A Cold Shower for the Hot Hand Fallacy: Robust Evidence that Belief in the Hot Hand is Justified

Joshua B. Miller[a] and Adam Sanjurjo[b] *†‡§¶

August 24, 2019

Abstract

The hot hand fallacy has long been considered a massive and widespread cognitive illusion with important implications in economics and finance. We develop a novel empirical strategy to correct for several fundamental limitations in the canonical study and replications, conduct an improved field experiment to test for the hot hand in its original domain (basketball shooting), and gather all extant controlled shooting data. We find strong evidence of hot hand shooting in every dataset, including on the individual level. Also, in a novel study of beliefs, we find that expert observers *can* predict (out-of-sample) which shooters are hotter.

JEL Classification Numbers: C12; C14; C91; C93; D03.
Keywords: Hot Hand Fallacy; Hot Hand Effect; Hot Hand Beliefs; Field Experiment.

*a: Department of Economics, University of Melbourne, b: Fundamentos del Análisis Económico, Universidad de Alicante. Financial support from the Department of Decision Sciences at Bocconi University and the Spanish Ministry of Economics and Competitiveness (Project ECO2015-65820-P) is gratefully acknowledged. We also gratefully acknowledge Generalitat Valenciana (Research Projects Grupos 3/086 and PROMETEO/2013/037. We thank José Valeiro, president of the Pabellón Municipal de Betanzos, for providing exclusive use of the basketball court, Javier Lopez for his jack-of-all-trades assistance in organizing and conducting experimental sessions, and Cristina Lopez for greatly improving the translation of our experiment's instructions to Spanish.

†Both authors contributed equally, with names listed in alphabetical order.

‡This draft has benefitted from helpful comments and suggestions and we would like to thank Pedro Albarran, Jim Albert, Alberto Alesina, Jose Apesteguia, David Arathorn, Jeremy Arkes, Olivier Armantier, Daniel Benjamin, Alberto Bisin, Marco Bonetti, Gary Charness, Lucas Coffman, Vincent Crawford, Carlos Cueva, Martin Dufwenberg, David Eil, Gigi Foster, Andrew Gelman, Nicola Gennaioli, Uri Gneezy, Matt Goldman, Robin Hogarth, Kyle Hyndman, Richard Jagacinski, Elliot Ludvig, Asier Mariscal, Daniel Martin, Daniel Miller, Raymond Nickerson, Daniel Oppenheimer, Raffaella Piccarreta, Giovanni Ponti, Justin Rao, Colin Raymond, Aldo Rustichini, Angel Sanjurjo, Andrei Shleifer, Connan Snider, Joel Sobel, Charles Sprenger, Hal Stern, Neil Stewart, Daniel Stone, Sigrid Suetens, Richard Thaler, Robert Wardrop, Thomas Youle, and Jeffrey Zwiebel. We would also like to thank participants at the SITE Psychology and Economics Workshop at Stanford University (2014), the SABE conference (2014), the North American Econometric Society Meetings (2014), the CAGE Conference, Warwick (2014), the ESA meetings in Honolulu, Santa Cruz and Zurich (2013-2014), and the BEELab conference in Florence (2013), as well as seminar participants at Pompeu Fabra University, Simon Fraser University, Bocconi University and the University of Alicante. All mistakes and omissions remain our own.

§We thank Thomas Gilovich for locating the paper records of his data and scanning them for us, as well as for making us aware of the work of Jagacinski, Newell, and Isaac (1979). We thank Richard Jagacinski for locating his computer punch card print-outs and having them converted to electronic format for us.

¶This manuscript previously circulated under the shorter title "A Cold Shower for the Hot Hand Fallacy."

A particular test can detect only a particular pattern or class of patterns, and complete randomness can therefore only be disproved, not proved. (Houthakker 1961)

1 Introduction

The *hot hand fallacy* refers to a belief in the atypical clustering of successes in sequential outcomes when there is *none*. It was introduced in the seminal study of Gilovich, Vallone, and Tversky (1985, "GVT"), in which they show that while observers, players, and coaches of basketball believe that some players exhibit "hot hand" or "streak" shooting, their analysis of shooting data demonstrates that the hot hand is a "myth" (Tversky and Gilovich 1989a,b).

Owing to the durability of the original conclusions,[1] and the sustained belief in the hot hand by professional practitioners nonetheless,[2] over the last three plus decades the hot hand fallacy has earned the reputation of "a massive and widespread cognitive illusion" (Kahneman 2011). Accordingly, the hot hand fallacy has been given considerable weight as a candidate explanation for various puzzles and behavioral anomalies identified in the domains of financial markets,[3] sports wagering,[4] casino gambling,[5] and lotteries.[6]

The consensus that the hot hand is a cognitive illusion has depended largely on the results of GVT's critical test, a controlled shooting field experiment with collegiate players, which was designed for the purpose of "eliminating the effects of shot selection and defensive pressure" (p. 34) present in game data.[7] In the experiment each of 26 players shot 100 times from a fixed distance,

[1] "Many researchers have been so sure that the original Gilovich results were wrong that they set out to find the hot hand. To date, no one has found it" (Thaler and Sunstein 2008).

[2] Indiana collegiate coach Bob Knight remarked, "There are so many variables involved in shooting the basketball that a paper like this doesn't really mean anything"(Moskowitz and Wertheim 2011, p.229). Legendary Boston Celtics coach Red Auerbach commented, "Who is this guy? So he makes a study, I couldn't care less"(Gilovich 2008, p.17). Amos Tversky remarked on the stubbornness of practitioners to heed the evidence, "I've been in a thousand arguments over this topic. I've won them all, and I've convinced no one"(Moskowitz and Wertheim 2011, p.229).

[3] See, e.g. Barberis and Thaler (2003); De Bondt (1993); De Long, Shleifer, Summers, and Waldmann (1991); Kahneman and Riepe (1998); Loh and Warachka (2012); Malkiel (2011); Rabin and Vayanos (2010)

[4] See, e.g. Arkes (2011); Avery and Chevalier (1999); Brown and Sauer (1993); Camerer (1989); Durham, Hertzel, and Martin (2005); Lee and Smith (2002); Paul and Weinbach (2005); Sinkey and Logan (2013)

[5] See, e.g. Croson and Sundali (2005); Narayanan and Manchanda (2012); Smith, Levere, and Kurtzman (2009); Sundali and Croson (2006); Xu and Harvey (2014)

[6] See, e.g. Guryan and Kearney (2008); Yuan, Sun, and Siu (2014)

[7] In addition to their controlled shooting field experiment GVT also tested for hot hand shooting in NBA game data and free throw shooting, which are vulnerable to challenging statistical issues that are not present in the controlled setting (see Appendix A). In Section 6 we briefly discuss recent analyses in both of these relatively non-controlled environments, which, with improved data and empirical approaches, find evidence of positive serial dependence in shooting performance.

yielding an estimate of the mean hot hand effect (across players) that was statistically indisdinguishable from zero. These results were later replicated with a near-identical design in Avugos, Bar-Eli, Ritov, and Sher (2013), but with Olympian rather than college players, and fewer shots per player.

However, the recent discovery of a surprising statistical bias in a measure of hot hand shooting used in GVT and subsequent studies leads to a reversal of their results (Miller and Sanjurjo 2018). In particular, simply adjusting the original analyses to correct for the bias yields average hot hand effects that are significant and substantial.

In the present study we begin by observing two fundamental limitations endemic in existing controlled (and non-controlled) shooting studies, which are separate from the bias. The first limitation with GVT and the studies that followed is that their tests lack the requisite power to detect the hot hand on the *individual level*.[8] This is crucial because people typically believe in the hot hand of a certain player, not, for example, in the hot hand shooting of an entire team or league. Further, the purported absence of an average hot hand effect across a group of players (Avugos et al. 2013; Gilovich et al. 1985; Koehler and Conley 2003) does not imply that there are no hot hands within the group. Rather, the presumed null effect could simply reflect, for example, a mixture of hot hands and their opposite. Second, previous studies' statistical measures cannot identify hot hand shooting because they do not separate streak shooting in makes from streak shooting in *misses*. For example, the standard measure of hot hand effect size is the difference between shooting percentage when on a streak of hits (makes) vs. when on a streak of misses. Thus, a player who shoots better than usual in both of these cases can be rendered indistinguishable from one who always shoots the same no matter what, and so on.[9]

To address each of these limitations and more accurately measure the extent of hot hand performance in the original controlled test environment, we introduce a set of statistical measures, as well as a test procedure, that increase statistical power, separate hot hand from cold hand shooting,

[8] Jagacinski et al. (1979) is an exception to this. See Appendix B for a power analysis of GVT's statistics vs. ours. The issue of power limitations in GVT's original study has been pointed out by other authors (Albert 1993; Albert and Williamson 2001; Dorsey-Palmateer and Smith 2004; Hooke 1989; Korb and Stillwell 2003; Miyoshi 2000; Stern and Morris 1993; Stone 2012; Swartz 1990; Wardrop 1999), with Wardrop (1999) being the only attempt to address it (see Footnote 11)

[9] The bias discovered in Miller and Sanjurjo (2018) is a separate issue that leads to the underestimation of performance when a shooter is on a streak of makes and the overestimation of performance when a shooter is on a streak of misses.

and are unbiased.[10] In addition, we conduct a controlled shooting field experiment (with expert shooters) that improves on the previous designs in terms of control, sample size, and out-of-sample prediction of hot hand performance, as it collects multiple shooting sessions (per player) across a six month period. This allows us to perform the first well-powered individual level tests of hot hand shooting. Further, we obtain and analyze the rich dataset of Jagacinski et al. (1979), which also allows for individual level testing, and was previously uncited in the hot hand literature. We conduct our improved statistical analysis on each of the three extant controlled shooting datasets, and again after pooling the three datasets together.[11]

Finally, in addition to our analysis of hot hand performance, we perform a novel investigation of hot hand beliefs. In particular, in our experiment we record expert beliefs on the degree of hot hand shooting in each of their teammates, which allows for out-of-sample tests of beliefs on performance.

We find strong evidence of large hot hand effects on the individual level in each of the two datasets that are sufficiently powered for such testing. This is the first evidence of its kind. Further, we find persistent individual hot hands across sessions in both of these datasets. This is also the first evidence of its kind, and accords with the hot hand beliefs that players hold about certain individuals. In addition, we observe that the degree of hot hand shooting varies considerably across shooters, which suggests that decision makers (e.g. teammates and coaches) may have incentive to correctly predict which shooters have a greater tendency to become hot. Accordingly, we find strong evidence that expert players' beliefs successfully predict (out-of-sample) which of their teammates have more (or less) of a tendency to become hot. This is the first evidence of its kind on beliefs. Finally, the considerable variation in the magnitude of the hot hand effect (which is sometimes even negative) that we observe across players implies that a pooled analysis could conceal the hot hand effects observed in our individual-level analysis. Nevertheless, while not directly relevant for

[10] The test procedure that we outline in Section B.1 and apply in Sections 4 and 5 was not practically feasible at the time GVT conducted their analysis, given limitations in computational power.

[11] Aside from the two controlled shooting datasets that we were able to gather, in addition to our own, the only other controlled shooting experiment that has been cited in the hot hand literature is Avugos et al. (2013). While the authors declined to make their data available to us, a proper bias correction can be applied directly to the data provided in the tables of their paper, which reveals a point estimate for the hot hand effect that is substantial(Miller and Sanjurjo 2018). Wardrop (1999) provides a case study involving a single shooter, but after personal communication with the shooter who conducted the informal study herself—sometimes rebounding her own shots and recording her own shot outcomes, and other times having someone else do both for her—we viewed it as not having sufficient control to be included in our analysis.

the hot hand beliefs that pertain to individual shooters, we surprisingly find evidence of hot hand shooting even on the aggregate level in each of the three datasets, as well as when pooling across the datasets.

Thus, the present results suggest that the hot hand shooting that Miller and Sanjurjo (2018) find in the data from GVT's critical controlled shooting experiment is not anomolous, but instead robust to variations in the experimental environment, and persistent across time. In terms of methodology, our statistical approach allows for exact tests of hot hand (and cold hand) performance, and is in no way restricted to basketball data. It can be used to perform either a stand alone analysis, or to complement other approaches such as regression analyses. Given the findings on hot hand performance it becomes natural for the literature to now shift more toward investigating the nature of hot hand beliefs, and their fitness. We provide such a test of beliefs in the present study, and discuss others in Section 6.

Section 2 explains the design of our field experiment, as well as those of the other two datasets we analyze. Section 3 explains our statistical approach. Section 4 reports our results on hot hand performance, Section 5 our results on hot hand beliefs, and Section 6 further discusses the implications of our results.

2 Design

We first describe the designs of three controlled shooting field experiments that generate the data we analyze in Sections 4 and 5, then discuss the novel features of our design with respect to previous work.

2.1 Our controlled shooting field experiment

Our design consists of two phases, conducted six months apart: Phase One tests whether any individual shooter in our sample has the hot hand, as well as whether the hot hand happens to be an average effect in our pooled sample of shooters (despite potential heterogeneity across shooters). Phase Two tests whether the hot hand effect can be predicted out of sample. To this end we had players from Phase One return for multiple additional sessions, to see if those with the hot hand in Phase One also have the hot hand in Phase Two, as well as whether any average hot hand effect

in Phase One would re-occur.

Setting and participants

We recruited players from the semi-professional basketball team Santo Domingo de Betanzos, in the Spanish province of Galicia, by dropping by at the end of a team practice and inviting all players to participate in a scientific study of basketball shooting with financial incentives.[12] While player interest was unanimous, it was not possible to accommodate all players given their limited time availability and our limited set of available time-slots. In total, eight players were able to participate in both phases of our panel.[13] The players averaged 24 years of age, and 14 years of experience playing in competitive, organized, basketball leagues. The experiment was conducted on their home court, the Pabellón Municipal Polideportivo de Betanzos, where they both practice and host opposing teams in league games. All shooting sessions were video-recorded.

Design of the shooting session

Upon arrival at the scheduled time the shooter (subject) was given several minutes to warm up by shooting however he liked. The experimenter observed the shooter in order to gauge from what distance he would make around 50 percent of his shots (in order to maximize the variance of shot outcomes for statistical testing purposes). The experimenter then used a strip of masking tape to mark the shooting location from which that player would take all 300 shots.[14] Next, the shooter was led to a closed room, where the experimenter read the instructions aloud as the shooter read silently (see Appendix C.1 for the shooter's instructions). The shooter was informed that he would be taking 300 shots, and that in addition to a 5 Euro participation fee, 10 of these shots would be selected at random to determine his payoffs. For the 10 randomly selected shots, he would receive 6 Euros for each shot that he hit and 0 Euros for each shot that he missed. He was also informed

[12] In Spain there are five rated categories of organized, competitive basketball. The top level is the ACB, in which the top earners make millions of euros each year. The second level is also professional, in the sense that players make enough money not to need other forms of employment to live comfortably. Levels three through five are considered semi-professional in the sense that while players have all of their basketball-related expenses paid for them, and may earn some take-home earnings on top of this, it is typically not enough to live comfortably on without additional employment. Santo Domingo de Betanzos was the best team in the 5th category the year that we conducted the experiment, for their region, so was invited to move up to the 4th category the following year.

[13] One of these players had recently left Santo Domingo to play professionally in the 2nd category (see previous footnote), but continued to train frequently with Santo Domingo.

[14] The shooting location was kept constant for the purpose of controlling a player's probability of hitting a shot. While the distance from the rim selected for each shooter varied, all selected locations were straight in front of the rim, meaning that they were situated on the imaginary axis which bisects the width of the court.

that the 10 shots had already been selected, printed on a sheet of paper, and sealed in an envelope. The envelope was shown to the shooter and left in his field of vision for the duration of the session. Upon completing the instructions the shooter was given an opportunity to ask questions before returning to the court, where he was then allowed two minutes of practice shots from the marked location before beginning the paid task.

After the shooter had an opportunity to warm up, each of the 300 shot attempts went as follows: a trained rebounder, who was unaware of the purpose of the study, held the ball from a fixed location near the basket, which was also marked on the floor.[15] When the experimenter initiated a computer-generated tone, the rebounder passed the ball to the shooter in a precisely prescribed way (see Appendix C). Once the shooter received the ball, the rebounder turned his back to the shooter and the shooter was allowed to choose the timing of his shot without constraint, though the shooters typically shot within 1-2 seconds after receiving the ball. After the shot, the rebounder collected the ball, returned to his marked location as quickly as possible, and awaited the same computer-generated tone to signal the beginning of the next shot attempt, which occurred, on average, 7 seconds after the previous tone. The task continued in this way, with the experimenter calling out after each block of 50 shots was completed. The duration of each 300 shot session was approximately 35 minutes, which was calibrated to avoid fatigue effects.[16]

Phase One and Phase Two

In Phase One each of ten shooters participated in a single session consisting of 300 shots. In Phase Two, six months later, each shooter participated in multiple additional shooting sessions. Eight of the ten Phase One shooters were available to participate in Phase Two; we refer to these eight shooters as the *panel*.[17] At the conclusion of Phase One subjects filled out a short questionnaire. Then, following Phase Two they filled out a more detailed questionnaire, which we discuss in

[15] The rebounder was informed only of his own task, and that the shooter would shoot 300 times.

[16] In order to minimize the possibility of a potential fatigue effect from over-exertion we settled on 300 shots after running pilots with ex-basketball players that were not in basketball shape. These players reported no problem shooting 300 times under the conditions of our design (with a rebounder). It is safe to say that less than one quarter of each session was spent in physical movement for the shooters. In a post-experiment questionnaire our subjects reported below average levels of fatigue. Players commented that they shoot around the same number of shots on a daily or near daily basis. In Section 4.1 we find that there is no evidence of fatigue effects (or warm-up effects) in our data.

[17] One of the two shooters that did not participate in Phase Two canceled at the last minute (the other was out of the country), so three additional players who were aware of the experiment and eager to participate contacted us and were given the canceled time slots. While these players were not part of our two-phase design, we nevertheless include their shots on our pooled analysis in Section 4.3.

Section 5.

Before Phase Two, we conducted a statistical analysis of the shooting data from Phase One (see Section 4.2), which identified one of our shooters, "RC," as the hottest shooter in the sample. Further, in questionnaire responses, his teammates—who had not observed RC's shooting session, but averaged 800 hours of previous playing experience with him—ranked RC as by far the hottest shooter of the group. On this basis, in Phase Two we allocated a larger number of our scarce session timeslots to RC (5) than to the other shooters in the panel (3 each), in order to maximize the power of our test of whether RC had the hot hand, and in order to observe if the effect is persistent across many sessions. Phase Two sessions were conducted using a design and protocol identical to those of Phase One.

2.2 Jagacinski et al. (1979) controlled shooting field experiment

Each of 6 collegiate-level players (one had been an all-American) shot in nine sessions of 60 shots each. Each player took all shots from a single location, at a distance from which the experimenters thought he would make around 55 percent of his shots. Players were paid \$.05 for each made shot.[18]

2.3 Gilovich et al. (1985) controlled shooting field experiment

Each of 26 players from the Cornell University Mens' (14) and Womens' (12) basketball teams participated in one session of 100 shots. For each player, the experimenters drew two symmetric arcs, one facing the left side of the basket and one facing the right, at a distance from which they thought the shooter would make around 50 percent of his/her shots. The player had to change location behind the arc after each shot was taken, and 50 shots were taken from behind each arc. Before every shot the player bet on whether the upcoming shot would be a make or a miss. If she bet "high" she would win \$.05 for a hit and lose \$.04 for a miss. If she instead bet "low" she would win \$.02 for a hit and lose \$.01 for a miss. Each player was paired with one other and bet on both her own shots and those of the other. Total payoffs were equal to \$2 plus or minus the amount of money won or lost in the betting tasks.

[18] The subjects shot under three different conditions: "On," "Off," and "Int." We present an analysis of the "On" condition only, as the "Off" and "Int" conditions involved turning the gym's lights off immediately after the player released the ball, and thus were not comparable to other controlled shooting studies (the Jagacinski et al. (1979) study was designed to investigate the role of movement and ball trajectory information in the prediction of shooting performance).

2.4 Discussion of designs

Our controlled shooting design improves on GVT's in several ways: (i) our shooters are strategically isolated; they only shoot, whereas GVT's were forced to bet before each shot, which allowed for possible interaction effects between shooting and betting,[19] (ii) our shooters have constant, non-negligible performance incentives, whereas GVT's players received earnings that accumulated over time (wealth effects), and changed in accordance to their bets, (iii) our shooters always shoot from the same location, whereas GVT's were forced to move after each shot, which enables us to control for the possibility that a player's hit probability varies solely due to shot location,[20] (iv) for each shooter we collect 300 shots per session, rather than GVT's 100, which gives our tests substantially more statistical power (see Appendix B.3), (v) we are able to collect multiple sessions of 300 shots from our shooters, across a six month period, in contrast to GVT's single session for each subject. This distinguishes our design by allowing us not only to test for the existence of hot hand shooting on the individual level, but also whether hot hand performance can be successfully predicted out of sample.

When compared to the design of Jagacinski et al. (1979), ours improves upon it by: (i) controlling for wealth effects and increasing incentives, (ii) improving statistical power by collecting 300 shots per session rather than 90, and (iii) spreading sessions for each shooter across a six month period in order to test whether hot hand shooting is persistent across time.

[19] For example, by forcing a player to predict his own performance before each shot, a separate task he may not be used to, one may divide his attention (Kahneman 1973) or disrupt his flow or rhythm (Csikszentmihalyi 1988), arguably making it less likely that hot hand performance emerges.

[20] Shooters have different probabilities of success at different locations (distance and shot angle). Thus, having players shoot from multiple locations would require one to control for their probability of success at each location. Fixing the shot location increases the power of the design for a given sample size. One might argue that if a player shoots from the same location, he may be able to use the outcome of the previous shot to calibrate the next shot, which could artificially induce serial correlation in a way that would not be reproducible in a game setting (aside from free throw shooting). This concern should be lessened by the fact that, using the empirical strategy we outline in Section 3, we find evidence of a hot hand effect in the original study of GVT, where the shot location changed after each shot. Further, in our design, if a shooter could calibrate, we would expect to see him improve over time, rather than remain at the target 50 percent hit rate for the task, but this is not what we find. In particular, in Section 4.1 we find no evidence of improvement (although we find that players shoot significantly worse in their first 3 shots, both in our study and all extant studies).

3 Empirical Strategy

3.1 Defining and detecting the hot hand

While the verbal descriptions that players and fans give for the hot hand are not always easy to formalize, all convey the notion that when a player has the hot hand his or her probability of success (i.e. shooting ability) temporarily elevates above baseline.[21] Thus, regardless of the underlying mechanism, periods of hot hand shooting should be expected to translate into a tendency for hits to cluster.[22]

We define three statistics that measure hit clustering patterns typically associated with hot hand shooting. Loosely stated, our statistics measure: (i) how often a player is on a "hot" streak,[23] (ii) a player's shooting percentage conditional on having a recent streak of success,[24] and (iii) the length of a player's most exceptional hot streak.[25] In addition, we employ a commonly used measure of first-order serial dependence, the number of runs. For each statistical measure, we test for the hot hand by comparing its realization to the sampling distribution under the null hypothesis that a shooter has a fixed (i.i.d.) probability of a hit.

Definitions

Let S be the set of shot indices and $\{x_s\}_{s\in S}$ the sequence of shot outcomes, where $x_s = 1$ if shot s is a hit, and $x_s = 0$ if it is a miss.

Our first statistic, the *hit streak frequency statistic*, is a measure of how often a player is on a "hot" streak; it is defined as the (relative) frequency of shots that immediately follow a streak of consecutive hits:

[21] Gilovich et al. (1985) report: "These phrases express a belief that the performance of a player during a particular period is significantly better than expected on the basis of the player's overall record."

[22] Candidate mechanisms include: (i) positive feedback following recent success (due to, for example, positive feedback leading to increased confidence [Bandura 1982]), or (ii) spontaneous shifts in performance state (due to, for example, internal or external factors leading to increased focus, attention, or motor control [Churchland, Afshar, and Shenoy 2006; Csikszentmihalyi 1988; Kahneman 1973]).

[23] While Gilovich et al. (1985) do not test for this pattern, they do mention that "the hot hand" and "streak shooting" should imply that "the number of streaks of successive hits or misses should exceed the number produced by a chance process with a constant hit rate."

[24] Gilovich et al. (1985) note that a player "who has better chance of hitting a basket after one or more successful shots than after one or more misses" can be said to have a "hot hand" or be described as a streak shooter

[25] Gilovich et al. (1985) note that "A player who produces longer sequences of hits than those produced by tossing a coin can be said to have a 'hot hand' or be described as a 'streak shooter.'

$$H_F := \frac{\#S_H}{\#S}$$

where $\#S$ is the number of shots, and $\#S_H$ is the number of shots that immediately follow a streak of consecutive hits, i.e. $S_H := \{s \in S : \prod_{t=s-\ell}^{s-1} x_t = 1 \text{ for some } \ell \geq k\}$, where k is the threshold that defines whether a block of ℓ consecutive hits, $\prod_{t=s-\ell}^{s-1} x_t = 1$, constitutes a streak.[26,27]

Our second statistic, the *hit streak momentum statistic*, is a measure of a player's probability of success conditional on having a recent streak of success; it is defined as the player's shooting percentage immediately following a streak of hits:[28]

$$H_M := \frac{\sum_{s \in S_H} x_s}{\#S_H} \quad \text{if } \#S_H > 0 \text{ (otherwise undefined)}$$

Computing the hit streak frequency and momentum statistics requires the determination of a threshold past which consecutive hits are defined to be a streak. While higher thresholds provide a better signal of hot hand shooting than lower thresholds,[29] they result in smaller sample sizes. With this trade-off in mind, in the main analysis below we adopt the threshold of $k = 3$, as in our experimental design this allows us a sample size ($\#S_H$) of more than 30 shots per session (given our target of 50 percent hits).[30] This threshold is also commonly used in previous studies (Avugos et al. 2013; Gilovich et al. 1985; Koehler and Conley 2003; Rao 2009b), and happens to coincide with the threshold that human observers use to categorize consecutive outcomes as a streak (Carlson and

[26] Because it is impossible for the first few shots to immediately follow a hit streak, in the results section we report $H_F := \frac{\#S_H}{\#S-k}$, where k is the threshold used to define a streak.

[27] While the calculation of $\#S_H$ involves overlapping shot windows, it can be shown to have a normal asymptotic distribution (see Appendix B).

[28] Here, $\#H_M$ is undefined if $\#S_H = 0$. Because H_M is equal to the ratio of two asymptotically normal random variables (see Appendix B), there is no guarantee it will be normal. Nevertheless, in simulations its distribution is relatively symmetric for 300 shot sequences.

[29] To illustrate how a higher threshold provides a better signal of the hot hand, suppose that a player's hit rate is .7 in the "hot" state, .3 in the "normal" state, and that the player is in the hot state on 10 percent of her shots. The likelihood of her hitting k in a row is $(.7/.3)^\ell$ times higher when she is in the hot state. Thus, upon observing ℓ hits in a row, the odds in favor of the player being in the hot state must increase by this factor. With prior odds $1:9$ in favor of being in the hot state, the posterior odds in favor the hot state for $\ell = 1, 2, 3$ are approximately $2:9$, $5:9$, and $13:9$ respectively.

[30] In any case, when applying the measures of the hot hand effect used in previous studies, we find a significant and substantial estimate of the hot hand effect that is robust to the threshold ($k = 2, 3, 4$). For the statistics introduced in this section, the results presented in Section 4 also hold if the threshold for a streak is set at $k = 4$ successive hits. For $k \geq 5$, the tests are underpowered, because fewer than 9 observations (in 300 shots) are expected from a 50 percent shooter. For $k \leq 2$, the results are mixed, as may be expected because a miss followed by a hit, or a miss followed by two hits is not as diagnostic of hot hand shooting as a streak of three or more hits.

Shu 2007).

Our third statistic, the *hit streak length statistic*, measures the length of a player's most exceptional hit streak, i.e. the length of the longest *run* of hits, where a run is a subsequence of consecutive hits flanked by either misses or the start or end of the sequence.[31,32]

As we will be testing several statistics, we address the issue of multiple comparisons with a composite statistic H equal to the first principal component of the three hit streak statistics.[33] This approach yields greater statistical power than conservative corrections for multiple comparisons (such as Bonferroni), but without compromising on the type I error rate, as it accounts for the correlations between the three hit streak statistics.[34]

In addition to the three hit streak statistics, we consider a statistic that appears prominently in the hot hand literature (Gilovich et al. 1985; Koehler and Conley 2003; Wardrop 1999), the *runs statistic*, R, which is equal to the total number of runs of hits and misses together.[35] This statistic is a measure of first order serial dependence.[36]

The exact distribution of each of these test statistics, under the null hypothesis that a player has a fixed probability of success, can be approximated to arbitrary precision with a Monte-Carlo permutation test. We describe this procedure in detail in Appendix B.1.

3.2 Discussion of our Empirical Strategy

Previous controlled shooting studies typically consider three types of tests: (i) a "conditional probability test," which tests whether players' shooting percentages differ on the shots taken immediately following a streak of k or more hits vs. the shots taken immediately following a streak of k or more

[31] For example, the sequence of shot outcomes 0011010111 has $\#S = 10$ shots and three runs of hits: 11, 1, and 111.

[32] The distribution of H_L can be approximated using a normal distribution (see Appendix B). In our statistical tests below we instead numerically approximate (to arbitrary precision) the exact distribution of H_L, as the convergence to a symmetric normal distribution is slow.

[33] We compute H from the joint distribution of the statistics for a given player's data (as generated by our permutation scheme, which is defined below). The first eigenvector from the decomposition weights each statistic nearly equally, though the hit streak length statistic is weighted a bit less.

[34] The higher the player's hit rate, the higher the value taken by each statistic. To illustrate, even a player with a fixed probability of success yields an average pairwise correlation between the statistics of around .5 (using the joint distribution under the null, as generated by the permutation method outlined below).

[35] For example, the sequence of shot outcomes 0011010111 has $\#S = 10$ shots, and six runs total, consisting of three runs of hits: 11, 1, and 111, and three runs of misses: 00,0, and 0.

[36] In particular, the number of runs in a sequence of binary outcomes is equal to the number of alternations (pairs of consecutive shots with different outcomes) plus 1, i.e. $R := 1 + \sum_{s=1}^{\#S-1}[1 - x_s x_{s+1} - (1 - x_s)(1 - x_{s+1})]$.

misses, for $k = 1, 2, 3$,[37] (ii) first order serial correlation, ρ, and (iii) the number of runs, R. The first test provides the only hot hand effect size measure. The second two tests are redundant with the $k = 1$ version of the first test, as they all measure the same pattern in shooting (see Appendix B.2 for a demonstration).[38]

The statistics that we propose offer several advantages over those of previous studies: (i) because each hit streak statistic has a symmetrically defined miss streak statistic, we can separate hot hand shooting from cold hand shooting, whereas the statistics used in GVT and replications detect streakiness in hits *or* misses, so cannot,[39] (ii) the hit streak momentum statistic (H_M), by benchmarking a player's shooting percentage immediately following a streak of hits to its distribution under the null, rather than comparing it to the shooting percentage immediately following a streak of misses, eliminates the severe bias present in the conditional probability test of GVT and replications (see Miller and Sanjurjo [2018]), (iii) we can compare our hit streak momentum statistic (H_M) to a shooter's performance following all other recent shot histories, which allows us to use all of a player's shots, thereby offering greater power than the conditional probability tests employed in GVT and replications, which use only around 25-30 percent of the shots taken in each shooting session when $k = 3$ (see Appendix B.3),[40] and (iv) the hit streak length (H_L) and frequency (H_F) statistics can detect hot hands in certain patterns of shooting behavior that could otherwise go undetected by first order serial correlation, and other measures, e.g. when hit streaks exhibit persistence and miss streaks exhibit reversal.[41]

Our empirical approach further improves statistical power relative to previous studies by: (i) collecting between nine and eighteen times as much data per player as the original GVT study, which in turn had more data than its replications, and has been noted for lacking the requisite power

[37] In particular, a paired t-test is conducted in which each player's two shooting percentages are compared to each other.

[38] In addition to these three tests (Gilovich et al. 1985) also consider the variation in the hit rate in four-shot windows, which is substantially more underpowered relative to the other tests (Albert 1993; Albert and Williamson 2001; Dorsey-Palmateer and Smith 2004; Hooke 1989; Korb and Stillwell 2003; Miyoshi 2000; Stern and Morris 1993; Swartz 1990; Wardrop 1999).

[39] While Avugos et al. (2013) is the only strict replication of GVT's controlled shooting study, for simplicity of exposition we refer to Koehler and Conley (2003)'s finding of no hot hand shooting in the less controlled NBA three point contest also as a replication of GVT.

[40] The increase in power relative to GVT is typically around 10-20 percentage points. Avugos et al. [2013] use the same test as GVT, but with only 40 shots per player, and Koehler and Conley [2003] performed a similar analysis, with a median of 49 shots per player.

[41] Another example arises when extended hit streaks and excessive alternations between hits and misses both occur in the same sequence. We illustrate this case in our power analysis (see Appendix B.3).

to detect a variety of alternative hot hand hypothesis models (see Footnote 8, and Appendix B.3 for a discussion of power), and (ii) performing pooled analyses across sessions, players, and even different studies, whereas GVT and replications can only perform (under-powered) tests on the individual shooter/session level.[42]

Finally, to the extent that the hot hand exists, its relevance for decision making depends largely on the magnitude of its effect. While we can measure standard deviations from the mean of hit streak statistics under the null, the hit streak momentum statistic (H_M) also allows us a direct estimate of magnitude, albeit a conservative one. In particular, H_M measures the shooter's hit rate in a potentially "hot" state (immediately following three or more hits in a row), which can then be compared against the shooter's hit rate in a presumably non-hot state (immediately following any other recent shot history). This comparison is conservative because: (i) it will lead to a downward biased estimate of the hot hand effect for the same reason that GVT's conditional probability test is biased, though the bias is relatively smaller under our experimental design (-2 percentage points for a 300 shot session), and (ii) the true effect size of the hot hand, should it exist, will be larger than the estimated effect size because not every shot taken immediately following three or more hits in a row occurs while a player is in a hot state (classification/measurement error), which leads to an attenuation bias in the estimate (see Appendix B.3; also see Arkes 2013, Green and Zwiebel 2017, and Stone 2012 for a discussion in relation to previous work).[43]

4 Results: Shooting

In Section 4.1 we provide a brief summary of overall performance in our panel of shooters. In Section 4.2 we test whether the hot hand effect exists on the level that hot hand beliefs typically pertain to—that of the individual shooter—and whether it re-occurs within individuals across time. In addition, in Section 4.3 we examine whether, despite the variation in the size and sign of the hot hand effect across shooters, the hot hand is also a property of the average shooter.

[42] We can pool all of a player's shooting data, compute the hit streak statistic for each session, standardize it by subtracting the mean and dividing by the standard deviation for that session, then calculate the average of the standardized statistics across sessions, and then generate the distribution of this average by permuting shots within session strata, to assure that the results are not being driven by good day/bad day effects. For pooling across players we average across these player averages, permuting shots within all strata defined by each player and session.

[43] As an example, Stone [2012] has demonstrated that a player can simultaneously have a large serial correlation in shooting ability and a sample serial correlation that is, asymptotically, orders of magnitudes smaller (e.g. .4 and .06, respectively).

4.1 Overall performance

The average shooting percentage across players in the (balanced) two-phase panel (3 sessions) was 50.08 percent, with a 7.7 percent standard deviation (our design target was 50 percent). By allowing the shooters to warm up before each session, and setting the length of the shooting session to 300 shots, the evidence indicates that players were not subject to warm-up or fatigue effects.[44] In particular, in a post-experiment questionnaire the average reported level of experienced fatigue by shooters was less than 5, on a scale of 1 to 10. Moreover, whereas fatigue might be expected to cause a decline in shooting performance towards the end of a session, we find no evidence of this, as in the first 150 shots players shoot at 49.7 percent, and in the second 150 shots 50.4 percent.[45]

4.2 Identification of the hot hand at the individual level

We use the empirical strategy outlined in Section 3 to test for hot hand shooting on the individual level in each of the three controlled shooting datasets described in Section 2. Notice that because the hot hand fallacy is an extreme notion, by which people believe in the hot hand although it is a "myth" [Tversky and Gilovich 1989a,b], evidence of persistent hot hand shooting in just one player—among the possibly many they observe—would be enough to conclude that their belief in the hot hand is not fallacious. The large number of sessions, and shots per session, for each individual in our experiment and that of Jagacinski et al. [1979] allow for well-powered individual level testing for hot hand shooting, and its possible persistence. By contrast, because the GVT experimental design is underpowered, with its single session per individual (see Section 3), the individual level analysis we perform on GVT's data is merely suggestive.

Evidence from our study

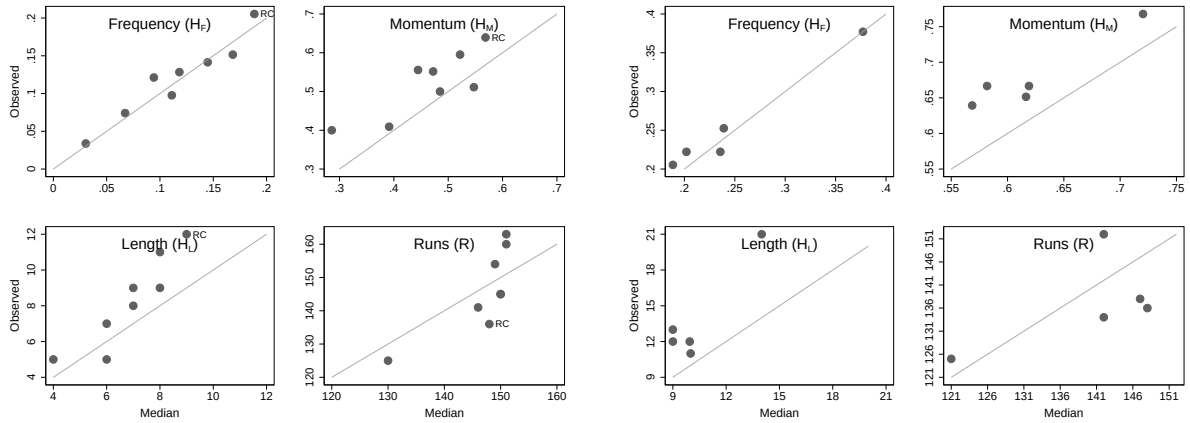
Figure 1a reports the hit streak statistics from Phase One of our panel, in which each of eight shooters performed a single 300 shot session. In each cell, each shooter's corresponding hit streak statistic is plotted against its median under the null (based on the number of hits in his 300 shot attempts). One shooter, whom we refer to as *RC*, and whose statistics are labeled in the

[44] The one minor exception is that for the first three shots there did appear to be a warm-up effect: players shoot significantly worse (37 percent) than the remainder of the session. This is also true in the data of GVT and Jagacinski et al. [1979].

[45] In addition, the average of the session-level differences for each of the three possible comparisons was not significant. If we instead divide the sessions into three sets of 100 shots, the averages were 50.5, 49.8, and 50.0 percent, respectively.

Figure 1: *Scatter of the three hit streak statistics, and number of runs, with observed value vs. median value under the null distribution for the session.*

(a) *Panel, Phase One (8 shooters, 1 session each)* **(b)** *Shooter "RC" (five sessions)*



figure, stood out in Phase One because he had the most extreme hit streak statistics among the shooters, and because his statistics were significant in the single session.[46] Further, in a multi-item questionnaire administered to the team, RC was identified by his teammates as the player with the greatest potential for the hot hand, based on their previous playing experience with him.[47] Thus, when we followed up with the players in our panel six months later, it became possible to use RC's Phase Two shooting data to test if teammate perception of in-game hot hand performance, as well as hot hand performance in the shooting study six months prior, could predict hot hand performance (out of sample). To maximize the identifying power of our test, without informing RC of the purpose, at the end of Phase Two we solicited more sessions from him than from the other players.

Each cell of Figure 1b plots one of RC's hit streak statistics, across each of five sessions, against its median under the null (based on the overall number of hits in that session).[48] RC's hit streak length (H_L) and momentum statistics (H_M) are greater than their respective medians under the

[46] RC's Phase One composite hit streak statistic, H, had a p-value of .08.

[47] The teammates were not aware of the goal of the study, did not witness RC shoot in his sessions, and were not informed of his performance.

[48] Due to a time conflict, in one instance RC requested to shoot two of his sessions successively in a single day. While we granted his request, we exclude this session from our analysis because it is conducted under conditions not strictly identical to those of his other sessions. If we instead include this "double session" in the pooled analysis of RC's sessions the significance of his statistics do not change. See Figure 9 in Appendix D for a graph similar to that in Figure 1b, but with this data included (five other shooters in our panel were also granted their request to shoot two successive sessions in a single day, and in these cases we followed the same procedure as with RC).

Table 1: *Average values for each of the three hit streak statistics, and number of runs, for the shooter RC from our study, and the shooter JNI6 from Jagacinski et al. (1979), with p-values in parentheses.*

	Shooter RC			Shooter JNI6
	session 1	sessions 2-5	All	9 sessions
Hit Streak Frequency (H_F)	.21	.27	.26	.27***
	(.197)	(.335)	(.134)	(.0011)
Hit Streak Momentum (H_M)	.64	.69***	.68***	.71***
	(.105)	(.009)	(.003)	(.0001)
Hit Streak Length (H_L)	12.00	14.25**	13.50**	9.56***
	(.132)	(.037)	(.019)	(.0005)
Total Runs (R)	136.00*	137.25	137.00	26.67**
	(.090)	(.442)	(.168)	(.0262)

Permutation test (50,000 permutations, with session strata)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ (one-sided)

null (gray lines) in every session; this would occur in only around three out of every hundred studies, for each statistic, if the null hypothesis were true ($p = .031$, binomial test).[49]

By pooling the shots from all of RC's sessions, we can perform a more powerful test of the hot hand effect. The third column of Table 1 reports each of RC's statistics, averaged across all five of his shooting sessions, with corresponding p-values in parentheses.[50] All of RC's statistics are in the direction predicted by the 'hot hand' hypothesis, and consistent with the results of the binomial test: the hit streak length statistic is significantly larger, by 2.7 shots, than its mean under the null ($H_L = 13.5$, $p = .019$; permutation test, session strata) and the hit streak momentum statistic is significantly larger, by +6 percentage points, than its mean under the null ($H_M = .68$, $p < .01$), while the frequency and runs statistics are in the predicted direction, but not significant ($p < .20$ for each). Further, the composite hit streak statistic, H, which controls for multiple comparisons, is highly significant ($p < .01$).

While this is the first study to clearly identify the hot hand at the individual level, an important question is whether the effect size is also non-negligible in size. To answer this question, Figure 2

[49] The hit streak frequency (H_F) and runs (R) statistics are not significant ($p = .50$, binomial test). As discussed in Section B.1, the runs statistic is essentially equivalent to serial correlation for binary data. It is not significant for RC because his shot outcomes exhibit persistence after streaks of hits and reversals after streaks of misses.

[50] The p-value of each statistic comes from a permutation test of the sum of each statistic across sessions (stratified at the session level). Each reported p-value is an approximation of the exact p-value under the exchangeability hypothesis. We do not report Clopper-Pearson binomial confidence intervals for the p-values because with 50,000 permutations, the intervals have a width of less than .001 for most statistics, and a width less than .01 for all.

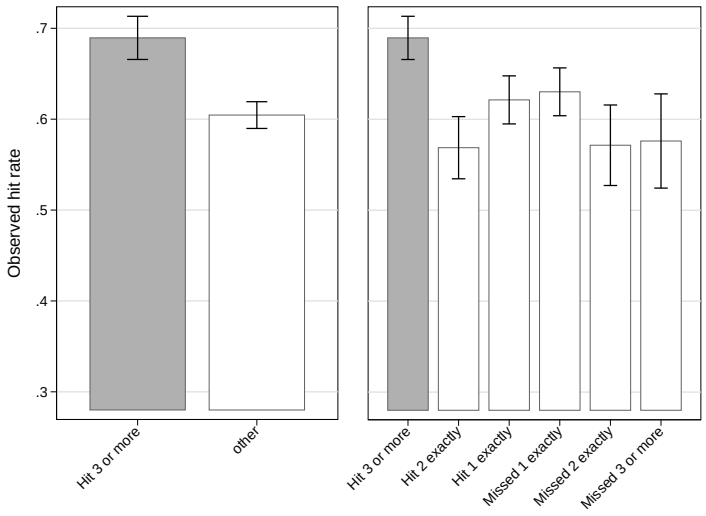


Figure 2: *The shooter RC's hit rate, immediately following 3 or more hits in a row, is higher than after any other recent shot history (with standard errors).*

compares RC's hit rate immediately following a streak of hits with his hit rate immediately following other recent shot histories (with standard error bars). The left panel shows that RC's hit rate increases substantially, by around +9 percentage points, immediately following three or more hits in a row, as compared to any other recent shot history ($p < .01$, two-sample test of proportions).[51,52] To put this effect size into perspective, the difference between the median and the very best NBA three point shooter in the 2015-16 season was +12 percentage points.[53]

Because this effect can be driven by a hot hand or a cold hand (an increased miss percentage immediately following a streak of misses), in the right panel of Figure 2 we categorize shots that do not immediately follow a streak of three or more hits into five mutually exclusive, and exhaustive, recent shot histories: hit the previous two shots but no more, hit the previous shot but no more, missed the previous shot but no more, missed the previous two shots but no more, and missed the previous three or more shots. We observe that RC's conditional hit rate immediately following a run of at least three hits is significantly larger than his conditional hit rate immediately following

[51] Correcting for the downward bias mentioned in Section 3, RC's difference is larger than +10 percentage points.

[52] The same results hold if one defines a hit streak as beginning at four hits in a row. A benefit of the test reported here is that it includes all of a shooter's data, unlike GVT's conditional probability test.

[53] ESPN, "NBA Player 3-Point Shooting Statistics - 2015-16." http://www.espn.com/nba/statistics/player/_/stat/3-points [accessed September 24, 2016].

each of the other five recent shot histories.[54] This indicates that the overall contrast observed in the left panel is driven by the hot hand and not the cold hand. We can also test if RC's *miss streak* statistics, which are symmetrically defined to his hit streak statistics, are significant; they are not, corroborating that RC has the hot hand, and not the cold hand.[55]

The analysis of RC's shooting data demonstrates that an individual can have a substantial hot hand effect that not only systematically re-occurs across time, but can also be correctly predicted—either on the basis of observed performance in previous shooting sessions, or teammate perception of the shooter's in-game performance (see Section 5).

A further question of interest is the extent to which there is evidence that other individuals in our panel have the hot hand. Though we have found that the hot hand exists, there is no reason to expect to see hot hand effects from each player in our panel; the subjects were not selected on the basis of shooting ability, or for a reputation of streak shooting, but rather solely on the basis of availability (they were all from the same team). Nevertheless, Figure 1a shows that in Phase One the hit streak length and momentum statistics are each above the median under the null for 7 out of 8 shooters (p=.035 for each, binomial test). Figure 3a presents a similar plot of the hit streak statistics for each player in the panel, instead with each statistic averaged across the three sessions conducted under identical conditions; the hit streak frequency and momentum statistics, as well as the runs statistic, are on the predicted side of the median for 7 out of 8 shooters (p=.035 for each, binomial test), while the hit streak length statistic is above median levels for 6 out of 8 shooters (p=.145, binomial test). In Table 3 of Appendix D we report the percentage of each player's shots that are hits, both immediately following three or more hits in a row, and after all other recent shot histories.

Evidence from Jagacinski, Newell, and Isaac [1979]

Jagacinski, Newell, and Isaac [1979, "JNI"] provides the richest dataset, after our own, for individual level testing. We focus here on one of JNI's six subjects, *JNI6*, because the standard metric for overall hot hand performance is substantial, significant, and robust to multiple comparisons.

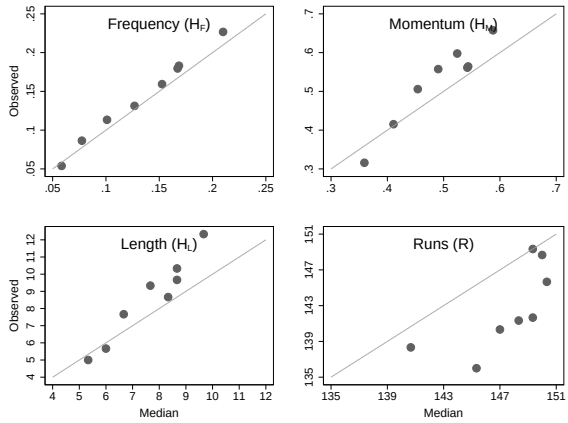
We first report JNI6's streak statistics as we did with RC. Figure 4a reports that in all nine of

[54] The respective p-values for the two-sample test of proportions are: .019, .008, .047, .027, and .002

[55] As a robustness check for possible session effects, in Appendix D.1 we estimate a linear model of RC's hit probability with session fixed effects (which downward bias the estimated effect size by 1.2-2.5 percentage points). The results of the proportion tests reported in Figure 2 are corroborated, again suggesting a hot hand, and not a cold hand, effect.

Figure 3: *The panel of eight shooters (three sessions per shooter).*

(a) *In each cell, each shooter's statistic (averaged across sessions) is plotted against its respective median under the null.*



(b) *Average hit streak statistics for each shooter in Phase 1, Phase 2, and Overall across both phases (p-values in parentheses).*

	Overall	Phase 1	Phase 2
Hit Streak Frequency (H_F)	.51***	.28	.63***
	(.006)	(.215)	(.006)
Hit Streak Momentum (H_M)	.48***	.66**	.39*
	(.008)	(.028)	(.057)
Hit Streak Length (H_L)	.43**	.54*	.38*
	(.021)	(.069)	(.071)
Total Runs (R)	−.58***	−.10	−.83***
	(.002)	(.387)	(.000)

50,000 Permutations (session strata)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ (one-sided)

JNI6's sessions the hit streak frequency statistic is above median levels (p=.002, binomial test), in eight of nine sessions the hit streak momentum statistic is above median levels (p=.02, binomial test), and in seven of nine sessions the hit streak length statistic is above median levels (p=.09, binomial test), while the runs statistic displays no particular (directional) pattern (p=.5, binomial test).

As in the analysis of RC, by pooling the shots from all of JNI6's sessions we can perform a more powerful individual-level test of the hot hand effect. As such, for each hit streak statistic we average its value across all nine of JNI6's sessions, weighting the value from each session equally. The last column of Table 1 (p. 17) reports that each of JNI6's average hit streak statistics is highly significant (p-values in parentheses).[56] To get a feel for the magnitude of each hit streak statistic, one can compare it against its median value under the null: for frequency (.27 vs. .22), momentum (.71 vs. .56), and length (9.6 vs. 7.0).

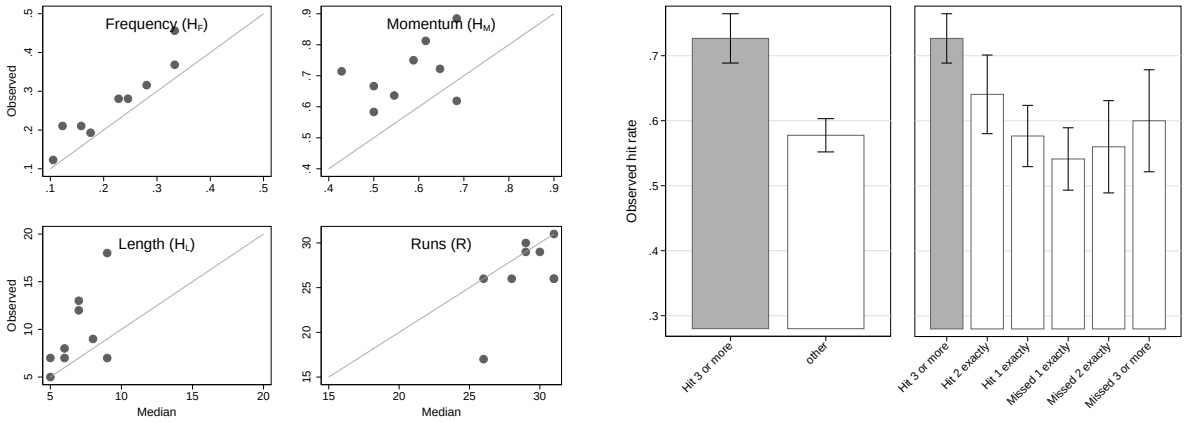
As in the case of RC, another way of checking whether JNI6's hot hand effect is not only highly significant, but also non-negligible in size, is by comparing his conditional hit rate immediately following a sequence of at least three hits with his conditional hit rate immediately following any other recent shot history. The left panel of Figure 4b shows that JNI6's conditional hit rate increases

[56] JNI6's hit streak statistics all remain highly significant even after a Bonferroni correction for JNI6 being only one of six players.

Figure 4: *Evidence of the hot hand from shooter "JNI6" of Jagacinski et al. [1979].*

(a) *In each cell, a statistic for each of the shooter's sessions is plotted against its respective median under the null.*

(b) *JNI6's hit rate, immediately following 3 or more hits in a row, is higher than after any other recent shot history (with standard errors).*



substantially, by around +15 percentage points, when immediately following three or more hits in a row, as compared to his conditional hit rate following all other recent shot histories (p=.001, two-sample test of proportions). This is larger than the aforementioned difference between the median and the very best NBA three point shooter (+12 percentage points), and is robust to multiple comparisons.

Also as in the case of RC, we can check whether JNI6's substantial changes in performance after streaks is being driven by the cold hand. The right panel of Figure 4b confirms that JNI6's performance differences are indeed being driven by the hot hand, and not by the cold hand, as his conditional hit rate immediately following a run of three or more hits exceeds his conditional hit rate immediately following each of the five other types of recent shot histories.[57,58] Finally, JNI6's miss streak statistics are not significant, which further corroborates that JNI6 has the hot hand, and not the cold hand.

While, for the purposes of hot hand beliefs, the presence of one player out of six with a substantial and persistent hot hand may already be more than expected (this would translate to roughly 2.5 such shooters on an NBA team), in addition to JNI6, a second player from the JNI study also

[57] The corresponding p-values for the test of proportions are .107, .006, .001, .015, and .062, respectively.

[58] As a robustness check for possible session effects we estimate a fixed-effects linear probability model of JNI6's hit rate, which corroborates these results (see Appendix D, Table 5).

exhibits significant hot hand effects, with significant hit streak frequency and runs statistics (.05 level for each).[59] The probability that at least 2 out of 6 players with constant hit rates cross this significance threshold is .03 (binomial test).

Evidence from Gilovich, Vallone, and Tversky [1985]

As mentioned in Section 3, at just one 100 shot session for each individual, GVT's shooting data is severely underpowered for testing for the hot hand at the individual level. Nevertheless, when comparing the hit rate after hitting 3 or more shots in a row to any other shot history, for each player, our tests reveal evidence of the hot hand effect in GVT's data, with a magnitude that would be substantial if it were to maintain in a larger sample. In particular, the conditional hit rate of 8 out of their 26 shooters increases by at least 10 percentage points immediately following a sequence of three or more hits, relative to any other recent shot history. By comparison, an i.i.d Bernoulli shooter's hit rate is expected to decrease by 4 percentage points, due to the bias identified in Miller and Sanjurjo [2018]. In Figure 5 we present the (uncorrected) performance of four such shooters, whose percentage point increases were +40, +28, +25 and +22 respectively—with the increases significant in all four of these shooters (.05 level, two-sample proportion test).[60] The probability of this occurrence under the null is $p = .039$ (binomial test). Moreover, the runs statistic for 5 out of 26 exceeds the one sided .05 significance threshold—an event that occurs with a binomial probability of $p = .009$ under the null.[61]

Discussion of individual level results

To put these individual-level results into perspective, the validity of GVT's conclusion was temporarily called into question by a study which claimed that a single player, Vinnie "the Microwave" Johnson, had the hot hand while playing for the Detroit Pistons in the 1987-1988 season (Larkey, Smith, and Kadane 1989), until the study was later called into question due to data coding errors (Tversky and Gilovich 1989b). Unlike in the case of Vinnie Johnson, whose shooting data was

[59] The hit streak momentum and length statistics for this player are not significant, though in the top quintile under the null.

[60] For the player with a +40 percentage point increase in shooting percentage, 30 of his shots were taken immediately following a run of three or more hits. For the other players the number of shots taken with this recent shot history were 15, 13, and 14, respectively.

[61] A further reason why individual-level tests are underpowered in the GVT data is that many of their players had a low overall hit rate (less than 35 percent), which reduced the amount of testable data for these players. There is some evidence of the cold hand in these players, e.g. one player never made more than two consecutive shots.

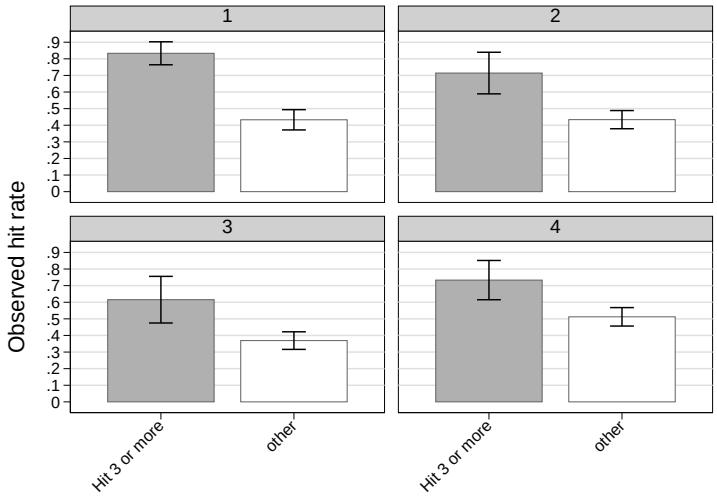


Figure 5: *The hit rates immediately following 3 or more hits in a row, for four shooters from GVT's study, are higher than after any other recent shot history (with standard errors).*

selected from a population of more than 300 players precisely because he was widely believed to be one of the hottest shooters in the NBA, we find clear evidence of sizeable hot hand shooting among players belonging to much smaller groups, who were selected only on the basis of availability.

4.3 Pooled Analysis: (Player-clustered) Average Effects

The individual-level analysis reported in Section 4.2 not only allows for a test of the existence of the hot hand in individuals, but also provides evidence of the heterogeneity of the hot hand effect across individuals. While we have demonstrated that some players systematically get the hot hand, other players appear to shoot with roughly a fixed hit rate, and still others actually under-perform immediately following a streak of hits. For these reasons a pooled test can only provide limited information about the existence of hot hand shooting in individuals; if one observes a pooled hot hand effect, then this suggests that at least one individual in the sample has the hot hand, whereas if no pooled effect is observed, without further information, one cannot know whether or not there are individuals with the hot hand in the sample. While hot hand beliefs are typically held with relation to individual shooters, as this is the level on which beliefs are likely most decision-relevant, a pooled analysis of shooting data can answer the question of whether the hot hand also happens to be a property of the average shooter in our sample.

Table 2: *Average values of the (normalized) hit streak statistics across players in our study, Gilovich et al. (1985), and Jagacinski et al. (1979), with p-values in parentheses. The hit streak statistics are computed by first standardizing the statistic from each player session using its null distribution for that session, then averaging across sessions for each player, and finally averaging across players.*

	Panel	GVT	JNI	Pooled †
Hit Streak Frequency (H_F)	.51***	.42**	.24**	.29**
	(.006)	(.017)	(.038)	(.013)
Hit Streak Momentum (H_M)	.48***	.37**	.06	.24**
	(.008)	(.031)	(.341)	(.027)
Hit Streak Length (H_L)	.43**	.24	.13	.16
	(.021)	(.109)	(.164)	(.112)
Total Runs (R)	−.58***	−.21	−.37***	−.21**
	(.002)	(.144)	(.003)	(.046)

Permutation test (50,000 permutations, with session strata)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ (one-sided)

†The panel has 3 sessions of 300 shots per shooter (8 shooters). GVT has 1 session of 100 shots per shooter (26 shooters). JNI has 9 sessions of 60 shots per shooter (6 shooters). The pooled analysis of all studies includes ad-hoc 300-shot sessions that were conduced with players who did not participate in the two-phase panel (5 shooters).

Table 2 reports the across-player average of the hit streak statistics (in standardized units), for each controlled shooting study, with the shooters from our panel in Column 1, GVT's shooters in Column 2, JNI's shooters in Column 3, and all shooters together in Column 4. Each hit streak statistic in the pooled data is generated by first standardizing the session level statistic for each player by subtracting the mean and dividing by the standard deviation for that session (under the null distribution generated for that session), then averaging across sessions for each player, and finally averaging across players. The null distribution is generated under the assumption of exchangeability within sessions (but not across, to avoid good-day/bad-day effects).

When considering just the players in our panel we find highly significant evidence of the hot hand in our pooled measures. In particular, all three of the hit streak statistics, as well as the runs statistic, are highly significant. Further, this effect is not driven by any single session; to the contrary, Figure 3b (p. 20) shows that hot hand performance in Phase One predicted the presence of hot hand performance in Phase Two (out of sample), six months later. When considering the players in the other two datasets we find that GVT's shooters have significant hit streak frequency and momentum statistics, and JNI's shooters have significant hit streak frequency and the runs statistics. Finally, we pool all available shooting sessions from all three studies (including ad-hoc

sessions conducted with players not included in our panel). This reveals a modest average hot hand effect across players of +.2 to +.3 standard deviation increase in performance, which is comparable to the average effect size found in pooled analyses of game data (Arkes 2010; Bocskocsky, Ezekowitz, and Stein 2014; Yaari and Eisenmann 2011). This result is consistent with the substantial and persistent hot hand effects observed among individual shooters in Section 4.2, which are diluted by the considerable variation in hot hand effect sizes observed across shooters.

5 Results: Beliefs

At the completion of Phase Two of the experiment a detailed questionnaire was administered in order to collect information on shooters' hot hand beliefs, and later test these beliefs against actual shooting performance. In particular, each shooter was asked to assess the degree of hot hand shooting in his teammates in our shooting task, in which, as described previously, each player shot from a distance at which he was expected to hit 50 percent of his shots (the observed average hit rate was 50.08 percent). Because a shooter's session was always observed by himself (and in some instances also by one other member of the panel), we test beliefs also after excluding a subject's beliefs about a shooter if he observed any of that shooter's sessions (see footnotes below).[62]

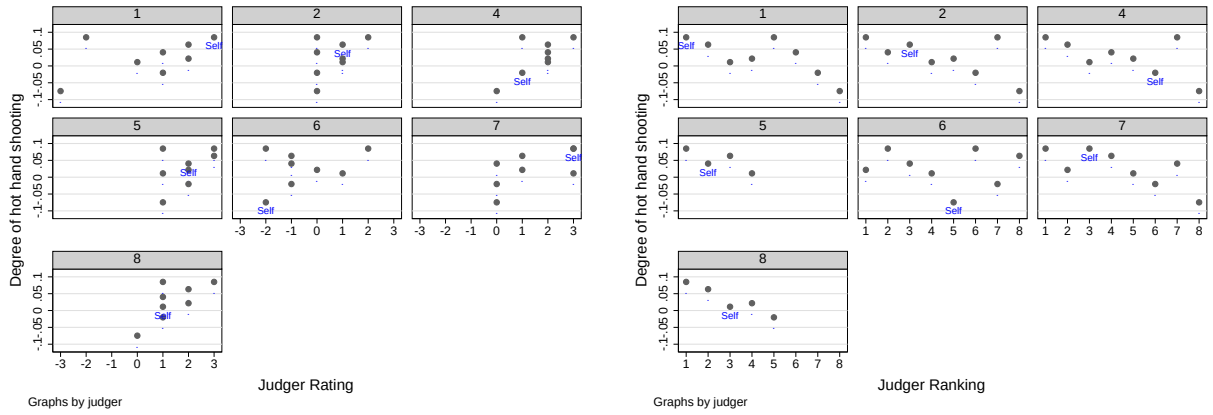
Subjects were first asked to create a rating of how much more or less likely each of the shooters in the panel is to make the next shot following three or more made shots in a row (relative to how each usually shoots) on a scale of -3 (much less likely) to +3 (much more likely). Each cell in Figure 6a contains a plot of the actual hot hand performance of each shooter (defined as the difference between the shooter's conditional hit rate immediately following three or more hits in a row and his overall hit rate) on the vertical axis against the corresponding ratings of a particular subject (henceforth "judger") on the horizontal axis.

Judgers' ratings indicate that they all believe that at least one other member of the panel has a hot hand, but only one judger believes that all eight shooters do. On average, they rate other shooters as slightly more likely to hit the next shot immediately following three or more hits (+ 1.0), and the shooter RC is rated as much more likely (+ 2.8). Further, six of the seven

[62] As a precautionary measure, despite subjects not being aware of the purpose of the study, before each shooting session we stressed in the instructions that they not communicate any information about the shooting session to others. Further, subjects were not provided any information about teammates' performance—hot hand related or otherwise—prior to the survey.

Figure 6: *Judger ratings and rankings predict which players have the tendency to exhibit a hot hand, as well as the relative degrees of hot hand shooting across teammates.*

(a) *Hot hand performance vs. judger ratings* **(b)** *Hot hand performance vs. judger rankings.*



shooters that judgers' rate (on average) as having a hot hand, do directionally exhibit a hot hand in performance. Likewise, the one shooter that judgers rate as having the opposite of a hot hand (an "anti-hot" hand) does directionally exhibit an anti-hot hand.[63] In addition, judgers' ratings of players are highly correlated with the players' actual degrees of hot hand shooting performance in the shooting task. In particular, eight out of eight judgers' correlations are positive—an event which has a binomial probability of .004 under the assumption that judger ratings are made at random. Further, the average judger correlation of 0.49 is found to be highly significant ($p < .0001$) in a permutation test conducted under the same null, and stratified at the judger level.[64,65]

Judgers were later asked to create a ranking of all of the shooters in the panel in terms of whose shooting percentage rises the most immediately after having made at least three shots in a row.[66] The names of the players were presented visually in a randomly scattered fashion so as to

[63] One can conduct a test of whether the average ratings of shooters (across judgers), correctly predict whether each shooter's performance directionally exhibits the hot hand or the opposite (the "anti-hot" hand). The sign of the hot hand effect is correctly predicted for seven out of eight shooters, which is significant under the null hypothesis that the sign of the average rating is determined at random (p=.035, binomial test). When we exclude judger-observed ratings, the sign is correctly predicted for six out of eight shooters (p = 0.144, binomial test), as one of the seven directionally hot hand shooters is believed by judgers to be anti-hot.

[64] When we exclude judger-observed sessions, the average rating for all shooters is +0.8 (for RC it remains +2.8). Further, six out of seven subjects' correlations are positive (p=.064, binomial test), with a highly significant average correlation of .38 ($p = .016$, stratified permutation test).

[65] Here it is assumed that judgers uniformly randomize among the ratings that we observed them make at least once. If we instead assume that they randomize across all possible ratings the statistical significance is even stronger.

[66] After the rating task, but before the ranking task, judgers were additionally given a slightly more involved series of questions in which they were told each shooter's base rate (overall hit rate), and asked to provide a numerical

minimize the possibility of priming effects (see Figure 7). Figure 6b is analogous to 6a, but with the degrees of hot hand performance across shooters now plotted against each judger's hot hand rankings, rather than ratings. Consistent with responses in the ratings questions, judgers' by far rank RC as the hottest shooter (1.1 average rank vs. 3.4 for next highest average rank), and overall rankings are highly inversely correlated with shooters' respective degrees of hot hand shooting in our experiment. In particular, seven of seven judgers' correlations are negative (p = .008, binomial test), and the average correlation is $-.60$ ($p < .0001$, stratified permutation test).[67]

Figure 7: *Scattered display of the eight shooter names used for the ranking task (generic names used here to protect subject anonymity).*

Overall, we find that the beliefs of our expert players about the hot hand shooting of their teammates, which were formed during *prior* playing experience, correctly predict which of their teammates have a greater tendency to exhibit the hot hand in our controlled shooting task (out-of-sample). Thus, not only are players correct to believe in the hot hand, but this is the first

percentage estimate of the shooter's conditional hit rate immediately following three or more hits in a row. Interestingly, the implied difference in hit rates is not correlated with the responses provided in either the ranking or rating task, despite these being highly correlated to one another. One possible explanation, which is supported by observed attrition rates, is that numerical estimates are less natural and more difficult for players than rankings or ratings. In particular, in GVT only 5 of the 8 subjects provided numerical estimates, whereas at least 7 (and typically 8) responded to each of the other questions, and in our survey 1 of 8 judgers stopped filling in the questionnaire precisely at this point.

[67] Excluding judger-observed sessions, the judgers' average ranking of RC is 1.2, with the next highest average rank being 2.8. Six out of six judgers' correlations are negative ($p = .016$, binomial test) and the average correlation is $-.60$ ($p < .01$, stratified permutation test)

evidence to suggest that they can identify it, and even relative degrees of it in different shooters, accurately.[68]

6 Conclusion

We generate and gather the richest datasets to date in order to provide a robust test of whether belief in the existence of the hot hand is a fallacy in its canonical domain. That we can test for the hot hand in individual performance is crucial, as this is the natural subject of hot hand fallacy beliefs (as opposed to whether groups of players "on average" have the hot hand). The result of our improved statistical approach is strong evidence of substantial, and persistent, momentum in shooting performance on the individual level that is driven by hot hand (and not cold hand) shooting. This evidence robustly invalidates the notion that belief in the existence of a substantial hot hand is a fallacy. Also, the method we introduce to identify hot hand performance (separating it from cold hand performance), along with the other notable features of our statistical approach, readily lend themselves to use in future studies of momentum, and in particular the hot hand, in both human and non-human performance.

Given the variation in the sign and magnitude of individual hot hand performance that we observe there is little reason to expect a hot hand effect when the data from all shooters is pooled. Nevertheless, we find a hot hand effect across all extant controlled shooting studies. This evidence is consistent with the pooled hot hand effects observed in relatively less controlled environments such as the ("semi-controlled") NBA three point contest (Miller and Sanjurjo 2015),[69] NBA free throw shooting (Arkes 2010; Yaari and Eisenmann 2011), and NBA live ball shooting (Bocskocsky et al. 2014).

With the robust evidence of the hot hand, and the consequent invalidation of the hot hand *fallacy*, it becomes natural to begin a more direct exploration of the fitness of hot hand beliefs themselves. This is an area of investigation that has long laid relatively dormant given that the

[68] In addition, these results on beliefs also suggest that: (i) the variation we observe in the estimated hot hand effect across players is due to heterogeneity in their tendency to exhibit hot hand shooting, and (ii) players who have a tendency to get (more) hot in controlled shooting environments also have a tendency to get (more) hot in practice and games.

[69] The NBA three point contest is less controlled than a shooting experiment yet also has some of the same important features. For an analysis of 29 years of shootout data using the empirical approach introduced here see Miller and Sanjurjo (2015).

fallacy paradigm only logically required the absence of evidence of hot hand performance in order for the widespread belief in it to remain fallacious. Our novel hot hand beliefs data reveal that players are able to identify which of their teammates have more (or less) of a tendency to become hot. This is the first evidence of its kind, and is clearly inconsistent with the fallacy view. In particular, GVT's betting study on hot hand beliefs, which Gilovich has referred to as "...the most important bit of evidence against the hot hand," found that players are unable to successfully predict the shot outcomes of their teammates.[70] However, it has recently been shown that the statistical tests used in GVT's betting study are severely underpowered, and the effect sizes misinterpreted (Miller and Sanjurjo 2017). In fact, an improved re-analysis of their data yields strong evidence that players are skilled at predicting their teammates' shot outcomes (Miller and Sanjurjo 2017), which is consistent with the results on beliefs what we observe here.

Of course, while our evidence on beliefs indicates no signs of bias in terms of ability to identify the hot hand, or even recognize which players tend to have more or less of one, it does not preclude other possible biases in beliefs, such as people (even experts) systematically over (or under) estimating the frequency of the hot hand, and/or its magnitude when it occurs.[71] We hope that the evidence of hot hand performance presented in this paper, along with our results on beliefs, encourage further exploration of the so-far relatively understudied beliefs side of the hot hand debate.

[70] The content is from a 2015 letter that Gilovich granted us permission to share publically.

[71] For recent evidence of over-reaction among professional dart players see Jin (2017). For recent evidence of under-reaction in NCAA basketball see Stone and Arkes (2017).

References

Aharoni, G. and O. H. Sarig (2011): "Hot hands and equilibrium," *Applied Economics*, 44, 2309–2320.

Albert, J. (1993): "Comment on "A Statistical Analysis of Hitting Streaks in Baseball" by S. C. Albright," *Journal of the American Statistical Association*, 88, 1184–1188.

Albert, J. and P. Williamson (2001): "Using Model/Data Simulations to Detect Streakiness," *The American Statistician*, 55, 41–50.

Angrist, J. D. and J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Arkes, J. (2010): "Revisiting the Hot Hand Theory with Free Throw Data in a Multivariate Framework," *Journal of Quantitative Analysis in Sports*, 6.

——— (2011): "Do Gamblers Correctly Price Momentum in NBA Betting Markets?" *Journal of Prediction Markets*, 5, 31–50.

——— (2013): "Misses in 'Hot Hand' Research," *Journal of Sports Economics*, 14, 401–410.

Avery, C. and J. Chevalier (1999): "Identifying Investor Sentiment from Price Paths: The Case of Football Betting," *Journal of Business*, 72, 493–521.

Avugos, S., M. Bar-Eli, I. Ritov, and E. Sher (2013): "The elusive reality of efficacy performance cycles in basketball shooting: analysis of players' performance under invariant conditions," *International Journal of Sport and Exercise Psychology*, 11, 184–202.

Bandura, A. (1982): "Self-Efficacy Mechanism in Human Agency," *American Psychologist*, 37, 122–147.

Barberis, N. and R. Thaler (2003): "A survey of behavioral finance," *Handbook of the Economics of Finance*, 1, 1053–1128.

Bocskocsky, A., J. Ezekowitz, and C. Stein (2014): "The Hot Hand: A New Approach to an Old 'Fallacy'," 8th Annual Mit Sloan Sports Analytics Conference.

Brown, W. A. and R. D. Sauer (1993): "Does the Basketball Market Believe in the Hot Hand? Comment," *American Economic Review*, 83, 1377–1386.

Camerer, C. F. (1989): "Does the Basketball Market Believe in the 'Hot Hand,'?" *American Economic Review*, 79, 1257–1261.

Carlson, K. A. and S. B. Shu (2007): "The rule of three: how the third event signals the emergence of a streak," *Organizational Behavior and Human Decision Processes*, 104, 113–121.

Churchland, M. M., A. Afshar, and K. V. Shenoy (2006): "A Central Source of Movement Variability," *Neuron*, 52, 1085–1096.

Croson, R. and J. Sundali (2005): "The Gambler's Fallacy and the Hot Hand: Empirical Data from Casinos," *Journal of Risk and Uncertainty*, 30, 195–209.

Csikszentmihalyi, M. (1988): "The flow experience and its significance for human psychology," in *Optimal experience: Psychological studies of flow in consciousness*, ed. by M. Csikszentmihalyi and I. S. Csikszentmihalyi, New York, NY, US: Cambridge University, chap. 2.

De Bondt, W. P. (1993): "Betting on trends: Intuitive forecasts of financial risk and return," *International Journal of Forecasting*, 9, 355–371.

De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann (1991): "The Survival of Noise Traders In Financial-markets," *Journal of Business*, 64, 1–19.

Dixit, A. K. and B. J. Nalebuff (1991): *Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life*, W.W. Norton & Company.

Dorsey-Palmateer, R. and G. Smith (2004): "Bowlers' Hot Hands," *The American Statistician*, 58, 38–45.

Durham, G. R., M. G. Hertzel, and J. S. Martin (2005): "The Market Impact of Trends and Sequences in Performance: New Evidence," *Journal of Finance*, 60, 2551–2569.

Ernst, M. D. (2004): "Permutation Methods: A Basis for Exact Inference," *Statistical Science*, 19, 676–685.

Gilovich, T. (2008): *How we know what isn't so*, Simon and Schuster.

Gilovich, T., R. Vallone, and A. Tversky (1985): "The Hot Hand in Basketball: On the Misperception of Random Sequences," *Cognitive Psychology*, 17, 295–314.

Goldman, M. and J. M. Rao (2012): "Effort vs. Concentration: The Asymmetric Impact of Pressure on NBA Performance," 6th Annual Mit Sloan Sports Analytics Conference.

Good, P. (2005): *Permutation, Parametric, and Bootstrap Tests of Hypotheses.*, New York: Springer.

Green, B. and J. Zwiebel (2017): "The hot-hand fallacy: cognitive mistakes or equilibrium adjustments? Evidence from major league baseball," *Management Science*, 64, 5315–5348.

Guryan, J. and M. S. Kearney (2008): "Gambling at Lucky Stores: Empirical Evidence from State Lottery Sales," *American Economic Review*, 98, 458–473.

Hooke, R. (1989): "Basketball, baseball, and the null hypothesis," *Chance*, 2, 35–37.

Houthakker, H. S. (1961): "Systematic and Random Elements in Short-Term Price Movements," *American Economic Review*, 51, 164–172.

Jagacinski, R. J., K. M. Newell, and P. D. Isaac (1979): "Predicting the Success of a Basketball Shot at Various Stages of Execution," *Journal of Sport Psychology*, 1, 301–310.

Jin, L. (2017): "Evidence of Hot-Hand Behavior in Sports and Medicine," *Working Paper*.

Kahneman, D. (1973): *Attention and Effort.*, Prentice Hall.

——— (2011): *Thinking, Fast and Slow*, Farrar, Straus and Giroux.

Kahneman, D. and M. W. Riepe (1998): "Aspects of Investor Psychology: Beliefs, preferences, and biases investment advisors should know about," *Journal of Portfolio Management*, 24, 1–21.

Koehler, J. J. and C. A. Conley (2003): "The "hot hand" myth in professional basketball," *Journal of Sport and Exercise Psychology*, 25, 253–259.

Korb, K. B. and M. Stillwell (2003): "The Story of The Hot Hand: Powerful Myth or Powerless Critique?" Working Paper.

Larkey, P. D., R. A. Smith, and J. B. Kadane (1989): "It's Okay to Believe in the 'Hot Hand'," *Chance*, 2, 22–30.

Lee, M. and G. Smith (2002): "Regression to the mean and football wagers," *Journal of Behavioral Decision Making*, 15, 329–342.

Loh, R. K. and M. Warachka (2012): "Streaks in Earnings Surprises and the Cross-Section of Stock Returns," *Management Science*, 58, 1305–1321.

Malkiel, B. G. (2011): *A random walk down Wall Street: the time-tested strategy for sucessful investing*, New York: W. W. Norton & Company.

Miller, J. B. and A. Sanjurjo (2015): "Is it a Fallacy to Believe in the Hot Hand in the NBA Three-Point Contest?" OSF Preprints: https://doi.org/10.31219/osf.io/dmksp.

——— (2017): "A Visible (Hot) Hand? Expert Players Bet on the Hot Hand and Win," OSF Preprints: https://doi.org/10.31219/osf.io/sd32u.

——— (2018): "Surprised by the Hot Hand Fallacy? A Truth in the Law of Small Numbers," *Econometrica*, 86.

Miyoshi, H. (2000): "Is the "hot hands" phenomenon a misperception of random events?" *Japanese Psychological Research*, 42, 128–133.

Mood, A. M. (1940): "The Distribution Theory of Runs," *The Annals of Mathematical Statistics*, 11, 367–392.

Moskowitz, T. and L. J. Wertheim (2011): *Scorecasting: The Hidden Influence Behind How Sports are Played and Games are Won*, New York: Crown Archetype.

Narayanan, S. and P. Manchanda (2012): "An empirical analysis of individual level casino gambling behavior," *Quantitative Marketing and Economics*, 10, 27–62.

Paul, R. J. and A. P. Weinbach (2005): "Bettor Misperceptions in the NBA: The Overbetting of Large Favorites and the 'Hot Hand'," *Journal of Sports Economics*, 6, 390–400.

Rabin, M. and D. Vayanos (2010): "The Gambler's and Hot-Hand Fallacies: Theory and Applications," *Review of Economic Studies*, 77, 730–778.

Rao, J. M. (2009a): "Experts' Perceptions of Autocorrelation: The Hot Hand Fallacy Among Professional Basketball Players," Working Paper.

——— (2009b): "When the Gambler's Fallacy becomes the Hot Hand Fallacy: An Experiment with Experts and Novices," Working Paper.

Sinkey, M. and T. Logan (2013): "Does the Hot Hand Drive the Market?" *Eastern Economic Journal*, Advance online publication, doi:10.1057/eej.2013.33.

Smith, G., M. Levere, and R. Kurtzman (2009): "Poker Player Behavior After Big Wins and Big Losses," *Management Science*, 55, 1547–1555.

Stern, H. S. and C. N. Morris (1993): "Comment on "A Statistical Analysis of Hitting Streaks in Baseball" by S. C. Albright," *Journal of the American Statistical Association*, 88, 1189–1194.

Stone, D. F. (2012): "Measurement error and the hot hand," *The American Statistician*, 66, 61–66, working paper.

Stone, D. F. and J. Arkes (2017): "March Madness? Underreaction to Hot and Cold Hands in NCAA Basketball," *Working Paper*, –.

Sundali, J. and R. Croson (2006): "Biases in casino betting: The hot and the gambler's fallacy," *Judgement and Decision Making*, 1, 1–12.

Swartz, T. (1990): "Letter to the editor: More on the "hot hand"," *Chance*, 3, 6–7.

Thaler, R. H. and C. R. Sunstein (2008): *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Yale University Press.

Tversky, A. and T. Gilovich (1989a): "The cold facts about the "hot hand" in basketball," *Chance*, 2, 16–21.

——— (1989b): "The"Hot Hand": Statistical Reality or Cognitive Illusion?" *Chance*, 2, 31–34.

Wardrop, R. L. (1995): "Simpson's Paradox and the Hot Hand in Basketball," *The American Statistician*, 49, 24–28.

——— (1999): "Statistical Tests for the Hot-Hand in Basketball in a Controlled Setting," *Unpublished manuscript*, 1, 1–20.

Xu, J. and N. Harvey (2014): "Carry on winning: The gambler's fallacy creates hot hand effects in online gambling," *Cognition*, 131, 173 – 180.

Yaari, G. and S. Eisenmann (2011): "The Hot (Invisible?) Hand: Can Time Sequence Patterns of Success/Failure in Sports Be Modeled as Repeated Random Independent Trials?" *PLoS One*, 6, 1–10.

Yuan, J., G.-Z. Sun, and R. Siu (2014): "The Lure of Illusory Luck: How Much Are People Willing to Pay for Random Shocks," *Journal of Economic Behavior & Organization*, forthcoming.

A Appendix: challenges in studying game shooting data

Live ball data

The in-game play-by-play data that GVT analyze (Study 2: 76ers, 1980-81 season: 9 players, 48 home games) is not ideal for the study of hot hand shooting. Perhaps the most notable concern is that the opposing team has incentive to make *costly* strategic adjustments to mitigate the impact of a "hot" player (Dixit and Nalebuff 1991, p. 17). This concern has been emphasized by researchers in the hot hand literature (Aharoni and Sarig 2011; Green and Zwiebel 2017), and is not merely theoretical. For example, while GVT observed that a shooter's field goal percentage is lower after consecutive successes, they also acknowledged the possibility that the result was being affected by omitted variable bias (shot selection and defensive pressure). To this end, subsequent studies have introduced partial controls for defensive pressure (and shot location), and as a result found that the aforementioned anti-hot hand effect disappears (Bocskocsky et al. 2014; Rao 2009a). Further, evidence of specific forms of strategic adjustment has been documented (Aharoni and Sarig 2011; Bocskocsky et al. 2014).

Dead ball data

The in-game free throw data that GVT analyze (Study 3: Celtics, 1980-81, 1981-82 seasons: 9 players), while controlled, is not ideal for the study of hot hand shooting for a number of reasons: (i) hitting the first shot in a pair of isolated shots is not typically regarded by fans and players as hot hand shooting (Koehler and Conley 2003), presumably due to the high prior probability of success ($\approx .75$), (ii) hitting a single shot is a weak signal of a player's underlying state, which can lead to severe measurement error (Arkes 2013; Stone 2012), (iii) there is a potential for omitted variable bias, as free throw pairs are relatively rare, and shots must be aggregated across games and seasons in order to have sufficient sample size. In any event, subsequent studies of free throw data have found evidence that is inconsistent with the conclusions that GVT drew from the Celtics' data (Aharoni and Sarig 2011; Arkes 2010; Goldman and Rao 2012; Wardrop 1995; Yaari and Eisenmann 2011).

B Appendix: Statistical tests

B.1 Permutation Test Procedure

Under the null hypothesis (H_0) that a player does not get hot (or cold), the player's shooting performance is a sequence of i.i.d. Bernoulli trials with a fixed probability of success. While a player's true success rate is unknown to the experimenter, conditional on the number of successes $N := \sum_{s \in S} x_s$, the shot outcomes are *exchangeable*, i.e. all orderings of the shot outcomes are equally likely under H_0. This means that for a single player's realization of a sequence of 300 shots, an exact (discrete) distribution exists for each statistic outlined above under H_0, by exchangeability; enumerating each permutation of the player's shots, and calculating the value of the test statistic gives this distribution. For a test statistic T defined on the sequence of shot outcomes $\{x_s\}_{s \in S}$, the (one-sided) test with a significance level α is defined by the family of critical values $\{c_{\alpha,k}\}$; if the sequence has $N = k$ successes, $c_{\alpha,k} \in \mathbb{R}$ is the smallest value such that $\mathbb{P}(T \geq c_{\alpha,k} | N = k, H_0) \leq \alpha$. Therefore, $\mathbb{P}(\text{reject} | H_0) = \sum_{k=1}^{\#S} P(T \geq c_{\alpha,k} | \sum_{s \in S} x_s = k, H_0) P(N = k | H_0) \leq \alpha$. While the enumeration required to calculate $\mathbb{P}(T \geq t | N = k, H_0)$ is computationally infeasible, the exact distribution of the test statistic can be numerically approximated to arbitrary precision with a Monte-Carlo permutation of that player's shot outcomes (Ernst 2004; Good 2005).[72,73]

B.2 Distributions

Claim: The hit streak frequency statistic (H_F) has a normal asymptotic distribution (as $\#S \to \infty$).

Proof: It is sufficient to show that $\#S_H = \#\{s \in S : x_{s-1} = x_{s-2} = x_{s-3} = 1\}$ has a normal asymptotic distribution. Among the $m := \#S - 1$ shots that are not the final shot let n_1 be the number of hits, r_{1j} be the number of runs of hits of length j, and s_{1k} be the number of runs of hits of length k or more.[74] Clearly, the number of (overlapping) sequences of 3 hits in a row satisfies

[72] We performed simulations to verify that our code accomplished this (not reported here). For example, for each statistic we determined the 5 percent critical threshold $c_{.05,k}$ by permutation for each k of the 300 possible realized success rates and with this we find that in slightly fewer than 5 percent of our simulated $Bernoulli(p)$ 300-trial experiments with a fixed theoretical success rate p the null was rejected using a test that, in each experiment, selects the permutation test's critical threshold $c_{.05,k}$ corresponding to the player's realized success rate $k/300$ generated by the Bernoulli process for that experiment. This holds for a range of underlying fixed Bernoulli success rates and critical region sizes.

[73] For the test statistic H_M, if the number of successes k is small, then for a fraction of permutations H_M may not be computable as a player may never hit three (or more) shots in a row in a sequence of 300 shots.

[74] Note that $s_{1k} = \sum_{j=k}^{m} r_{1j}$.

$\#S_H = \sum_{j=3}^{m} r_{1j}(j-2)$, which can easily be simplified to $n_1 - 2r_{12} - r_{11} - 2s_{13}$. Theorem 1 of Mood (1940) shows that for n_1/m fixed, (r_{11}, r_{12}, s_{13}) has a (multivariate) normal asymptotic distribution. Therefore, as a corollary, the asymptotic distribution of $\#S_H$ is normal.

■

Claim: H_M is a ratio of two asymptotically normal random variables.

Proof: $H_M := \sum_{s \in S_H} x_s/(\#S_H - 3)$. The proof of the previous claim establishes that the denominator is asymptotically normal. The numerator satisfies $\sum_{s \in S_H} x_s = \#\{s \in S : x_s = x_{s-1} = x_{s-2} = x_{s-3} = 1\} = \sum_{j=4}^{n} r_{1j}(j-3)$, where $n := \#S$, and we can apply the same reasoning as above.

■

Claim: The probability of the hit streak length statistic (H_L) at each value in the discrete support can by approximated using a normal distribution.

Proof: Let $n := \#S$ be the number of shots, n_1 be the number of hits, r_{1j} be the number runs of hits of length j, and s_{1k} be the number of runs of hits of length k or more. The support of H_L is $\{1, 2, \ldots, n_1\}$, and its discrete distribution is given by:

$$\mathbb{P}(H_L = \ell) = \begin{cases} \mathbb{P}(s_{12} = 0) & \text{if } \ell = 1 \\ \mathbb{P}(r_{1\ell} \geq 1, s_{1\ell+1} = 0) & \text{if } 1 < \ell < n_1 \\ \mathbb{P}(s_{1n_1} = 1) & \text{if } \ell = n_1 \end{cases}$$

Theorem 1 of Mood (1940) shows that for n_1/n fixed, for $1 \leq \ell < n_1$, $(r_{11}, r_{12}, \ldots, r_{1\ell}, s_{1\ell+1})$ has a (multivariate) normal asymptotic distribution. Therefore, each probability can be approximated from the associated multivariate normal distribution (with a continuity correction).

■

Claim: The first order serial correlation statistic and the runs statistic yield the same test for a shooter with a 50 percent hit rate.

Proof: Let $n := \#S$ be the number of shots. The first order serial correlation statistic is defined as $\hat{\rho}_{t,t+1}(\boldsymbol{x}) := \sum_{s=1}^{n-1}(x_s - \bar{x})(x_{s+1} - \bar{x}) / \sum_{s=1}^{n}(x_s - \bar{x})^2$, and the runs statistic can be represented as the number of switches, plus one: $R(\boldsymbol{x}) = 1 + \sum_{s=1}^{n-1}[1 - x_s x_{s+1} - (1 - x_s)(1 - x_{s+1})]$. It is easy to show that $R(\boldsymbol{x}) = (n+1)/2 - (n/2)\hat{\rho}_{t,t+1}(\boldsymbol{x})$ when the hit rate is fixed at .5, which implies that

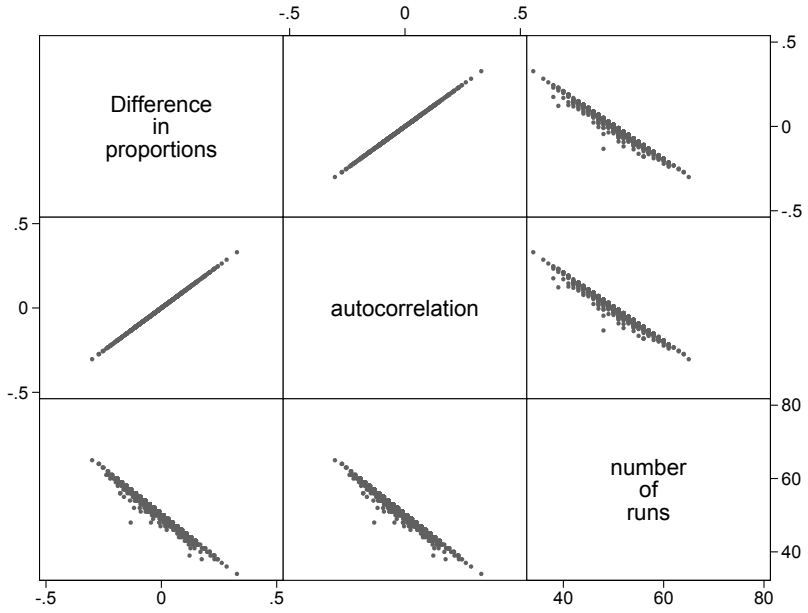


Figure 8: *The scatter plot matrix for the difference in proportions $\hat{p}(hit|1\ hit) - \hat{p}(hit|1\ miss)$ conditional probability test for $k = 1$, first order autocorrelation, and the number of runs, generated from 1000 simulated sequences of 100 flips of a fair coin.*

across all permutations, ρ and R are perfectly correlated; therefore, a runs test will reject the null if and only if the serial correlation test rejects the null. ■

In the general case shooting performance may instead deviate from the design target of 50 percent. Nevertheless, in this case the correlation between the statistics remains nearly perfect because $\tilde{R}(\boldsymbol{x}) := -2(n-1)\hat{\sigma}^2\hat{\rho}_{t,t+1}(\boldsymbol{x}) + (n+1)/2$ is a close approximation to R, and is perfectly correlated with ρ. In addition, the absolute difference satisfies $|R(\boldsymbol{x}) - \tilde{R}(\boldsymbol{x})| = |2\hat{\mu} - 1|(x_1 + x_n)$, where $\hat{\mu}$ is the fraction of hit shots and $\hat{\sigma}^2$ is the standard deviation, both fixed across permutations. This implies that GVT's runs test and serial correlation test are redundant. In addition, GVT's conditional probability test for $k = 1$ also happens to be equivalent to these two tests because the serial correlation is nearly identical to the associated difference in the conditional probability test. To illustrate further, in Figure 8 we provide a scatter plot matrix that shows how tightly related these statistics are in a comparison based on 1000 simulated sequences of a 100 flips of a fair coin.

B.3 Power & Effect Size

Below we report the power of the hit streak statistics (see Section 3) to detect departures from the null hypothesis of consistent shooting at the $\alpha = .05$ significance level for three alternative hot hand models, across a range of parameters we select. Further, we consider two tests commonly used in the previous literature: the runs test and the conditional probability test (which compares the hit rate immediately following three consecutive hits to the hit rate immediately following three consecutive misses).

Of the statistics we propose, H_F and H_M outperform the runs and conditional probability tests across all alternative models of hot hand shooting, and the difference in power is particularly large in the case that a shooter has taken 300 or fewer shots. While all statistics are underpowered at detecting the hot hand when an individual shooter has taken 100 shots as in GVT (regardless of the model and chosen parameters), with 300 shots, H_F and H_M (but not H_L) are typically adequately powered (though this is only true for certain models of hot hand shooting). These results illustrate the importance of having a particularly large sample size when testing for the existence of the hot hand on the individual level.

We can estimate the magnitude of the hot hand effect by comparing the hit streak momentum statistic H_M to performance when the shooter has not just hit three or more consecutive shots:

$$\text{Magnitude of hot hand effect } = H_M - \frac{\sum_{s \in S_H} x_s}{\#S_H^C}$$

As discussed below, most tests dramatically underestimate the true magnitude of the hot hand effect, as modeled.

Below we detail these results for the three alternative models of the data generating process that we consider: the regime shift model, the positive feedback model, and the hit streak model.

Regime Shift Model

We consider a player with a baseline state ($Y = b$), with hit probability p, and a hot state ($Y = h$) with hit probability $p + \delta$, where $\delta > 0$. The transition probability $q_{ij} := \mathbb{P}(Y = j | Y = i)$, where $i, j \in \{b, h\}$, satisfies $0.75 < q_{hh} < q_{bb}$, with q_{bb} large (near 1). Let π be the stationary distribution, i.e. the fraction of the time in the baseline state. In our simulations we vary $p \in \{.4, .5, .6\}$, $\delta \in \{.1, .2, .3, .4\}$, $q_{bb} \in \{.97, .98, .99\}$, and $\pi \in \{.80, .85, .90\}$, and find that even with 900 shots, all

statistics have low power on average, i.e. below .5. If we restrict our parameters to more extreme values $p \in \{.5, .6\}$, $\delta \in \{.3, .4\}$, $q_{bb} \in \{.99\}$, and $\pi \in \{.80, .85\}$, for 900 shots, H_F, H_M and H_L have an average power of .73, .80 and .76 respectively, whereas the runs tests and conditional probability tests have an average power of .6. In these most extreme cases, in which the hot hand increases shooting percentage by 30 or 40 percentage points, the estimated magnitude of the hot hand effect is severely downward biased, with only 9 and 17 percentage point increases in shooting percentage, respectively (for a true effect of 20 the estimated effect is only 4).

Positive Feedback Models

We consider two models of positive feedback and vary the strength of the serial dependence that each employs. Let p be a player's hit rate.

For the first feedback model, which we term the "feedback streak" model, the probability of a successful shot is defined as follows:

$$\mathbb{P}(x_s = 1) = \begin{cases} p + \delta & \text{if } x_{s-1} = 1, x_{s-2} = 1, x_{s-3} = 1 \\ p & \text{otherwise} \end{cases}$$

In our simulations we vary $p \in \{.4, .5, .6\}$ and $\delta \in \{.1, .2, .3, .4\}$. Unsurprisingly, the Runs statistic is poor at detecting the hot hand in this case: the average power is .48 for 300 shots, and .67 for 900 shots. On the other hand, the H_F and H_M statistics are more powerful at detecting these departures from hot hand shooting (.6 and .75, respectively, for 300 shots, and .8 and .9, respectively, for 900 shots), and also more powerful than the conditional probability test (.57 for 300 shots, .8 for 900 shots).[75] In this model the estimated magnitude of the true hot hand effect is not biased downward, as a streak of three or more hits is precisely the cause of the increase in a player's hit rate.

In the second feedback model, which we term the "feedback ladder" model, the probability of

[75] H_L has a power of .6 for 900 shots.

a successful shot is defined as follows:

$$\mathbb{P}(x_s = 1) = \begin{cases} p & \text{if } x_{s-1} \neq 1 \\ p + k\delta & \text{if } \exists k < K : x_{s-1} = 1, \ldots, x_{s-k} = 1 \ \& \ x_{s-k-1} \neq 1 \\ p + K\delta & \text{if } \exists k \geq K : x_{s-1} = 1, \ldots, x_{s-k} = 1 \ \& \ x_{s-k-1} \neq 1 \end{cases}$$

with $k, K \in \{1, 2, \ldots, n\}$, where n is the number of shots.

In our simulations we vary $p \in \{.4, .5, .6\}$, $(K, \delta) \in \{3\} \times \{.05, .10, .15\} \cup \{5\} \times \{.02, .03, .04, .05\}$. For $K = 3$, and 300 shots, unsurprisingly, given the first order serial dependence, the runs test performs well with a power of .8. In comparison, the H_F and H_M statistics each have an average power that is as large or larger (.85 and .80, respectively) with H_F outperforming the runs test by up to 10 percentage points for certain parameter values (H_L has a power of .5). The conditional probability test, on the other hand, has a power of .61. All tests are highly powered with 900 shots (except the one based on H_L). The estimated magnitude of the hot hand effect is downward biased, reflecting only 2/3 to 4/5 of the true effect size in this case. For $K = 5$ the results are similar, except H_F outperforms the runs test by 10 percentage points on average (and H_M outperforms it by 5 percentage points). In this case, the estimated magnitude of the hot hand effect has a slight downward bias (9/10) with respect to the true effect size.

Hit Streak Model

We consider a player with a baseline state ($Y = b$), with hit probability p, and a hot state ($Y = h$) with hit probability p_h. If the player is in a baseline state, the player enters into a hot state with probability q, and remains there for exactly K shots, at which point he returns to the baseline state, with the possibility of entering into the hot state again. In our simulations we vary $p \in \{.4, .5, .6\}$, $p_h \in \{.9, .95, 1\}$, $q \in \{.01, .02, .03\}$, and $K \in \{10, 15, 20\}$. For 300 shots, the runs and conditional probability tests have an average power of .62 and .64, respectively, whereas H_F, H_M and H_L have an average power of .75, .82, and .70, respectively. For 900 shots, the runs and conditional probability tests have an average power of .86 and .84, whereas H_F, H_M and H_L have an average power of .93, .96, and .76, respectively.

C Appendix: Experimental Procedures & Instructions

C.1 Instructions: Shooting Experiment

Shooting

INSTRUCTIONS: SHOOTERS

Your Task:

You will shoot 300 times with your toes just behind the line marked with tape point to line. There is no need to move from the line because the rebounder will rebound each of your shots and pass the ball back to you once I signal him to do so via a clear audible signal. Once you have received the ball you can shoot when you like.

As you shoot I will be sitting on the side of the court recording whether you make or miss each shot. The video cameras you see will be filming your shots.

Payoffs:

Of your 300 shot opportunities 10 of them have been selected at random as shots for which you will be paid. For each of these 10 selected shots that you make you will be paid 6,00 Euros. For each of these 10 selected shots that you miss you will be paid 0,00 Euros. Thus you can earn between 0,00 and 60,00 Euros for your shooting. The 10 paid shots were chosen by a random number generator this morning and are in this sealed envelope which I will leave here and which we will open together before calculating your payoffs.

Independent of how many shots you make, you will receive 5,00 Euros for participating. This means that in total you can be paid up to 65,00 Euros (60,00 Euros for your shots + 5,00 Euros for participating).

Once you finish your 300 shots you and I will calculate your payoffs. We will do this by first opening (together) your envelope with the 10 randomly selected shot numbers, then seeing which of the corresponding shots you made and missed, with a 6,00 Euro payment for each make. Then I will pay you your money and you will be free to leave.

Communication is Prohibited:

While you are shooting please do not directly communicate with the rebounder or with me in any way.

Summary:

You will now shoot 300 times with your toes just behind the line marked with tape. Once you have finished you and I will calculate your payoffs. Then I will pay you.

Do you have any questions? If so, please ask now because once you start shooting I will not be able to answer any of your questions.

Thank you for your attention.

You are now free to start shooting your shots. I will announce to you when you have completed your 50th, 100th, 150th, 200th, 250th, and 300th shot so you know how many shots remain.

Rebounding

INSTRUCTIONS: REBOUNDER

You will be asked to rebound each of the 300 shots performed by the shooter. You have a line marked with tape from where you will always pass the ball to the shooter, while facing the shooter squarely. I ask that you always deliver the same pass, a two-handed bounce pass originating from the head. I ask that you try to be as mechanical and repetitive as possible, so the situation changes as little as possible for the shooter, from shot to shot. Before the shooter's first shot, you will stand on your marked line, facing him squarely, and once you hear a clear audible signal you will deliver him the two-handed bounce pass originating from the head. Once you have passed the ball you rotate 180 degrees so that your back is now facing the shooter. You prepare to recover the rebound from the shooter's shot. Once you see the ball make or miss quickly grab the rebound and come back to the marked line. You will wait there, facing the shooter, until I give you a clear audible signal. When you hear the signal this means that you should deliver the two-handed bounce pass originating from the head, to the shooter. You then immediately rotate 180 degrees to await the next rebound, and so forth. I will announce to you when the shooter has completed his 50th, 100th, 150th, 200th, 250th, and 300th shot so you know how many shots remain.

Finally, please avoid any type of direct communication with the shooter, as any such communication can corrupt the scientific validity of the study.

Do you have any questions? If so, please ask now because once the experiment has started I will not be able to answer any of your questions.

Thank you for your attention.

D Appendix: Supplementary tables and figures

D.1 Supplementary Analysis

In Section 4.2 we mention that the results of a two-sample proportion test (see Figure 2 for graph) could be driven by selection bias at the session level. To control for this possibility we estimate the marginal effect of having just completed a run of three or more hits on RC's probability of hitting the next shot using a linear probability model with session fixed effects. Under the null hypothesis, the indicator variable for hitting the previous three or more shots is a treatment that is assigned at random to the player. In the first column of Table 5, we present the coefficient corresponding to the marginal effect of hitting three or more shots in a row (Hit 3+) on the probability of hitting the next shot (p-value in parenthesis).[76] These marginal effects clearly corroborate the estimates presented in Figure 2, as well as the associated two-sample proportion test on their differences.[77] In column two of Table 5 we estimate the coefficients of a fixed-effects linear probability model with indicators variables corresponding to the five mutually exclusive shooting situations in the right panel of Figure 2 (Hit 2, Hit 1, $Miss$ 1, $Miss$ 2, $Miss$ 3+).[78] When controlling for these session level effects, the results of the proportion test are corroborated: RC still has a significantly lower shooting percentage in all five of these shooting situations, with the significant coefficients on Hit 2 and Hit 1, suggesting that this is a hot hand effect and not a cold hand effect. These effect size estimates are biased downward from the true effect for two reasons: (i) with player fixed (mean) effects, when Hit 3+ = 1 the player has performed at above mean levels in the previous three shots from the same finite dataset that the mean is calculated for, and therefore the effect size will have a downward bias, (ii) not all shots taken when Hit 3+ = 1 are taken in a hot state, and therefore the true difference between a hot state and cold state is underestimated.

[76] The p-values are computed from the distribution of the coefficients under the null, where the approximated distribution is generated to arbitrary precision via a session-level permutation of shots. We also estimated the p-value for the model using asymptotic assumptions and robust standard errors (if there are session level effects on a player's hit rate, errors will be heteroskedastic if shots are iid Bernoulli at the session level). The asymptotic p-values computed using robust standard errors were much lower then the approximation of the exact p-values under the null reported here, but with only 5 clusters, there is a potential for robust standard errors to be severely biased (Angrist and Pischke (2008), pp. 221-240).

[77] The marginal effects and their significance do not differ substantively under a logit model (which should be expected, Angrist and Pischke (2008), p. 103).

[78] Hitting three or more shots in a row serves as the base category.

D.2 Supplementary Tables & Figures

Table 3: *For each shooter in the three session panel, and in each phase (and overall), the hit percentage is reported after having hit 3 or more in row and after all other recent shot histories. The number of shots is reported in brackets.*

	Phase 1		Phase 2		Overall	
Shooter	Hit 3 or more	other	Hit 3 or more	other	Hit 3 or more	other
1†	63.9	55.5	66.7	58.5	65.8	57.5
	[61]	[236]	[141]	[453]	[202]	[689]
2	55.2	47.0	55.7	50.4	55.6	49.2
	[29]	[268]	[88]	[506]	[117]	[774]
3	50.0	49.2	59.2	58.0	57.1	54.9
	[38]	[258]	[125]	[469]	[163]	[727]
4	40.9	41.1	41.8	44.2	41.6	43.1
	[22]	[275]	[55]	[539]	[77]	[814]
5	55.6	44.4	47.7	47.4	50.5	46.5
	[36]	[261]	[65]	[529]	[101]	[790]
6	40.0	31.5	28.9	42.4	31.3	38.7
	[10]	[286]	[38]	[556]	[48]	[842]
7	59.5	51.8	61.0	52.2	60.6	52.1
	[42]	[255]	[100]	[494]	[142]	[749]
8	51.1	56.3	59.1	53.9	56.9	54.7
	[45]	[252]	[115]	[479]	[160]	[731]

†Shooter 1,"RC", had two additional identically conducted sessions. In these sessions his hit percentage was 72.5 [178] and 65.4 [416] respectively.

Table 4: *Linear probability model of the Panel's hit rate over in each Phase (with fixed session effects, permutation p-values equal proportion of permuted (session strata) data where coefficient exceeds the realized coefficient (one-sided))*

	Phase 1		Phase 2		Overall	
	Main	All	Main	All	Main	All
constant	.471	.519	.510	.541	.497	.533
	(.9142)	(.9590)	(.9283)	(.9681)	(.9753)	(.9942)
Hit 3+	.048**		.031**		.035***	
	(.0441)		(.0407)		(.0086)	
Hit 2		−.058*		−.021		−.032*
		(.0696)		(.1825)		(.0593)
Hit 1		−.053*		.001		−.017
		(.0533)		(.4287)		(.1370)
Missed 1		−.017		−.051***		−.038**
		(.2700)		(.0100)		(.0139)
Missed 2		−.083**		−.030*		−.047**
		(.0152)		(.0940)		(.0100)
Missed 3+		−.053*		−.064***		−.059***
		(.0535)		(.0034)		(.0013)

p-values in parentheses

50,500 Permutations (session strata)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ (one-sided, right)

Table 5: *Linear probability model of hot streak performance (with fixed session effects, permutation p-values equal proportion of permuted (session strata) data where coefficient exceeds the realized coefficient (one-sided))*

	Main Effect					Categories				
	RC	JNI6	Panel	GVT	JNI	RC	JNI6	Panel	GVT	JNI
constant	.609	.585	.497	.478	.622***	.678	.706	.533	.529	.587
	(.991)	(.971)	(.975)	(.864)	(.004)	(.996)	(1.000)	(.994)	(.991)	(.621)
Hit 3+	.069***	.120***	.035***	.050**	−.036					
	(.005)	(.001)	(.009)	(.014)	(.674)					
Hit 2						−.109***	−.068	−.032*	−.021	.044
						(.004)	(.113)	(.059)	(.213)	(.815)
Hit 1						−.055**	−.125***	−.017	−.069***	.071
						(.048)	(.007)	(.137)	(.009)	(.963)
Missed 1						−.046*	−.160***	−.038**	−.052**	.012
						(.083)	(.002)	(.014)	(.029)	(.303)
Missed 2						−.098**	−.121**	−.047**	−.008	.043
						(.021)	(.036)	(.010)	(.244)	(.587)
Missed 3+						−.078*	−.082*	−.059***	−.079***	−.031**
						(.059)	(.068)	(.001)	(.002)	(.014)

p-values in parentheses

50,000 Permutations (session strata)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ (one-sided, right)

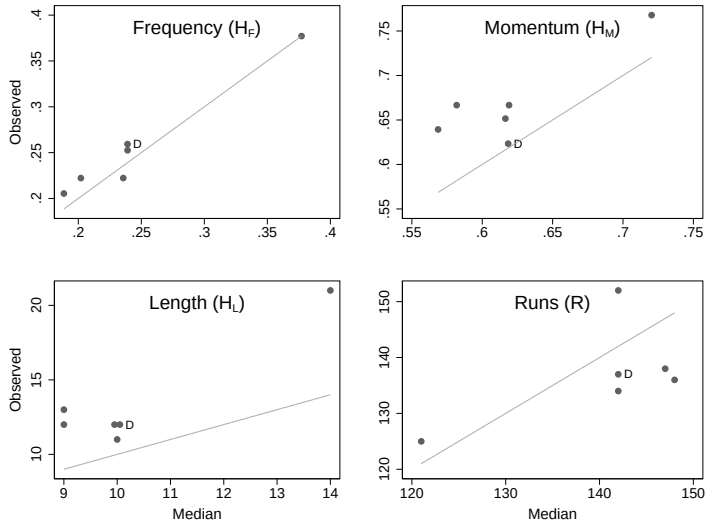


Figure 9: *Observed vs. median hit streak statistics for the player RC, where median is based on the exchangeability assumption (the session labeled was a doublesession)*