Key words and phrases. hot hand, performance streaks, amateurs versus professionals

JEL: C1, H3.

Cotton: Queen's University, Department of Economics, Kingston, Ontario K7L 3N6, cotton@econ.queensu.ca. McIntyre: Rutgers Business School. Price: Brigham Young University.

1. Introduction

One of the most persistent beliefs in sports is that athletes sometimes have "hot hands," during which time their probability of making a shot or otherwise performing well increases. Those who experience the hot hand are said to be "in the zone," when their focus or awareness of the game has temporarily increased, raising their performance above its typical level.[1] In statistical terms, the hot hand represents positive serial correlation in performance outcomes beyond what may be explained by an athlete's ability and chance alone.

Despite the popular belief in the hot hand, there is no conclusive evidence that it actually exists.[2] In their seminal paper, Gilovich, Robert, and Amos [1985] test for the hot hand among professional basketball players. They point out that players who make each shot with some constant probability will occasionally have a game in which they make nearly every shot, and some games in which they miss nearly every shot. They argue that the relevant question in testing for the hot hand is not whether players experience streaks of above average and below average performance, but rather whether players experience such streaks more often than expected if a player has a constant probability of making a shot. Gilovich et al. [1985]'s primary contribution was to show that this is not the case. The streaks of better-than-average and worse-than-average performance exist no more often in the data than would be predicted given players' underlying skill and chance alone. They coin the term "hot hand fallacy" to describe the popular belief in a hot hand which, they argue, does not really exist.

Following Gilovich et al. [1985], a substantial literature has developed testing for the hot hand in a variety of settings.[3] Reifman [2012] provides a thorough review of this literature. In most settings, the researchers find no significant evidence in favor of the hot hand, showing that streaks take place no more often than predicted by average skill and chance. In other settings, however, researchers find evidence supporting a small hot hand effect. For example, Livingston [2012], and, to a certain extent, Pope and Schweitzer [2011] find evidence that a hot hand may exist in professional golf. Klaassen and Magnus [2001] find evidence of a small hot hand in professional tennis, and Dorsey-Palmateer and Smith [2004] find that some professional bowlers may experience hot hands. When a paper presents evidence in favor of the hot hand, the magnitude of the effect tend to be small.

Our analysis begins by highlighting two substantial issues with the existing hot hand literature. First, the literature typically employs tests for the hot hand that to work properly require each player to attempt a large number of "shots" per match, game or

[1]The hot hand is psychological. The increase in ability associatd with the hot hand is distinct from the increase in ability that is associated with an athlete working to increase her underlying skill or physical health.

[2]See discussion along these lines in Camerer [1989], Camerer and Loewenstein [2004], McFadden [2006] and Reifman [2012].

[3]The application to sports includes baseball [Albright, 1993], putting and dart throwing [Gilden and Wilson, 1995], horseshoes [Smith, 2003], bowling [Dorsey-Palmateer and Smith, 2004], PGA golf [Clark, 2005], and tennis [Klaassen and Magnus, 2001]. Additionally, [Cheng et al., 1999] look at the hot hand performance of mutual funds. [Hendricks et al., 1993] also consider the performance of mutual funds, but use a different definition of the hot hand compared with the other studies.

tournament. In the data, players are not attempting nearly enough shots each game for the traditional tests to return unbiased results.

Some papers address this concern by aggregating attempts across multiple games, matches, or days in order to observe a large enough number of sequential attempts to reduce the bias in the hot hand estimators [e.g. Green and Zwiebel, 2015]. The problem with doing this is that player skill or physical condition often changes from one day to the next. A player's ability often changes over the course of and across seasons. This is true even of the professional athletes typically analyzed in hot hand analyses. A basketball player who spends a game not shooting as well as he or his team expects him to shoot will work with coaches and practice his shot to get back to his potential. A tennis player who is not hitting her backhand as strongly as she would like will watch video and practice in an attempt to strengthen her swing. A golfer who spends a tournament consistently slicing his drive or missing putts will work to correct the problem before the next tournament. An athlete with an injury is likely to recover at least in part from one game to the next. In these situations, an athlete who consistently misses many shots in one tournament may consistently make more shots in the next; performance that traditional analyses may misinterpret as evidence that the player experiences a hot hand, since hits are clustered with other hits and misses with other misses even across days of play.

To illustrate this issue, we run traditional tests on simulated data that closely resembles the real performance data for participants in American Junior Golf Association (AJGA) tournaments. Although we are certain that no hot hand exists in this simulated data, the traditional tests find evidence that they do. This suggests that we should not trust results based on the traditional tests for the hot hand.[4]

A second concern with the hot hand literature is that it almost exclusively looks at professional or advanced amateur athletes taking actions that they have likely taken thousands of times before. It is exactly this group of people for whom we expect emotion, confidence and performance to be least effected when they make or miss a shot, or score above or below expectation on any given attempt. Because of this, we are unsurprised that these previous studies find little evidence of the hot hand. Even if we trusted the empirical methodology used to test for the hot hand, evidence suggesting that it does not exist in professional sports should not be interpreted as evidence that it does not exist. It could reasonably still exist among novice or amateur athletes, even if it does not exist among professionals.

After explaining our concerns regarding existing hot hand studies, we present an analysis of performance in youth golf designed to account for both concerns. First, to correct for the biases inherent in the traditional tests, we adapt standard dynamic panel data techniques designed to consistently estimate auto-correlation in small samples with

[4]Recent work by Green and Zwiebel [2015] finds that performance among baseball papers across many games is autocorrelated. This long trend hot hand is an interesting phenomenon but not the same as the short term effect we are interested in here. Miller and Sanjurjo [2015] discuss some problems with the early work on basketball players. They perform experiments with long series of free-throw shots for professional basketball players in Spain to show that there do exist some players with hot hand effects. They further show in Miller and Sanjurjo [2016] that there are econometric problems with standard methods of computing the hot hand. Though the presentation is different, the essential mechanism seems to be the same that gives us the induced negative autocorrelation in short panel fixed effects estimates we discuss below.

fixed effects. These tests are easily replicated in standard software packages. When we apply the tests to simulated data for which we know no hot hand exits, we correctly find no evidence in favor of the hot hand. This is in contrast to the traditional tests which consistently returned evidence of a small hot hand effect.

Second, we apply the unbiased test for the hot hand to a large dataset on the hole-by-hole performance of young, amateur golfers. The golfers in our data range in age from 12 to 17. Some of the participants are serious golfers who we observe in numerous tournaments per year across multiple years, and who go on to play golf in college or professionally. Others take the tournaments less seriously, only competing in one tournament. There is also wide variation in golfer skill. This allows us to look for the hot hand amongst young, amateur athletes of various ability levels, whom we speculate are more likely than professional athletes to experience the hot hand.

With a consistent test for the hot hand, we find no evidence that the hot hand systematically exists among the AJGS golfers. This is in contrast to the weak positive effects that we find when we employ the biased traditional estimators. Correcting for the bias in the hot hand estimators eliminates the evidence in support of the hot hand. We continue to find no support for the hot hand, even when we look at subgroups based on age or ability level. This means that not even the youngest golfers in our data systematically experience the hot hand.

These results do not rule out the hot hand completely, however. Although the golfers in our data are young amateurs, they also tend to be relatively serious golfers. This is particularly true of the golfers that participate in tournaments year after year. Recognizing this, we restrict attention to the golfers who participate in no more than one year of tournaments. These are the golfers that we refer to as the "most-amateur" golfers in our data. When we do this, we observe weak evidence suggesting that the most-amateur golfers experience small hot hands, on average. This suggests that casual golfers very well may experience a small hot hand effect; but this effect does not persist among the golfers who stick with competition.

Based on these results, we see no reason to expect the hot hand to exist among more serious amateur and professional athletes, even if we do see some suggestive evidence for it among the most-amateur athletes.

2. Data

We use both simulated and real data on performance in youth golf tournaments. We begin with a description of the real data, as the simulated data is designed to look as similar to the real world data as possible while ensuring performance independence across holes.

2.1. **Youth Golf Data.**

We use data we gathered from the American Junior Golf Association (AJGA) website on the hole-by-hole performance of individual golfers in golf tournaments from 2002 through 2009. The AJGA is the largest organization of youth golfers in the United States and runs roughly 75 open and invitational tournaments per year throughout the country. Participants in AJGA tournaments range in age from 12 to 17, and can participate in a maximum of five open tournaments, and an unlimited number of invitational tournaments per year. Each tournament consists of up to four rounds (18 holes each) spread across multiple days. All golfers participate in the first three rounds of the tournament, and depending on the tournament, participation in

subsequent rounds may depend on one's earlier performance. Thus we always limit our analysis to the first three rounds.

Golfers attempt to minimize the number of strokes taken in each round of golf. The golfer with the lowest score in each round wins the round. For each hole, a par is assigned which reflects the expected number of strokes required to complete the hole. The par on a given hole can be 3, 4, or 5 strokes. A golf course is typically made up of multiple holes at each level of par, with par-4 holes being the most common. A player's performance on each hole is judged relative to par. A player makes par on a hole when they complete the hole in par or fewer strokes. This is our "good" outcome. A "bogey" is one or more shots more than par, which we will use as our bad outcome to test for "cold hand".[5] Players make par on about 62% of holes and 38% (the rest) are bogeys. We also need to define what we mean by a hot or cold hand. In this case, we look at outcomes after a player has succesfully made par three times in a row and then test for evidence that they do particularly well on the next hole. Evidence that they do would be evidence of hot hand. Conversely, a cold hand is when a player does poorly on a hole after three or more consecutive bogeys.

While we do not have data on the performance of players prior to 2002, we do have a list of participants for all of the tournaments in 2001. We exclude all players who participated in two or more tournaments in 2001 from our analysis. This assures that we have an accurate measure of a player's actual experience in the AJGA. Our analysis includes 8436 individual golfers, 53% of whom participate in more than one year of tournaments. In testing for hot and cold hands, we consider the probability of success on a total of 1,316,718 player hole observations.

2.2. **Simulated Data.** We generated simulated data designed to broadly reconstruct the patterns in the data, but known to have independence in unobservables across holes. Our basic strategy is to create a set of data composed of holes with the same difficulty as that observed in our data with players of the same ability, and then simulate tournaments for them where we impose that unobserved outcomes are random and uncorrelated, i.e., no hot or cold hand. The appendix provides a detailed explanation for the construction of the simulated dataset. The summary statistics for the data look remarkably like those for the real data. Except that in the simulated data we are confident that no hot hand exists.

Table 1 compares the simulated and real data. The first section lists fixed characteristics that are by construction identical across the two data sources. We have 8436 golfers and the typical par is just under 4. 40% of our observations on players come from their first tournament, 24% of players are only seen in one tournament, 47% show up only in one year of data, and 28% are active in three or more years.

In the real data, the average player achieves par 61.7% of the time, while in the simulated data, the average player achieves par a slightly higher 63.2% of the time. This slight difference is driven by the par rate among golfers with the lowest ability level. In the real data, the tournament participants in the bottom 10% of ability achieve par 38.5% of the time, while in the simulated data, they achieve par 42.1% of the time. In the real data, the golfers achieve par on the most recent three holes in a tournament 31.2% of

[5]To maximize power, and provide a clear description, we do not discuss or make use of any extra information in double bogeys or scoring one or more below par. We simply refer to every outcome as either "par" or "bogey".

TABLE 1. Real Data vs. Simulated Data

	Real	**Simulated**
Number of players	8436	8436
Average par	3.98	3.98
Player's first tournament	40%	40%
Players younger than grade 9	9%	9%
Players with 3 or more total years of experience	28%	28%
Players with only 1 year of total experience	47%	47%
Players only seen in one tournament	24%	24%
Players with multiple tournaments	6381	6381
Average player par rate	61.7%	63.2%
High ability (Top 10%) par rate	79.9%	79.8%
Medium ability (Middle 80%) par rate	64.2%	64.3%
Low ability (Bottom 10%) par rate	38.6%	42.1%
3-par Streaks	31.2%	31.1%
3-bogey Streaks	5.4%	4.8%

the time, compared to an almost identical 31.1% of the time in the simulated data. They miss three holes in a row 5.4% of the time in the real data, and 4.8% of the time in the simulated data.

As we discuss in the appendix, recreating the data can be a tricky process. While matching means is easy, we also want a reasonable approximation of the full distribution of tournament outcomes. Although there are some minor differences in golfer performance in our real and simulated data sets, the broad outlines are the same, letting us use our simulated data in the coming analysis. Our objective in constructing the simulated data is simply to have a data set that captures the essentials of the real data, but for which we are confident that no hot hand exists. We will use the simulated data to illustrate biases in traditional hot hand tests, by showing that they produce evidence of a hot hand in the simulated data set, where we know no hot hand exists.

3. Analysis

We present a number of tests for the hot hand using both the real junior golf data, and the simulated data for which we are certain that performance is independent across holes. Although some of the traditional tests return evidence in support of performance streaks, they do so in both the real and simulated data. This means that the tests find evidence in support of a hot hand, even when no hot hand exists. After the traditional analysis, we propose an alternative test for streaky performance, designed to correct for the biases present in the traditional tests.

In addition to considering the hot hand, we also look for evidence of a cold hand, defined as in increased probability of poor performance (e.g. a bogey or worse) following poor performance on recent holes.

3.1. **Traditional tests.** We are interested in determining if good performance on a given hole leads to good performance on the next hole. A number of different methods are available for testing this kind of dependence, but unfortunately not all of them give satisfactory results. The chief problem is that a round, or at most a tournament of golf is (or should be) the basic unit of analysis and it is only 18 or 54 holes (and thus 18 or 54 observations) long. This is too few observations to take advantage of many asymptotic results. Similar problems are present or even worse in other settings in which the hot hand has been studied; in basketball the average NBA player attempts 11 shots and bowlers complete only 10 frames per game.

To illustrate these issues, we simulate data with known properties and test common estimation techniques on the simulated data for which we are certain that players do not experience hot and cold hands more often than predicted by their ability, and across game performance trends. We find that many common methods of analysis require assumptions about the data generating process that are not likely to be met for young players. This includes an assumption that player unobservable characteristics are fixed over one or more years. Allowing for more flexible unobserved player ability controls leads to small sample problems per panel.

To deal with the limited number of observations per event, others have pooled performance data across multiple events, considering for example whether there is evidence that a player exhibits hot and cold hands across an entire season of data. Such an approach, however, requires that players maintain a constant ability across each of the multiple games or tournaments. It does not allow for the possibility, for example, that a player who sees a weakness in his or her game will work to improve on that aspect of their performance in between events.

When an athlete's performance improves across games or tournaments due to practice and hard work, the better performance could be interpreted as an increase in ability or skill rather than evidence of a hot hand. Traditional tests of the hot and cold hand conducted across multiple tournaments of data will suggest that a player who improves over time is a streaky player, cold in earlier tournaments and hot in later tournaments. Other factors may cause players' ability to decrease across games or tournaments; these include injury, fatigue, or distractions that arise in their personal life. Traditional analyses may suggest that these athletes experienced hot hands when healthy and cold hands during the times of injury, fatigue or personal crisis.

Given these issues, we find that common estimation techniques are prone to finding both hot and cold hand effects that are statistical artifacts or that are perhaps not best described as hot hand, but rather as improvement.

3.1.1. *Wardrop tests.* A first, simple analysis is based on a 2x2 table where we look at the probability of making par given that the player made par on at least the last three holes. Our choice of three holes is partially motivated by the prior literature. Obviously, requiring longer prior streaks makes one more confident that the remaining test cases are a good sample for testing streaks. On the other hand, the number of observations available drops precipitously.[6]

[6]Requiring shorter streaks before testing for a hot hand does not do much to change the results we get, since we find so little evidence to begin with. Longer streaks for bogeys become problematic as they happen increasingly rarely, but we do consider long streaks for par. We look at the performance of a player on the next hole after having made the prior six holes for our preferred regression estimator in Table 6

In this basic analysis, we do not account for player identity, let alone the fact that an individual's ability may change over time. The evidence, presented in Table 2, suggests the existence of hot hand regardless of whether we use the simulated or real data.

TABLE 2. Pearson Chi-Square Table Aggregated Over All Players

	Simulated Data		**Real Data**	
	Par	*Not*	*Par*	*Not*
Prior 3-Hole Par Streak	36.2%	63.8%	35.3%	64.7%
Prior Hole Not a Streak	29.4%	70.6%	30.0%	70.0%
N	1,316,718		1,316,718	
Chi(1)	5951		3512	

Note: Pearson's chi-square test on a 2x2 table to see if the probability of success or failure is independent of previous streaks. Results aggregated over all players and tournaments.

Using the simulated data, Table 2 shows that the probability of achieving par on a given hole is substantially higher following a string of 3 pars (or better), than following three holes that were not all successes. A player who has had three recent successes has a 36.2% chance of achieving par on a given hole, compared to 29.4% chance of par for a player who did not immediately previously have a string of three successes. The gap is slightly smaller, but still highly significant for the real data. Our Chi-Square test statistics are 5951 and 3512, when we would have rejected the null at any value over 4. This result would be evidence in support of the hot hand in youth golf, except for our suspicions about the appropriateness of the methodology. In particular, we know that the hot hand did not exist in the simulated data.

A possible driver of these results is that fact that each player has a different chance of making par, and so if the golfer made par on the last hole, that makes it more likely that he is a good player (Wardrop (1995) discusses this problem in detail). This suggests that there is a "player fixed effect" confounding our attempt to uncover serial correlation. Worse, a given player may improve over time, thus a player's idiosyncratic ability may not be constant over all tournaments. These results illustrate how a traditional test of the hot hand can identify evidence in support of performance streaks in situations in which no such streaks exist.

Recognizing that differences across individuals may be driving the results, Wardrop (1995) proposes that a similar analyses be run on the performance data of each individual in the analysis, and determine the portion of individuals that seemingly exhibit hot or cold hands. As discussed previously, we are concerned that some players may improve or stagnate from one tournament to the next, particularly considering the age and experience level of some of the athletes in our sample. We do not want to interpret such trends as evidence in favor of a hot hand. We therefore consider the holes for a given player at a given tournament. Given 18 holes per round of golf, and three rounds

and find no evidence of hot hand in either the simulation or the real data, with tight standard errors of 0.7%. Thus the main results do not appear to be overly sensitive to streak length.

per player per tournament, this gives us 54 holes per player, and 51 total holes for which there exist three prior holes during the same tournament.[7]

TABLE 3. Wardrop Test: Illustrative Example

	Par	*Not*
Prior 3-Hole Par Streak	33%	67%
Prior Hole Not a Streak	20%	80%
N	51	
Chi(1)	0.9	
P-value	0.34	

Note: Illustration of Pearson's chi-square test for hot hand of hitting par. Based on a single player in a single tournament with 54 holes.

An example table for an individual player-tournament is given in Table 3. This example player performed better on the holes immediately following a string of pars, than he did otherwise. Getting three successes in a row seemingly increases this player's chances of achieving par on the next hole from 20% to 33%. Despite these large effects, 51 observations is not much to work with, and Pearson's Chi-square test makes clear that the results could well have occurred by chance in independent data (p-value of 0.34).

This test's lack of power can be readily dealt with. The same test can be repeated for every player-tournament combination. Under the null hypothesis of independence, each of these tests is a chi-square with one degree of freedom. Thus we can perform a joint test of the sum of them. Table 4 reports on the joint test across the entire data set.

For both the simulated and real data, we observe that the hot hand (evidence that there is a higher probability of par following a string of pars) occurs slightly but significantly more frequently than expected in independent data (5.3% and 5.6% compared to the expected 5%). We run a test on this overall sample and show that we can reject independence ($p < 0.001$) which suggests the existence of hot hand. Unfortunately, we get the same test rejection for both simulated and real data. Thus this test appears to once again be finding hot hand where it does not exist.

We see even odder behavior when we run the same analysis but look at cold streaks, i.e. the probability a player fails to achieve par after experiencing a bogey or worse on each of the previous three holes. Here, we find that evidence in favor of the cold hand occurs too infrequently, even compared to independence! The individual tests do not reject enough (1.7% and 2.3% instead of 5%) and give Chi square values that are in the wrong tail. This is a pretty clear signal that the test is misspecified. Further, we find the same phenomenon in both the real and simulated data, thus the cause of the problem is sufficiently widespread to affect data known to have no unobserved streaks.

This misspecification for the cold hand tests could be at least in part driven by the fact that bogey streaks precede only about 5% of holes. This implies that some of the cells of the individual 2x2 tables are barely populated when there are only 51 observations on

[7]We could do this by round as well, but then the small sample problems this test has with empty cells (discussed below) would become substantially more severe.

TABLE 4. Wardrop Test, Joint Across All Player-Tournament Pairs

	Rejection rate	Test statistic	Critical Value	P Value
Simulated Data:				
Par streak	5.3%	26631	25655	<0.001
Bogey streak	1.7%	13735	17607	>0.999
Real Data:				
Par Streak	5.6%	26813	25501	<0.001
Bogey Streak	2.3%	14339	17703	>0.999

Note: Chi-Square test adding up individual tests of each player-tournament pair. Done for both tests if par streaks (making par at least the last three holes) lead to another par and if streaks of three or more bogeys increase the probability of another bogey. Rejection rate is the fraction of individual Chi-Square tests (like the illustration in the prior table) that reject independence. Test Statistic is the sum of the Chi-Square test values across the data while the Critical Value and P-value give the test results.

an event that occurs one time in twenty. This is a well-known issue with Pearson's test which is based on an asymptotic approximation. To a lesser extent, this may also be a problem for some of our par streak data. Although there are small sample fixes, we will not pursue these here, as we have better options.

Given that the performance patterns in the real data also occur in the simulated data, we are not comfortable interpreting the results as evidence for or against the hot and cold hand. Overall, our simulation suggests that this is a poor test for low probability events, as expected. For par, the results appear more reasonable, but still unconvincing, since simulated data triggers the same rejection of the null when it should not.

3.1.2. *Runs test.* A somewhat different testing approach proposed in the literature considers whether or not golfers exhibit "runs" of good or bad performance. This is a chi-square test based on the number of switches between good and bad performance for independent data. Note that here we have some concerns about small sample problems, because the natural unit of observations is going to be 18 holes of golf, which will not generate very many switches between par and bogey runs. A symptom of this is that for some rounds of golf, the test statistic cannot be computed because there are no switches (when the golfer always makes par or never does).

Table 5 shows that the runs test does correctly identify a lack of hot hand in the simulation. Which is encouraging. 4.8% of the run sequences are individually rejected, but more importantly, the pattern of runs across the whole sample is consistent with a Chi-square distribution (p = 0.96). The real data also has the correct number of expected rejections (4.8%) but the distribution across all the runs is not consistent with a Chi-square (p=0.05). Thus this gives us our first indication of possible hot or cold hand.

Unfortunately, this test does not lend itself to easy interpretation. The Chi-square statistic is only marginally rejected on an exceptionally large dataset. Thus, this may either be a false positive, or there is some hot hand or cold hand, but barely enough

TABLE 5. Wald-Wolfowitz Runs Test

	Rejection rate	Test statistic	Critical Value	P Value
Simulated Data	4.6%	76469	77162	0.96
Real Data	4.8%	77701	77052	0.05

Note: Chi-Square test if number of observed runs of par or not par in a given player-tournament conforms with independence. Rejection rate is the fraction of individual Chi-Square tests that reject independence. Test Statistic is the sum of the Chi-Square test values across the data while the Critical Value and P-value give the test results.

to register. Further, because we are simply looking at streak lengths, we do not have any tools for differentiating between hot and cold hands. Once again, we could push farther on there results, providing subsample analyses or refining the tests in some way, but instead we will turn to a more natural toolkit for the applied economist and see what we can learn with regression analysis, where the coefficients will provide readily interpretable measures of how much hot or cold hand we are seeing.

3.2. **Regression Analysis.** An alternative approach uses regression analysis. As we noted before, we will still need a way to deal with unobserved differences across players. In OLS, in theory, this can be done using an adjusted OLS fixed effects strategy (for a logit method that deals with the problem – for which there are more stringent requirements on the data – see Honoré and Kyriazidou [2000]). These regression methods have distinct advantages in that they are readily understood by most researchers, they allow us to control for confounding effects such as changing par across holes, or changing secular patterns across the 18 holes of the round, and they immediately deliver interpretable coefficients of how much hot hand or cold hand there may be.

We specify an econometric model where doing well in the previous period is correlated with doing well this period. Thus we are looking for serial correlation in a model of the form:

$$G_{iht} = \alpha + \rho I(G_{iht-1} = G_{iht-2} = G_{iht-3} = 1) + \beta X_{iht} + \gamma Z_{it} + \epsilon_{iht}$$

where i is the player, h is the hole, and t is the tournament.[8] We are most interested in the coefficient ρ which tells us how the next hole is affected by making the last three holes or, in the bogey specification, by missing the last three. X_{iht} is the probability of making par (or hitting a bogey) for the given hole, as a more refined measure than par of the hole difficulty.[9] Z_{it} contains a measure of player ability such as average player performance or fixed effects as discussed below.

Table 6 reports the coefficient of interest, ρ, for seven alternative regression analyses. Note that, for ease of interpretation, we've multiplied all the coefficients in this and

[8]We have simplified the notation slightly here, since hole goes from 1-18 and each hole is played three times across three rounds of golf for a given tournament

[9]In earlier specifications, we used dummies for par itself instead of this ability measure, and the results were identical.

TABLE 6. Regression analysis under alternative specifications

Regression Type	**Par Streak**	**Bogey Streak**
Simulated Data		
No controls	5.27***	9.44***
	(0.11)	(0.24)
Linear control for ability	-0.63***	-0.68***
	(0.09)	(0.20)
Flexible polynomial control	-0.58***	-0.87***
	(0.09)	(0.20)
Flexible polynomial control (external ability)	0.91***	2.53***
	(0.09)	(0.22)
Player-tourn. fixed effects	-2.80***	-3.67***
	(0.09)	(0.20)
Player fixed effects	-0.66***	-1.00***
	(0.09)	(0.20)
Dynamic panel data model	**0.04**	**-0.21**
	(0.24)	(0.80)
Actual Data		
No controls	6.88***	13.57***
	(0.11)	(0.27)
Linear control for ability	0.54***	1.30***
	(0.09)	(0.20)
Flexible polynomial control	0.62***	1.04***
	(0.09)	(0.20)
Flexible polynomial control (external ability)	2.41***	5.13***
	(0.10)	(0.23)
Player-tourn. fixed effects	-2.32***	-2.65***
	(0.09)	(0.19)
Player fixed effects	0.57***	0.91***
	(0.09)	(0.20)
Dynamic panel data model	**0.07**	**0.43**
	(0.24)	(0.76)

Note: N = 1316718 for full sample. All coefficients and standard errors have been multiplied by 100. Coefficients are the ρ values from the regression specification discussed in text. Thus a coefficient of 5 indicates a 5 percentage point increase in the probabiilty of making par after a par streak. Standard errors clustered by player-tournament.

the following tables by 100, to make them percentages. Thus a coefficient of "1" means that making par in all of the last three rounds makes par one percentage point more likely this round. For example, a player who normally hit par 35% of the time would be expected to make par 36% of the time after a good streak. Column 1 reports how achieving par or better on each of the three most recent holes affects the probability of achieving par on the next hole. Column 2 reports how performing worse than par (i.e. a bogey or worse) on the past three holes affects the probability of achieving worse than par on the current hole.

We first consider the results for our simulated data, where we know that no hot hand actually exists. When we include no controls (in row 1), we find spurious evidence of a hot hand effect. Following a string of pars, there is a 5.27% increase in the probability of another par. Following a string of bogies, there is a 9.44% increase in the probability of another bogey. Both results are highly statistically significant, given our massive sample size. What we are picking up here is unobserved player ability (which we included as part of the simulation). In the second row, we add a simple linear control for the player's average performance (measured as their par rate or bogey rate respectively) across all their tournaments and the hot and cold hand results not only go away, but are also reversed, showing evidence of small negative hot and cold hand effects, implying that a string of past successes (failures) decreases the probability of another success (failure) by about 0.6%.

We can expand on this approach by allowing for more flexible ability controls. We include a fourth order quartic polynomial in our performance measure. This gives us essentially the same result – a light negative effect. One concern here is that the data used as a "control" include the same data used as the dependent variable and the auto-correlation data. This can create a bias in the data, thus we restrict the sample to only those who compete in two or more tournaments and use the player's performance in other tournaments as a control for the current tournament. We call this a measure of their external ability and we include it as a fourth order polynomial term. This flips both the par and bogey results for the simulation data to estimating small, but spurious, hot and cold hand effects (0.9% and 2.5% respectively).

We then turn to fixed effect approaches. The next two rows include fixed effects first by player-tournament and then by player. Player-tournament fixed effects estimate a fairly strong negative correlation. This is a well known problem with fixed effects in short panels and, in fact, with no other regressors the size of the bias on the autocorrelated term, when there is no actual correlation, is proportional to $1/T$ (times 100 for our table), the inverse of the length of each panel. In our case, the panels are 51 observations, after we lose one hole each round over a three round tournament for the lags. Thus the observed bias we get with fixed effects is about what we expect to be explained by a short panel bias. And since our earlier ability controls are really very similar to these fixed effects, they show the same bias. The one exception is the external ability control which evades this particular trap, though it is still not able to give consistent results in the simulated data.

One way to address the issue of short panel bias is to use an alternative set of fixed effects for which there are more observations within group. In Table 6, the next row includes player fixed effects rather than player-tournament fixed effects. This reduces the bias substantially, though it does not eliminate it entirely.

3.2.1. *Correcting for bias: Dynamic panel model.* Similar dynamic panels with potentially endogenous regressors have been extensively considered in other contexts; enough so that readily prepackaged estimators in STATA are available for dealing with the short panel problems that arise. Although the solution we pursue here is not the only available one, the fact that it will be easy to replicate by other researchers, and easily applied to other data, argues heavily in its favor.

We estimate a version of Arellano and Bond's (1991) estimator for panel data, where one estimates on differenced data using lags of the dependent variable to construct

instruments to overcome the inconsistency we've identified in the simulation. This procedure is easy to use in STATA and reasonably versatile. We constrain the estimator to only use three lags (i.e., outcomes from holes four, five, and six in the past, rather than all past lags) as multiplying the number of instruments can be problematic, both computationally and theoretically. As it turns out, we have more than enough power with these, as the standard errors are all quite small.

We modify Arellano and Bond's design to take advantage of improvements noted by Windmeijer (2005) that add efficiency through a two step procedure while accounting for possible bias introduced. This modification is also readily available in STATA's prepackaged routines and makes the standard errors robust to correlations within each player-tournament combination. Of course, any such correlation would be a sign of a hot or cold hand, but we find no evidence of such effects. This new estimator performs exceptionally well. It finds zero evidence of hot or cold hand in the simulation data, while still being only slightly less precise, as indicated by the standard errors, than prior estimators.

We take away from this a couple of lessons. First, given the potential pitfalls of this kind of estimation, any technique should be validated on a carefully constructed simulated sample to ensure that the estimator is, in fact, consistent. Second, in this particular application, the Arellano-Bond dynamic panel data approach is consistent and, thanks to modern software, fairly painless to implement. This is in sharp contrast to the OLS and the fixed effects estimators, both of which are biased, though not always in the same direction.

Turning to the real data, we see patterns of hot hand or cold hand that mimic our simulations. Depending on the regression, the analysis of the real data goes from suggesting the existence of large, small, and negative hot and cold hands. In some cases, the sign here is opposite from what we got with the simulation, but because these regressions produce untrustworthy results with the simulated data, we dismiss the significant positive and negative effects as potentially spurious.

However, we also identified a consistent estimator, which we can apply to the real data. Using the dynamic panel data estimator, the analysis returns very small point estimates for our coefficient ρ statistically inseparable from 0. Therefore, once we correct for bias in the analysis of the hot hand, we find no evidence that hot or cold hands exist among junior golfers. Given the small size of the coefficients and standard errors, and the large size of our data set, we are confident that on average, no hot or cold hand effects are present among the youth golfers in AJGA tournaments.[10]

This does not, however, establish that no hot or cold hands exists among any subgroup of golfers. Just that they do not exist on average. To explore this possibility, we estimate our dynamic panel data specification on different subsamples of our data defined by golfer experience and ability level. Table 7 reports these results.

Overall, we see little evidence of any hot or cold hand within the groups defined by age, or ability level. This means that even when we restrict attention to the youngest golfers, we see no evidence of hot and cold hands.

The only time we observe weak evidence of small hot and cold hand effects is when we restrict attention to those golfers that participate in only one year. These golfers are

[10] As noted previously, requiring par streaks of six instead of three does raise the standard errors, but still returns no evidence of hot hand in the simulated or real data.

TABLE 7. Dynamic Panel Data Estimator Coefficient (x100) on continuing a streak on the next hole by experience and ability using actual data

	Par Streak	**Bogey Streak**
Whole sample	0.07	0.43
	(0.24)	(0.76)
Those that eventually get at least 3 yrs experience	-0.02	0.17
	(0.29)	(1.19)
Younger than grade 9	-0.30	-1.37
	(0.84)	(2.26)
Max 1 yr experience	1.26*	2.14*
	(0.65)	(1.26)
1st tournament ever	0.58	1.38
	(0.68)	(1.28)
Only 1 tournament ever	1.80	3.14*
	(1.20)	(1.81)
More than one tournament ever	-0.09	-0.09
	(0.24)	(0.83)
High Ability	-0.21	-3.23
	(0.51)	(5.99)
Medium Ability	-0.01	0.32
	(0.27)	(0.96)
Low Ability	-2.25	-2.18
	(2.84)	(1.45)

Note: N = 1316718. All coefficients and standard errors have been multiplied by 100. Coefficients are the ρ values from the dynamic panel data regression specification discussed in text. Thus a coefficient of 5 indicates a 5 percentage point increase in the probabiilty of making par after a par streak. Standard errors clustered by player-tournament.

likely less experienced, and less committed to tournament play than the other golfers in the data. We think of these as the most-amateur of the amateur golfers that we observe. Although we hesitate to make too much out of these results, given that the effects are small (1-3%) and only marginally significant, the results for these most-amateur golfers are interesting nonetheless.

Specifically, the evidence suggests that those who participate in only a single season of golf experience small hot and cold hand effects, with a string of past successes increasing the probability of another success by 1.26%, and a string of past failures increasing the probability of another failure by 2.14%. For those who participate in only one AJGA tournament ever, the cold hand effect is slightly larger at 3.14%, and although the hot hand effect is insignificant it is positive and of a similar magnitude (1.8%). This suggests that some of the most-amateur golfers in our data do tend to experience small hot and cold hands. Or alternately, some of them experience large effects while others experience no hot or cold hand, averaging our results out to a small, statistically significant effect overall.

4. Discussion

The paper discusses the problems that arise with the traditional tests for the hot hand. We propose a consistent estimator, which corrects for these problems. When we apply this estimator to test for hot and cold hands among amateur, youth golfers, we find no evidence of either hot or cold hand effects among the golfers, on average. This is true even when we restrict attention to golfers of a given ability, or the youngest golfers in our data. This suggests that the hot hand does not exist, at least among the amateur golfers than participate in AJGA tournaments. And, if it does not exist within this group of athletes, we expect that it is even less likely in settings with more mature, and more experienced athletes.

When we restrict attention to the most-amateur of the golfers in our data (defined as those who participate in no more than one year of tournaments), we do see weak evidence of a small hot hand that is not present among the other groups of golfers in our data. This prevents us from ruling out the hot hand entirely. Rather, the evidence suggests that casual athletes may experience small hot hands, but that the effect does not persist among more serious athletes (not even among the youngest). This may give insight into why the belief in the hot hand in professional sports exists, even when the evidence suggest it does not: someone who has been a casual participant in an activity may recognize that they have experienced a hot hand, and expect that others (including the professionals engaging in the same activity) do as well.

Appendix A. Constructing the simulated data set

Here we explain our procedure for simulating a sample of data on junior golfers making par. We perform an exactly symmetric simulation to generate simulated data on bogeys.

We first construct an estimate of the empirical difficulty of each hole h at each tournament t, defined as the observed probability that players, on average, make par on that hole. This is, in essence, an empirical version of "par," but more fine-grained and we label it e_{th}. This is constant across players.

We first split the sample into two groups, those with only one tournament and those with multiple tournaments. Consider first the players with multiple tournaments where we have both their outcomes at this tournament, as well as their outcomes at other tournaments. We average a player's performance over the rest of their career– giving each player a measure of their probability of making par but one that excludes the current tournament. We refer to this as their "external ability." Next, we average a player's performance over the given tournament– giving each player a measure of their probability of making par in that tournament. This is what we refer to as their "internal ability."

We then estimate two probit models, one using internal ability and the other using external ability, of the probability of making par for a given hole for a given player where the latent probability index is defined as:

$$y_{ijh} = \gamma f(\theta_{it}) + \beta_1 X_{it} + \beta_2 e_{th} + \beta_3 P_{th} + \delta_1 e_{th}\theta_{it} + \delta_2 P_{th}\theta_{it} + \delta_3 e_{th} X_{it} + \delta_4 P_{th} X_{it} + \varepsilon_{ith}$$

where θ is either internal or external ability and $f(\cdot)$ a quartic polynomial in the player's ability. X_{it} are fixed effects for years of player i's experience (1,2,3+) at a given

tournament t. p_{th} are par fixed effects (4,5). e_{th}, our empirical measure of hole difficulty, and p_{th} are interacted with a linear term in the player's ability and the experience fixed effects. Thus players that are typically good might be exceptionally good at hitting par 3 holes or holes near the beginning of the round, or holes in the final round of the tournament. The simulated data would pick up this pattern, but would still not have any structural dependence between the error terms of consecutive holes.

If we just used external ability for our measure, this estimator of hit probability would fail to pick up the fact that players do well on some courses and poorly on others, or simply have good or bad tournament days, a phenomenon related to "hot hand", but not what we are attempting to measure here. Thus, we would expect this estimator to be right on average, but the variance across players will typically be too low for not picking up these patterns. [11] A strictly internal ability measure would catch the tournament to tournament patterns but, due to the simulation variance, have higher variance across players compared to the actual data. Thus we considered a mixture of the two estimators that would maintain the benefits of each, but would have the correct cross-player variance. To that end, we estimate both probit models to predict the probability a player will make par on a given hole. We then take a weighted average of the two probabilities, where the weights are the number of external tournaments competed in for the external ability measure and two for the internal measure. This gives more weight to the current tournament, but as the external measure becomes more precise we allow it to dominate.

For players who compete in only one tournament, we have no external measure of ability, so we perform a simple average of the player's internal ability and the average performance of all new players on the given hole. If we just used average performance we would get too little player-level variance in our data, but if we just use player outcomes we add simulation variance to our player outcome variance and end up with too much variance. We found that a simple average of the two balanced these forces fairly well. This gives us the right amount of streakiness that would be observed even without hot hand effects, accounting for characteristics of the holes and players.

Given this mixed estimator, we have a probability of making par for each player at each hole in the dataset. We draw a uniform random variable to see if the simulation indicates that that player made par on a particular hole by comparing the drawn random variable to the estimated probability. This gives us a dataset that matches the real data in many details, such as changes across par or across players, as well as having a similar distribution of player talent. We then do the exact same thing for our simulated bogey data, substituting the word "bogey" for "making par" in the above.[12]

References

S. Christian Albright. A Statistical analysis of hitting streaks in baseball. *Journal of the American Statistical Association*, 88:1175–1183, 1993.

Colin Camerer. Does the basketball market believe in the 'hot hand'? *American Economic Review*, 79(5):125–161, 1989.

[11] For example, if there are tournament specific effects that add variance to the outcomes.

[12] Since each outcome is simply one minus the other, the simulated par and bogey data are essentially mirrors of one another.

Colin Camerer and George Loewenstein. Behavioral economics: Past, present, future. In *Advances in Behavioral Economics*. Princeton University Press, New Jersey, 2004.

Louis T. Cheng, K. Pi Lynn, and Don Wort. Are there hot hands among mutual fund houses in Hong Kong? *Journal of Business Finance and Accounting*, 26(12):103–135, 1999.

Richard Clark. Examination of hole–to–hole streakiness on the PGA tour. *Perceptual and Motor Skills*, 100:806–814, 2005.

Reid Dorsey-Palmateer and Gary Smith. Bowlers' hot hands. *American Statistician*, 58(1): 38–45, 2004.

David L. Gilden and Stephanie Gray Wilson. Streaks in skilled performance. *Psychonomic Bulletin and Review*, 2:260–265, 1995.

Thomas Gilovich, Malone Robert, and Tversky Amos. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17:295–314, 1985.

Brett S. Green and Jeffrey Zwiebel. The hot-hand fallacy: Cognitive mistakes or equilibrium adjustments? evidence from Major League Baseball, May 2015. URL `http://ssrn.com/abstract=2358747`.

Darryll Hendricks, Patel Jayendu, and Richard Zeckhauser. Hot hands in mutual funds: Short–run persistence of relative performance. *Journal of Finance*, 48(1):93–130, 1993.

Bo E. Honoré and Ekaterini Kyriazidou. Panel Data Discrete Choice Models with Lagged Dependent Variables. *Econometrica*, 68(4):839–874, 2000.

Franc Klaassen and Jan Magnus. Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, 96:500–509, 2001.

Jeffrey A. Livingston. The hot hand and the cold hand in professional golf. *Journal of Economic Behavior & Organization*, 81(1):172–184, 2012.

Daniel McFadden. Free markets and fettered consumers. *American Economic Review*, 96 (1):5–29, 2006.

Joshua B. Miller and Adam Sanjurjo. Surprised by the Gambler's and Hot Hand Fallacies? A Truth in the Law of Small Numbers. 2016. URL `http://ssrn.com/abstract=2627354`.

Joshua P. Miller and Adam Sanjurjo. A Cold Shower for the Hot Hand Fallacy. 2015.

Devin J. Pope and Maurice E. Schweitzer. Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes. *American Economic Review*, 101(1):129–157, 2011.

Alan Reifman. *Hot hand: The statistics behind sports' greatest streaks*. Potomac Books, Virginia, 2012.

Gary Smith. Horseshoe pitchers' hot hands. *Psychonomic Bulletin and Review*, 10:753–758, 2003.