A Cold Shower for the *Hot Hand Fallacy*

*Joshua B. Miller and Adam Sanjurjo*

*Working Paper n. 518*

*This Version: June 13th, 2014*
A Cold Shower for the Hot Hand Fallacy

Joshua B. Miller and Adam Sanjurjo

June 13, 2014

Abstract

The hot hand fallacy has long been considered a massive and widespread cognitive illusion with important economic consequences. While the canonical domain of the fallacy is basketball, which continues to provide its strongest and most readily generalizable supporting evidence, the fallacy has been considered as a candidate explanation for various economic and financial anomalies. We demonstrate, in its canonical domain, that belief in the hot hand cannot be considered a fallacy. Our identification approach is to design a controlled shooting field experiment and develop statistical measures that together have superior identifying power over previous studies. We find substantial evidence of the hot hand, both in our study and in all extant controlled shooting studies, including the seminal study. In light of this discovery, we reexamine the evidence for the hot hand fallacy in other domains and reevaluate whether the hot hand fallacy is an economically meaningful cognitive illusion.

JEL Classification Numbers: C12; C14; C91; C93; D03.

Keywords: Hot Hand Fallacy; Hot Hand Effect.

---

*a: Department of Decision Sciences and IGIER, Bocconi University, b: Fundamentos del Análisis Económico, Universidad de Alicante. Financial support from the Department of Decision Sciences at Bocconi University, the Spanish Ministerio de Ciencia y Tecnología and Feder Funds (SEJ-2007-62656), and the Spanish Ministry of Economics and Competition (ECO2012-34928) is gratefully acknowledged.

†Both authors contributed equally, and alternate order of surnames across joint papers.

‡We thank José Valeiro, president of the Pabellón Municipal de Betanzos, for generously providing us with many hours of exclusive court access in order to prepare and conduct our experiments, Javier Lopez for his jack-of-all-trades assistance in organizing and conducting experimental sessions, Cristina Lopez for greatly improving the translation of our experiment’s instructions to Spanish.

§For helpful comments and suggestions we would like to thank Pedro Albarran, Jim Albert, Alberto Alesina, David Arathorn, Jeremy Arkes, Olivier Armanier, Daniel Benjamin, Alberto Bisin, Marco Bonetti, Gary Charness, Lucas Coffman, Vincent Crawford, Martin Dufwenberg, David Eil, Nicola Gennaioli, Uri Gneezy, Matt Goldman, Kyle Hyndman, Richard Jagacinski, Elliot Ludvig, Daniel Martin, Daniel Miller, Raymond Nickerson, Daniel Oppenheimer, Raffaella Piccarreta, Giovanni Ponti, Justin Rao, Colin Raymond, Aldo Rustichini, Angel Sanjurjo, Connan Snider, Joel Sobel, Charles Sprenger, Neil Stewart, Daniel Stone, Sigrid Suetens, Richard Thaler, Robert Wardrop and seminar participants at the University of Alicante, Bocconi University, BEELab 2013 in Florence, CAGE 2014 Conference at the University of Warwick ESA 2013 meetings in Zurich and Santa Cruz. All mistakes and omissions remain our own.

¶We thank Thomas Gilovich for locating the 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 data and having them converted to electronic format for us.
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

An individual believes in the hot hand if he or she thinks good outcomes have a tendency to cluster, either because a good outcome is perceived as more likely than usual after a streak of good outcomes, or because streaks of good outcomes appear as unusually long or frequent. This belief becomes a fallacy in an environment in which good outcomes arrive as if they had a constant rate of occurrence.

The hot hand fallacy was introduced in the seminal study of Thomas Gilovich, Robert Vallone and Amos Tversky (1985; henceforth GVT), in which they demonstrated that while observers, players, and coaches of basketball believe in a “hot hand” in shooting performance, the data fails to support these beliefs. In particular, the authors used evidence from in-game performance and controlled shooting experiments to demonstrate that observed sequences of shot outcomes were statistically indistinguishable from repeated realizations of a properly weighted coin, i.e. iid Bernoulli trials (Gilovich, Vallone, and Tversky 1985; Tversky and Gilovich 1989a,b). Subsequently, over the last three decades, the hot hand fallacy has earned the reputation of “a massive and widespread cognitive illusion” (Kahneman 2011), owing to the sustained divide between basketball professionals’ persistent and costly belief in the hot hand (Aharoni and Sarig 2011; Attali 2013; Bocskocsky, Ezekowitz, and Stein 2014; Cao 2011; Neiman and Loewenstein 2011; Rao 2009a), and the repeated inability of formal studies to discover a hot hand effect.¹

1The GVT study was a novel exhibit of how existing laboratory results on the systematic human misperception of randomness in sequential data (Tversky and Kahneman 1974; Wagenaar 1972)—results whose relevance had been questioned (Einhorn and Hogarth 1981; Lopes 1982; Morrison and Ordeshook 1975)—could be demonstrated in a natural setting.² GVT were also the first to identify

1Thaler and Sunstein (2008) summarized the hot hand literature in basketball with the statement: “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.” More recent studies of in-game data present clear evidence that players do not shoot with a constant rate of success, even when controlling for shot difficulty (Arkes 2010; Bocskocsky et al. 2014; Yaari and Eisenmann 2011). While these studies have limitations (see Section 2.1), the performance variation they document is consistent with the possibility of substantial hot hand effects, and contrast with the results from the original GVT study.

²The existence of a phenomenon known as the gambler’s fallacy in the laboratory “predicted” the presence of the hot hand fallacy in natural settings (Bar-Hillel and Wagenaar 1991; Rabin 2002). These early laboratory studies

2
a setting in which biased beliefs were both costly and acted upon by experts in their domain of expertise. While it is not unheard of for professionals to make systematic mistakes, including in the domain of sport, the hot hand fallacy is exceptional because professionals typically respond to incentives, learn from experience, and correct their mistakes when they are made aware of them (Hakes and Sauer 2006; List 2003). Because the perception of patterns in sequential data is essential in many domains of economic decision making, and the strength and persistence of this apparent irrationality is so striking, the hot hand fallacy has been given considerable weight as a candidate explanation of various puzzles and behavioral anomalies that have been identified in the domains of financial markets (De Long, Shleifer, Summers, and Waldmann 1991; Malkiel 2011; Rabin and Vayanos 2010), sports wagering (Camerer 1989), casino gambling (Croson and Sundali 2005), and lotteries (Guryan and Kearney 2008).

As the central contribution of this paper we devise a novel empirical strategy and conduct an incentivized controlled-shooting field experiment that together overturn what has long been considered the strongest evidence in support of the hot hand fallacy—that from the canonical domain of basketball. In light of this finding, we reconsider the relevance of the hot hand fallacy more generally. Upon evaluating the related literature, we find that the original GVT result remains unique in its evidence of the hot hand fallacy because of the clarity with which the conflict between beliefs and performance is demonstrated: the hot hand beliefs of decision makers are clearly identified, and the hot hand effect is not detectable in a multitude of performance environments that the beliefs are based on. By contrast, studies that attribute anomalous phenomena to the hot hand demonstrated that when people are told that a sequence of outcomes is generated by repeated coin flips, they expect outcomes to alternate more than they actually do, and for the relative frequency of an outcome in a finite sequence to closely match its underlying probability of occurrence. These behaviors are thought to lead to the gambler’s fallacy, in which people attach undue significance to streaks which happen to be typical of a random device, and strengthen their belief that a streak will end, the longer the streak. Further, when people are unsure about how data is generated (as in basketball), streaks that are typical of a random device lead people to incorrectly reject the possibility that the data is random, presumably because they instead believe that there is some persistent mechanism behind the streak. See Rabin and Vayanos (2010) for a model of this behavior.

There is recent evidence that NBA players can indeed be considered experts who optimize their shooting decisions (Goldman and Rao 2010).

Notably, in the domain of sport, Hakes and Sauer (2006) confirmed the existence of systematic and costly expert mistakes in the sport of professional baseball (the Moneyball anomaly [Lewis 2004]), and then demonstrated that experts corrected these mistakes as they became aware of them.

Indeed, we know of no other example in which experienced professionals exhibit a similar degree of irrationality.

Presuming that there is no hot hand effect, there is abundant evidence that hot hand beliefs have a costly impact on decision making, e.g. “hot” players take more difficult shots, coaches allocate more playing time to “hot” players, and opposing teams allocate more defenders to “hot” players (Aharoni and Sarig 2011; Attali 2013; Bocskocsky et al. 2014; Cao 2011; Rao 2009a).
fallacy in other domains, from gambling to financial investing, do not provide as strong a test as the original GVT study, both in terms of the identification of beliefs and the demonstration that beliefs are fallacious (see Section 5.1). Thus, the present finding that the hot hand effect exists in the domain of basketball, and in the original data of GVT, is central to the assessment of the robustness of the hot hand fallacy more generally.\footnote{In fact, GVT were careful to make this point and limit their conclusion that there is no hot hand effect to the domain of basketball because the hot hand fallacy was the general phenomenon of interest, and the evidence for the associated mistaken beliefs had been validated (Gilovich et al. 1985; Tversky and Gilovich 1989b).}

In Section 2 we present the design of our study, first motivating our approach with an account of the limitations inherent in the study of in-game shooting data. We focus on the identification issues that are particular to basketball shooting data. The general issues are standard and have to do with omitted variables and selection bias, which can lead to a spurious attribution of clustering in performance to a hot hand effect, or can mask it. Having recognized some of these issues, GVT conducted a controlled shooting experiment to complement their analysis of game data. While we agree with the spirit of this approach, we go further by concluding that the myriad confounds present in games actually make it impossible to identify or rule out the existence of a hot hand effect with in-game shooting data, despite the rich nature of modern data sets. In accepting this, one is forced to conclude that there has never been conclusive evidence for or against the hot hand fallacy based on in-game shooting data.

In contrast to in-game shooting, a controlled shooting experiment with expert players can provide the opportunity to identify, or rule out, the hot hand effect. In Section 2.2 we describe the field experiment we conducted with a group of Spanish semi-professional basketball players.\footnote{There are many studies that make an important contribution to the study of streaks and momentum in human performance, but they have had little influence on the hot hand fallacy literature because the beliefs and behavior of decision makers have not been identified (Alter and Oppenheimer 2006). A large literature exists on the presence or absence of the hot hand effect in other sports besides basketball, for a review, see Bar-Eli, Avugos, and Raab (2006), for a meta-analysis, see Avugos, Köppen, Czienkowski, Raab, and Bar-Eli (2013b). In the finance literature there is mixed evidence of a hot hand effect (or “performance persistence”) among fund managers (Carhart 1997; Hendricks, Patel, and Zeckhauser 1993; Jagannathan, Malakhov, and Novikov 2010).}

We improve on earlier controlled shooting tasks by removing design features that introduced confounds, such as players betting on their own shots and changing locations after each shot. In addition, we collect three times as many shots in each shooting session when compared to the original study of GVT, increasing the power of statistical tests. As an illustration of this increased power, when testing for momentum, our design is expected to generate a sample of more than 30 observations per

\textit{In the taxonomy of Harrison and List (2004) this is a framed field experiment.}
session for testing purposes, whereas GVT’s design (a single session study) is expected to generate a sample of only 12. Finally, we conduct multiple shooting sessions with each player, six months apart, which not only increases the power of tests for the existence of the hot hand effect, but also allows us to test whether the effect can be predicted out-of-sample.

In Section 3 we describe our empirical strategy, the types of hot hand mechanisms it is intended to capture, and how it improves on the work of GVT and others. We introduce a novel package of statistical measures that allow us to test for the existence of patterns in shooting performance that are more tightly related to the type of shooting typically associated with hot hand—streak shooting. In particular, our statistics measure the degree to which hit streaks (consecutive successful shots) have a tendency to persist, and whether hit streaks are longer or occur more often than should be expected given a player’s overall hit rate. Because these measures can be defined symmetrically for miss streaks, we can distinguish the “hot hand” from the “cold hand.” This makes our measures more capable of identifying hot hand effects than the statistics used in previous studies, which typically only measure patterns consistent with serial correlation or excessive volatility in performance. While previous studies have not found significant evidence of these patterns, this is consistent with the possibility that shooters exhibit positive serial dependence following streaks of hits, and negative serial dependence following streaks of misses. Moreover, even if significant serial correlation or excessive volatility had been detected, they could have been caused by a “cold hand” departure from normal shooting, without “hot hand” shooting having played any role. Thus, in order to improve identification, we account for these important issues when devising and interpreting our package of hit streak statistics. In addition, the test procedure we outline allows for a new approach in analyzing the shooting data from earlier controlled shooting studies.

In Section 4 we present our results. We find clear evidence of hot hand shooting at the individual level, and with substantial effect sizes. This is the first evidence of the existence of the hot hand, and more importantly, in individual shooters. In our individual-level analysis we place special focus on two expert shooters, one in our sample of eight, “RC”, who participated in six 300-shot sessions, and another, “JNI6”, from the study of six shooters by Jagacinski et al. (1979; henceforth JNI), who participated in nine 60-shot sessions. Because RC shot in two phases, six months apart, we were able to test whether the hot hand can be predicted out of sample. We find that it can be;
RC shows strong evidence of the hot hand in both phases, and because he participated in multiple sessions in the second phase, we observe that his hot hand systematically appears in nearly all of his sessions. Similarly, for JNI6, we find that his hot hand also systematically appears in nearly all of his nine sessions. In addition, we re-analyze the GVT dataset using our hit streak statistics, and find that the number of players with performance patterns in the direction predicted by the ‘hot hand’ hypothesis is significantly greater than would be expected if there were no hot hand effect, and that the size of this effect is substantial.

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 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 in a much smaller group, who were selected only on the basis of availability. In Section 2.1 we explain why any attempt at identifying the hot hand in an individual shooter with game data is inherently limited, because once one controls for even a subset of confounding factors—which neither of the Vinnie Johnson studies do—individual shooting data becomes too sparse for tests to have sufficient statistical power.11

Beyond the clear evidence of individual level hot hand effects is the unexpected finding of an average hot hand effect in a pooled analysis of shooters from all extant controlled shooting studies. Because basketball is a team sport with a high degree of specialization, there is substantial heterogeneity among shooters, and one should therefore expect heterogeneity in the hot hand effect. This is reflected in controlled shooting studies—many shooters actually perform worse after recent success, and therefore an average hot hand effect appears unlikely to emerge when pooling their shooting data with that of hot players. Instead, we find that the hot hand exists as an average effect.

11 In a recent study, Bocskocsky et al. (2014) used an extensive set of control variables that are generated by STATS Inc.’s SportsVu optical tracking system, which necessitated that they pool data from more than 300 players to have sufficient statistical power. In an earlier study, Rao (2009a) performed an individual-level analysis of 8 players of the Los Angeles Lakers in 60 games selected from the 1997-1998 season, using video of the games to manually code for a subset of variables that influence shot difficulty. Using these variables as statistical controls, a hot hand effect was not detected in any of the 8 players.
in our panel of shooters. Moreover, this effect is significant in both phases of the study, separated six months in time. Even more remarkably, when using our statistical measures to analyze the GVT data, which has long been considered conclusive evidence against the existence of the hot hand, we find that the hot hand is an average effect there as well. Finally, after also finding the hot hand in JNI’s data, we pool all extant controlled shooting data—ours, JNI’s, and GVT’s—and find that hot hand is an average effect in a pooled sample of all shooters from all extant controlled shooting experiments.\(^\text{12}\)

To put our pooled results into perspective, it is instructive to contrast them with the results from recent studies of in game data that have leveraged the increased size and richness of modern datasets. These studies have introduced control variables for an extensive set of within game factors that are known to confound the identification of the hot hand effect. They have found that in-game shooting exhibits significant clustering in performance, with players shooting better (worse) than predicted after they have recently shot better (worse) than predicted (Arkes 2010; Bocskosky et al. 2014; Yaari and Eisenmann 2011).\(^\text{13}\) While these findings are consistent with hot hand shooting, they have two limitations: (1) they do not control for omitted variables that may change a player’s probability of success from one game to the next (see Section 2.1), and (2) if one is willing to assume (1) is not an issue, it is still not clear if the observed effect is being driven by hot hand, or instead by cold hand shooting (with the player suffering a drop in ability after recent failure). By contrast, when one studies data from controlled shooting experiments, and uses appropriately averaged versions of the hit streak statistics we propose, one can analyze pooled data free of the selection bias issues present in game shooting, and can also be confident that the abnormal clustering in shot outcomes that is observed is indeed driven by periods of hot hand shooting, and not by some other factor.

Our body of results convincingly establish that, contrary to the conclusions drawn in the previous literature, but consistent with the ubiquitous belief among professional players, coaches, and

\(^{12}\)Surprisingly, only one other controlled shooting experiment, Avugos, Bar-Eli, Ritov, and Sher (2013a), has been cited in the hot hand literature. Their design was a close replication of GVT’s, but with fewer shots per session (40). The authors declined to make their data available. Wardrop (1999) provides a case study involving a single shooter, but after personal communication with the shooter who conducted the study herself, we viewed it as not having sufficient control to be included in our analysis.

\(^{13}\)The Arkes (2010) and Yaari and Eisenmann (2011) studies consist of a pooled analysis of pairs of unguarded foul shots (“free-throw” shooting). While this data offers a natural control for nuisance variables such as shot location and defensive pressure, it is still vulnerable to substantial selection bias arising from the data having to be pooled across a large number of games with different characteristics (see Section 2.1 for details).
lay-people alike, substantial hot hand shooting exists, and even appears to be a general property of the average shooter in all extant controlled shooting tasks, including the original GVT study.\footnote{Consistent with the previous literature, our subjects unanimously reported believing in the hot hand in a post-experiment questionnaire.} These results overturn the previous evidence for the hot hand fallacy in the basketball domain. A natural question arises: does this constitute evidence of a hot hand effect in games that can be detected by expert decision makers? In our view, the evidence suggests that we should not remain agnostic. Recent pooled analyses of in-game data have found a significant positive serial dependence in shooting performance (Arkes 2010; Bocskocsky et al. 2014; Yaari and Eisenmann 2011); while effect sizes in these studies appear modest and in game data does have limitations, their aggregate nature may also mask substantial individual hot hand effects due to the infrequency of hot hand shooting (Arkes 2013) and heterogeneity among players. Further, we find evidence of individual (and average) hot hand performance that is substantial in all extant controlled shooting designs (which all differ). The fact that shooting in a controlled environment is the same physical act as in game shooting, and that the hottest shooter in our panel, RC, was ranked as such by his teammates in a post-experiment questionnaire, based only on previous playing experience with RC, suggests that the hot hand effect is robust across shooting environments, including games, and that players can recognize it.\footnote{Without knowing what the experiment was about, or observing RC’s shooting sessions, the players were asked (in simple language) to rank the shooters by how much it was believed their conditional shooting percentage would exceed their unconditional shooting percentage, immediately following a sequence of three or more made shots.} Finally, in a game environment, coaches and players, who know each other well, have access to more than the outcomes of a “hot” player’s previous shots; in conjunction with this, they see the player’s shooting mechanics, ball trajectory, body language, and other signals of the player’s underlying mental and physical state.\footnote{Bill Russell, in his 1969 retirement letter, after 13 years as a player and 3 years as a player/coach (with 11 championships), wrote: “But nobody keeps statistics on other important things—the good fake you make that helps your teammate score; … the way you recognize when one of your teammates has a hot hand that night and you give up your own shot so he can take it” (Russell 1969).}

The implications of our results extend far beyond the domain of basketball, and present a serious challenge to the widely accepted notion of the hot hand fallacy as a massive and widespread, economically meaningful, cognitive illusion; we elaborate this argument in Section 5.2. Further, in Section 5.3 we elaborate on the portability of our results to a broader class of performance environments, and consider the implications for the study of momentum in human performance.
2 Design

We conduct a field experiment to test for the existence of the hot hand effect in basketball shooting. We first motivate why a controlled shooting experiment is necessary, then describe the setting and design of our experiment, and finally highlight novel design features with respect to the existing literature.

2.1 Why Controlled Shooting?

Challenging the hot hand fallacy in the domain in which the evidence has been considered the strongest necessitates the discussion of issues that are specific to the collection and analysis of basketball shooting data. In this section we detail why in-game data cannot provide conclusive evidence for or against the hot hand effect, and why a controlled shooting study is necessary.

If a careful researcher could control the conditions of a basketball game in any way desired, with the purpose of generating the best possible data with which to test for the existence of hot hand shooting, this researcher would observe expert shooters, shooting many times from the same location, with identical defensive pressure, with the same score in the game, surrounded by the same group of players, and with all other sources of shooting performance variability unrelated to hot hand shooting also held constant.

If this researcher does not control sufficiently for other sources of shooting performance variability, then the hot hand cannot be cleanly identified in game data, as any statistical test becomes vulnerable to false positives and false negatives. In the data a shooter may appear to have a period of hot hand shooting, but this may be because he is in the midst of a sequence of relatively easy shots, or because he shoots relatively better when the offensive strategy allows him to anticipate his shooting opportunities, the score of the game is lopsided, etc. Likewise, a shooting sequence may appear as if it were generated by a process no different than repeated coin flips but actually be generated by a shooter who, when hot, takes more difficult shots, is guarded by superior defenders, or passes the ball away relatively more often to take advantage of additional defensive pressure sent his way, etc. These concerns are not theoretical; previous studies have shown that when one does not attempt to control for shot difficulty, shooting performance exhibits negative serial dependence (Bocskocsky et al. 2014; Rao 2009a).
While the difficulty of controlling for the above mentioned within game confounds makes identification of the hot hand effect seem impossible, there are also across game confounds to account for, which guarantee that it is. In particular, because individual shooters do not shoot frequently enough in a single game for statistical tests to have sufficient power to detect abnormally streaky shooting, in-game shooting data must be pooled across a large number of games (Arkes 2010; Bocskocsky et al. 2014; Rao 2009a; Yaari and Eisenmann 2011).17,18 By pooling shooting data across games one introduces a selection bias that can easily lead to a clustering of shot outcomes that appears to be due to a hot hand, but actually has nothing to do with it; this bias can arise when periods of better-than-usual shooting performance are “selected” due to variables that are impossible to simultaneously control for, such as variation in the quality of opponents, player health, which teammates are playing, standings, contract situations, shot technique, offensive and defensive strategies, etc. For example, one can consider free-throw data, where players have been found to shoot, on average, around 2 percentage points higher after hitting their first shot than after missing it (Arkes 2010; Yaari and Eisenmann 2011).19 When analyzing the seven most recent seasons of professional basketball (NBA) free throw data, we find that free throw performance varies substantially over the course of a season. Namely, players have a 2 percentage point higher success rate in the second half of the season ($p < .0001$), meaning that if a player has hit the first shot, it is more likely that his shot is selected from the second half of the season.20 Further, when controlling for the possibility of fixed games effects due to omitted variables, we find that the performance after a hit is not significantly better than after a miss in six out of seven seasons, with a difference of one percentage point or less in most seasons (though the difference is significant when all seasons are pooled together).21 Yaari and Eisenmann (2011) and Bocskocsky et al. (2014) control

---

17 Further exacerbating this sample size issue, is the fact that if hot players exist, one should expect to see fewer shooting opportunities from these players due to strategic adjustments by the defense (Dixit and Nalebuff 1991).

18 The low frequency with which shots are taken raises another issue not discussed here. Suppose a player becomes “hot” and hits a shot, but there are not enough recent shots to associate with it (or none at all); how can one detect the hot hand in this case? Without a method of measuring factors that are more tightly related to a player’s hot state—such as the quality of a player’s shooting mechanics, current physiological state (e.g. neural noise [Churchland, Afshar, and Shenoy 2006]), or current psychological state (Bandura 1982)—detection of hot hand shooting in games is impossible, despite the richness of modern shooting data (e.g. STATS, Inc.’s SportVU optical tracking data).

19 Similar pooling issues present themselves when analyzing data from the NBA’s three-point contest, but in this case, because data is so sparse, it must be pooled across years (Koehler and Conley 2003). The three-point contest has further issues arising from the fact that shot locations differ and performance incentives are non-constant.

20 Goldman and Rao (2012) document a factor that leads to between game variation in performance: on average players have a higher free throw percentage at home, but in pressure situations this reverses.

21 Game-level controls do not influence the 2005-2006 estimate of Arkes (2010). We collected data from 2005-2006 to 2011-2012, and extend up until the present an analysis similar to that conducted in Arkes (2010) (using a linear
only for within-game sources of variation that are unrelated to hot hand shooting, and therefore their results are similarly vulnerable to omitted variables that may lead to across game differences in performance.

Furthermore, once one begins controlling for within game confounds, let alone across games confounds, shooting data becomes too sparse to allow statistical testing on the individual shooter level, due to insufficient power. Instead, shooting data must be pooled not only across games, but also across players (Arkes 2010; Bocskocsky et al. 2014; Yaari and Eisenmann 2011). Even if one is willing to ignore the aforementioned limitations, when shooting data is pooled across players, identification of the hot hand is confounded by player heterogeneity; some players may perform relatively better after recent failure than after recent success, and therefore pooling their shooting data with hot players will dampen the estimated hot hand effect, and could even reverse it.\(^{22}\)

Notice that if the shooting environment in a game were controlled enough so that the hot hand were identifiable, then it would no longer be a game at all, but rather a controlled shooting task. Therefore, we conduct a controlled shooting experiment.\(^{23}\)

### 2.2 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, and whether the hot hand is an average effect in our pooled sample of 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, and whether any average hot hand effect in Phase One would re-occur. We first describe the setting and design, then highlight novel features with respect to previous work.

**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 probability model). We find that the effect size reported in Arkes maintains when all seven seasons are pooled together, without fixed game effects.

\(^{22}\)We find evidence of this effect in our data. There are players that perform better after a streak of failures than after a streak of successes, presumably because they make some sort of adjustment after a streak of failures.

\(^{23}\)In Section 5.3 we discuss what we can infer about hot hand shooting in uncontrolled game environments on the basis of our controlled shooting results.
players to participate in a scientific study of basketball shooting with financial incentives. 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. 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 the purpose of statistical testing). The experimenter then used a strip of masking tape to mark the shooting location from which that player would take all 300 shots. Next, the shooter was led to a closed room, where the experimenter read the instructions aloud as the shooter read silently. 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 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.

---

24 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 this year is the best team in the 5th category, for their region, so can move up to the 4th category next year if they elect to.

25 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.

26 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.

27 See Appendix A.1 for the shooter’s instructions.
Each of the 300 shot attempts went as follows: a trained rebounder held the ball from a fixed location near the basket, which was also marked on the floor.\textsuperscript{28} When the experimenter initiated a computer-generated tone, the rebounder passed the ball to the shooter in a precisely prescribed way.\textsuperscript{29} 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.\textsuperscript{30} 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. 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.\textsuperscript{31}

\textit{Phase One and Phase Two}

In Phase One each of ten shooters participated in single session consisting of 300 shots. In Phase Two we conducted multiple shooting sessions, six months after Phase One. Eight of the ten Phase One shooters were available to participate in Phase Two; we refer to these eight shooters as the panel.\textsuperscript{32}

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 by far 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 the basis of this evidence, 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

\textsuperscript{28}The rebounder was informed only of his own task, and that the shooter would shoot 300 times.
\textsuperscript{29}The rebounder first lifted the ball above his head to signal the pass and then bounced the pass to the shooter so that it arrived in the shooter’s hands without the shooter having to move from his location.
\textsuperscript{30}The shooters typically shot within 1-2 sec. after receiving the ball.
\textsuperscript{31}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 (approximately 35 minutes) was spent in physical movement for the shooters. In a post-experiment questionnaire our subjects reported below average levels of fatigue. Several commented that they shoot around this many times 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.
\textsuperscript{32}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 new shooters who were eager to participate were given his time slots. The data of all shooters who did not participate in both phases of the experiment are included in the pooled analysis in Section 4.3.
Discussion of our shooting design

Because GVT’s controlled shooting study is by far the most well-known in the literature, and other controlled shooting designs relate closely to GVT’s, we ease exposition in this discussion by comparing our design solely to GVT’s. Then, in Appendix ?? we provide a more complete discussion of the relative advantages of our design over each of its predecessors.

Our controlled shooting design improves on GVT’s in several ways: (1) 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 of a player’s hit probability varying solely due to shot location; (2) 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; (3) 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; (4) for each shooter we collect 300 shots per session, rather than GVT’s 100, which gives our tests relatively more statistical power; (5) 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. As such, we are able to test not only for the existence of hot hand shooting on the individual level, but also whether hot hand performance can be successfully predicted out of sample.

---

33 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). However, 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. Further, 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.

34 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.
3 Empirical Strategy

3.1 Hot hand statistics

If the hot hand effect is real, the underlying mechanism is likely based on (1) non-stationarity, i.e. a shooter shifting to a relatively higher performance state (due to, for example, factors leading to increased focus, attention, or muscle control [Churchland et al. 2006; Csikszentmihalyi 1988; Kahneman 1973]), (2) serial dependence, i.e. a shooter performing better following recent success (due to, for example, positive feedback leading to increased confidence [Bandura 1982]), or some combination of (1) and (2).

For this reason we introduce three statistics to measure patterns in shooting performance related to both mechanisms, and, importantly, which represent the patterns typically associated with hot hand shooting. Loosely stated, our statistics measure: (1) how often a player is on a “hot” streak (non-stationarity), (2) a player’s shooting percentage conditional on having a recent streak of success (serial dependence), and (3) the duration of a player’s most exceptional hot streak (non-stationarity). In addition, we employ a commonly used measure of first order-serial dependence, the number of runs.

Definitions

Because our statistics are defined in terms of streaks, we first define streak. Let $S$ be the set of shots 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. A hit streak occurs at shot $s$ if a player has just hit three or more shots, i.e. if $x_s = x_{s-1} = x_{s-2} = 1$. We choose this definition based on evidence that the intuitive sense of the emergence of a streak generally begins at the third successive event, and the fact that previous studies have focused on streaks of at least three hits as indicative of hot hand shooting (Carlson and Shu 2007; Koehler and Conley 2003; Rao 2009b); this threshold allows us to expect a player to have more than 30 shots per session in this category, given our design target of 50 percent hits.

The first conception of the hot hand, that hit streaks occur more often for a player than would be expected if that player had a constant hit rate, is captured by the hit streak frequency statistic, for a discussion of the patterns, see Gilovich et al. (1985); Gould (1989); Tversky and Gilovich (1989a); Wardrop (1999). For a summary account see Reifman (2012).

The results we report also hold if the threshold for a streak is set at four successive hits; setting the threshold higher leads to insufficiently powered tests, because fewer than 9 observations are expected from a 50 percent shooter.
\(H_F\), defined as the relative frequency of shooting situations that immediately follow a streak of hits:

\[ H_F := \frac{|S_H|}{|S|} \]

where \(|·|\) counts the number of shots in a set of shots, and \(S_H\) is the subset of shots \(s\) that immediately follow a streak of hits, i.e. \(S_H := \{s \in S : x_{s-1} = x_{s-2} = x_{s-3} = 1\}\).\(^{37}\) To illustrate, a shooter with a constant 50 percent hit rate is expected to take approximately 37 of 300 shots immediately following a streak of hits, yielding \(H_F = .12\).\(^{38}\)

The second conception of the hot hand, that a player’s conditional hit rate immediately following a streak of hits is better than would be expected for a player with a constant hit rate, is captured by the hit streak momentum statistic, \(H_M\), defined as the shooting percentage on the shot immediately following a streak of hits (i.e. three or more hits in a row):

\[ H_M := \frac{\sum_{s \in S_H} x_s}{|S_H|} \]

Before defining the last two statistics we must set some additional notation. Call a sequence of consecutive hits flanked by misses (or the start or end of the session) a run of hits \(H_0 \subset S\).\(^{39}\) Further, let \(H_0 \subset 2^S\) be the collection of all runs of hits, and for \(H_0 \in H_0\), let \(|H_0|\) be defined as the length of the run, i.e. the total number of shots in the run of hits. To illustrate, the sequence of shot outcomes 0011010111 has \(|S| = 10\) shots, and three runs of hits \(\{3, 4\}\), \(\{6\}\), and \(\{8, 9, 10\}\), with lengths 2, 1, and 3 respectively.

The third conception of the hot hand, that hit streaks are longer than would be expected if a player had a constant hit rate, is captured by the hit streak length statistic, \(H_L\), defined as the length of the longest run of hits\(^{40}\):

\[ H_L := \max_{H_0 \in H_0} |H_0| \]

Because it is impossible for the first three shots to immediately follow a hit streak, in the results section we report \(H_F := \frac{|S_H|}{|S|-3}\).\(^{37}\)

Under the assumption that a shooter has a constant hit rate of \(p\), we derive \(E|S_H^\ell|\), the expected number of shots which immediately follow a sequence of \(\ell\) or more consecutive hits. We find that \(E|S_H^\ell| = \sum_{k=\ell}^{\log_2 |S|} p^k[|S|-k](1-p)+1\).\(^{38}\)

More formally \(H_0\) is an run of hits if it consists of consecutive indices \(s \in S\) where \(x_s = 1\) for all \(s \in H_0\) and \(x_s = 0\) for \(s \in \{s_{min} - 1, s_{max} + 1\} \cap H_0^C\), where \(s_{min} := \min_{s \in H_0} s\) and \(s_{max} := \max_{s \in H_0} s\).\(^{39}\)

This statistic was suggested by Wardrop (1999), who approximated its distribution numerically and showed that it was an effective test for the existence of hot hand shooting when it is driven by a non-stationary process.\(^{40}\)
$$H_L := \max_{H_0 \in \mathcal{H}_0} |H_0|$$

The expected maximum run length in a 300 shot session from a shooter who has made half of his shots is $E(H_L) = 7.25$, with $H_L$ exceeding 10 in fewer than 5 percent of sessions.\(^{41}\)

For each of the hit streak statistics there is a symmetrically defined miss streak statistic. If a player has significantly large hit streak statistics, but miss streak statistics that are not significantly different than expected, this allows us to conclude that a player has the hot hand and not the cold hand (see Section 4).

As we will be testing multiple statistics, we must address the issue of multiple comparisons. Because the hit streak statistics are not independent of one another we do not use standard corrections (such as Bonferroni), which are too conservative.\(^{42}\) Instead, because each statistic is intended to measure a different dimension of the same underlying construct, we employ a composite statistic, $H$, equal to the first principal component of the three hit streak statistics, which we compute from the joint distribution of the statistics for a given player’s data (as generated by our permutation scheme defined below).\(^{43}\)

In addition to the hit streak statistics, we consider a statistic that appears prominently in the hot hand literature, the runs statistic, $R$, which counts the total number of runs of hits and misses together:

$$R := |\mathcal{H}_0| + |\mathcal{M}_0|$$

where $\mathcal{M}_0 \subset 2^S$ is the collection of all the runs of misses, defined analogously to $\mathcal{H}_0$. This statistic is a measure of first order serial dependence, as the number of runs $R$, in $S$ trials, is the number of pairs of consecutive shots with different outcomes, plus 1.\(^{44}\) If it is more likely than expected that a hit follows a hit (or a miss follows a miss), the runs statistic will be lower than expected. The

\(^{41}\)Assuming all arrangements equally likely, we calculate this by using Theorem 1 from Mood (1940) to derive an asymptotic normal approximation of the distribution of $H_L$ (using a continuity correction). In our statistical tests below we instead numerically approximate the exact distribution of $H_L$.

\(^{42}\)If a player has a higher hit rate, each statistic will have a higher value. In fact, even for a player with a fixed probability of success, we find that the average pairwise correlation between the statistics is around .5 (as computed from the joint distribution under the null, generated by permutation using the methods outlined further below).

\(^{43}\)In practice this statistic weights each hit streak statistic nearly equally.

\(^{44}\)More precisely, $R := 1 + \sum_{s=1}^{\lfloor S/2 \rfloor - 1} [1 - x_s x_{s+1} - (1 - x_s)(1 - x_{s+1})]$. 

17
expected number of runs from a shooter who has made half of his shots, in a 300 shot session, is 
\[ E(R) = 150.5, \] with \( R \) falling below 136 in no more than 5 percent of sessions.\(^{45}\)

### 3.2 Statistical Test Procedures

Under the null hypothesis that a player is always the same shooter, the player’s shooting performance is a sequence of iid Bernoulli trials with a fixed probability of success. While a player’s true success rate is unknown to the experimenter, the shot outcomes are exchangeable, i.e. all orderings of the shot outcomes are equally likely under the null hypothesis. This means that for a single player’s sequence of 300 shots, an exact distribution exists for each statistic outlined above, by exchangeability. Given a particular player’s performance in a sequence of 300 shots, enumerating each permutation of the player’s shots, and calculating the value of the test statistic gives this distribution. While this enumeration 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.\(^{46}\) When we test for the existence of a pattern consistent with hot hand shooting and report a p-value for this pattern under the null exchangeability hypothesis, this is the p-value for the exchangeable model, but it also corresponds to the p-value of iid Bernoulli trials with an unknown and fixed probability of success.\(^{47}\) We report one-sided p-values throughout, as the alternative hypothesis of hot hand shooting establishes a clear ex-ante directional prediction.

\(^{45}\)Assuming all arrangements equally likely, there is a straightforward asymptotic normal approximation of the runs statistic (Mood 1940). The well-known Wald-Wolffowitz runs test uses this normal approximation. We instead use a numerical re-sampling scheme that allows us to approximate the exact p-values of each test statistic to any precision we choose (see Section 3.2). For the number of shots taken in a single session of the current study, the asymptotic normal approximation of the distribution of the runs statistic is adequate (the p-values it generates are nearly identical to the exact p-values). Nevertheless, when pooling a single player’s shooting data from different sessions the normal approximation cannot be used, as sessions are not independent.

\(^{46}\)Exchangeability is a sufficient condition for the permutation distribution of a test statistic to be exact and thus for a permutation test procedure to be an exact test (Ernst 2004; Good 2005).

\(^{47}\)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 \( c_{0.05,k} \) 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 \( c_{0.05,k} \) 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.
Discussion of our Empirical Strategy

Our empirical strategy improves substantially on earlier analyses of controlled shooting data by employing statistics that not only address both serial dependence and non-stationarity, but also, in a broader sense, more closely relate to the patterns of performance typically associated with hot hand shooting.\[^{48,49}\] In addition, because we can define analogous statistics for the symmetric patterns associated with cold hand shooting, we are able to \textit{separately} identify hot hand shooting and cold hand shooting, and therefore we can determine if significant variation in performance is being driven by the hot hand or the cold hand (or both).

The original study of GVT considered four statistics—(1) first order serial correlation, (2) the number of runs, (3) the conditional hit rate immediately following a streak of hits vs. immediately following a streak of misses, and (4) the variation in the hit rate in four-shot windows. We begin by noting that first order serial correlation, and the number of runs, both measure the same pattern in shooting, and therefore can be referred to interchangeably.\[^{50}\] We observe here that none of these four statistics are tightly related to what it means for a player to have a hot hand, and thus do not measure the patterns that are representative of hot hand shooting. The first order serial correlation (or runs test) does not measure hot hand shooting because (1) it is possible for a player’s shot outcomes to have a serial correlation of zero, and at the same time, for the player to be producing substantial hot hand shooting; this can happen if hit streaks exhibit persistence and miss streaks exhibit reversal,\[^{51}\] and (2) it is possible for a player’s shot outcomes to have substantial positive serial correlation without the player having a “hot hand”; this can happen if hit streaks neither persist nor reverse, but miss streaks exhibit persistence (“cold hand”). Similar identification issues are also present in GVT’s remaining two statistics. In the case of the conditional hit rate, by

\[^{48}\]Earlier studies include GVT, as well as the analysis of 3-point contest data by Koehler and Conley (2003), a recent replication of GVT by Avugos et al. (2013a), and the earlier work of Jagacinski et al. (1979).

\[^{49}\]Wardrop (1999) and Albert and Williamson (2001) considered the $H_L$ statistic and used a permutation procedure similar to ours to generate the null distribution. Albert (2008), in an analysis of in-game baseball data, considered a statistic that is highly correlated with our hit streak frequency statistic, the average length of all the runs of hits, $H_L^{mean} := \frac{\sum_{h_i \in H} |h_i|}{|S|}$.\[^{50}\]

\[^{50}\]In fact, we find that the first order serial correlation statistic ($\rho_1$) and the runs statistic ($R$) are perfectly linearly related in an experimental design where the shooter has made exactly half his shots, with $R = (1 - \rho_1)|S|/2 + 1/2$. This implies that across all permutations, $\rho_1$ and $R$ are perfectly correlated. When shooting performance deviates from the design target of 50 percent, the correlation between the statistics is still nearly perfect because $\tilde{R} := -2(|S| - 1)\hat{\sigma}^2\rho_1 + (|S| + 1)/2$, which is perfectly correlated with $\rho_1$, is a close approximation to $R$, with the absolute difference satisfying $|R - \tilde{R}| = 2\mu - 1)(x_1 + x_{|S|})$, where $\mu$ is the fraction of hit shots and $\hat{\sigma}^2$ is the standard deviation, both fixed across permutations.

\[^{51}\]In fact, this pattern of performance occurs in the prominently featured shooter, RC (see Section 4).
contrasting its value immediately following a streak of hits with its value immediately following a streak of misses (1) hot hand effects can be masked by a player’s ability to reverse miss streaks, and (2) a significant effect can be driven by the cold hand and not the hot hand. Finally, the variation in a player’s hit rate in a four shot window measures only the volatility of a player’s performance, which, if significant, could also be driven by periods of cold hand shooting.

While, to our knowledge, the identification issues that we address above have not been mentioned before in the literature, there are other important issues that have been pointed out in earlier work. Stone (2012) finds that a serious identification issue arises when one uses the serial correlation of binary outcomes to test for the serial correlation of the underlying probabilities which generate the outcomes. Because the observed percentage of a shooter’s hits is only a noisy measure of that shooter’s underlying ability parameter (hit rate), the sample serial correlation of observed shot outcomes is not only a biased measure of serial correlation in a shooter’s ability, but, as Stone found, it can be severely inconsistent. To illustrate, a relatively high underlying serial correlation in a shooter’s ability, (e.g. $\rho_a = .4$), can co-occur with a markedly lower sample serial correlation of shot outcomes (e.g. $\rho_s = .06$), even as the sample size gets arbitrarily large.

Stone’s work indicates that a test for the hot hand effect requires statistics that capture patterns in shooting data that are representative of hot hand shooting. Our statistics do precisely this, as they measure the patterns of performance that have typically been associated with the hot hand: (1) a player’s hit streaks being more frequent than what is expected, given that player’s historical performance, (2) a player’s shooting performance, following a streak of hits, being better than what is usual for that player, and (3) a player’s exceptional hit streaks being longer than what is expected, given that player’s past performance (Gilovich et al. 1985; Gould 1989; Tversky and Gilovich 1989a; Wardrop 1999).52

Aside from identification issues, that statistical tests often lack power when applied to binary data is another serious issue. The power of a statistical test, of course, is jointly determined by the statistic and the number of shots in the sample. Limitations on both dimensions have hampered the power of tests in previous studies. Regarding the selection of statistics, many authors have made clear that the statistical analysis employed by GVT, as well as similar analyses of binary data in

52We do not identify the mechanism that can generate these patterns; two candidates are (1) feedback of shot realizations into hit rates, e.g. from confidence, and (2) non-stationary shifts in a players hit rate, e.g. from fluctuation in a players mental or physical state.
other areas, are severely underpowered, and therefore unlikely to detect deviations from the null model (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). Regarding the number of observations in a sample, power will always be limited when studying a shooter’s data because the number of shots that a shooter can take under near identical conditions in a single session is limited. A further mitigator of the power of tests in previous analyses, such as GVT, is that they were unable to accommodate pooled data. By contrast, our controlled shooting design collects a greater number of shots per session than previous designs, and our empirical strategy not only allows us to perform tests on an individual shooter, by pooling shots across sessions, but it also allows us to conduct tests with data pooled across players.  

53 We take advantage of our ability to pool data across shooters—not only to increase the power of tests on GVT’s dataset, but also in our testing of a composite dataset, consisting of the data from all extant controlled shooting experiments (Stern and Morris 1993).

4 Results

In Section 4.1 we provide a brief summary of overall performance in our panel of shooters. In Section 4.2 we test if the hot hand effect exists at the individual level, and whether it systematically re-occurs across time. In an analysis of the shooting data from our own study, as well as the data from all extant controlled shooting studies, we find that certain individuals get the “hot hand,” and that the effect is systematic and substantial; shooter hit rates increase by between 8 and 40 percentage points immediately following a streak of hits.  

54 In order to examine whether the hot hand is an average effect among players, in Section 4.3 we extend our analysis to pooled shooting data. Because our experimental paradigm allows us to control for session and day-level variations in performance, our analysis is not vulnerable to the unavoidable endogeneity issues (selection bias) that arise when analyzing in-game shooting data.  

55 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.  

54 To put these numbers into perspective, the difference between the highest and lowest field goal percentages for NBA players in the 2012-2013 season was 25 percent (nba.com).  

53 Recall that this can lead to the spurious attribution of clustering in hits and misses to within-game variation in player performance.
In our study, we find an average hot hand effect among members of our panel in Phase One of our design, and then again in Phase Two, six months later. Moreover, we find an average hot hand effect in every extant controlled shooting study (including GVT), and again when we pool the shooting data from all studies together.

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: in a post-experiment questionnaire the average reported level of fatigue by players was less than 5, on a scale of 1 to 10. Further, the average shooting percentages in the first 150 and second 150 shots were 49.7 and 50.4 percent, respectively, and the average of the session-level differences 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, and the average of the session-level differences for each of the three possible comparisons was not significant.56,57

4.2 Identification of the hot hand at the individual level

Our Study

Figure 1a reports the hit streak statistics for 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 figure, stood out in Phase One because he had the most extreme hit streak statistics among the shooters, and because his statistics were nearly significant in the single session.58 Further, in a multi-item questionnaire

---

56 A single player in the panel (not the player RC discussed below) performed significantly better in the first half of his sessions (54 percent) than during the second half (46.2 percent), but this same player also performed better on average in his final 100 shots (49 percent) than his middle 100 shots (46 percent).

57 For player RC, whom we discuss extensively below, across his 5 identical sessions (1500 shots total), his first half and second half performance was not significantly different (61.7 percent vs. 63.3 percent), and his performance in each 100 shot window did not differ significantly either (61.6, 63.4, and 62.6 percent, respectively).

58 The p-value was \( p = .08 \) (one-sided) for the composite hit streak statistic, \( H \) (the first principal component of the standardized values of \( H_F, H_M, \) and \( H_L \)).
administered to the team, RC was identified by his teammates as the player with the greatest potential for the hot hand, based on their experience playing with him in games. 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, 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). RC’s hit streak length ($H_L$) and momentum statistics ($H_M$) are greater than their respective medians under the null (gray lines) in every session; this would occur in around three out of every hundred studies, for each statistic, if the null hypothesis were true ($p = .031$, one-sided binomial test). The hit streak frequency ($H_F$) and runs ($R$) statistics are on the predicted side of the median in the majority of
Table 1: The average value of the hit streak statistics for the player RC and player 6 from JNI (p-values in parentheses).

<table>
<thead>
<tr>
<th></th>
<th>The player RC</th>
<th></th>
<th>JNI6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>session 1</td>
<td>sessions 2-5</td>
<td>All</td>
</tr>
<tr>
<td>Hit Streak Frequency ($H_F$)</td>
<td>.21</td>
<td>.27</td>
<td>.26</td>
</tr>
<tr>
<td></td>
<td>(.197)</td>
<td>(.335)</td>
<td>(.134)</td>
</tr>
<tr>
<td>Hit Streak Momentum ($H_M$)</td>
<td>.64</td>
<td>.69***</td>
<td>.68***</td>
</tr>
<tr>
<td></td>
<td>(.105)</td>
<td>(.009)</td>
<td>(.003)</td>
</tr>
<tr>
<td>Hit Streak Length ($H_L$)</td>
<td>12.00</td>
<td>14.25**</td>
<td>13.50**</td>
</tr>
<tr>
<td></td>
<td>(.132)</td>
<td>(.037)</td>
<td>(.019)</td>
</tr>
<tr>
<td>Total Runs ($R$)</td>
<td>136.00*</td>
<td>137.25</td>
<td>137.00</td>
</tr>
<tr>
<td></td>
<td>(.090)</td>
<td>(.442)</td>
<td>(.168)</td>
</tr>
</tbody>
</table>

p-values in parentheses
50,000 Permutations (session strata)
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ (one-sided)

sessions, but not to a significant extent ($p = .50$, one-sided binomial test). By pooling the shots from all of RC’s sessions together, 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. 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 ($H_L = 13.5$, $p = .019$; one-sided permutation test, session strata) and the hit streak momentum statistic is significantly larger, by 6 percentage points, than its mean under the null ($H_M = .68$, $p = .003$), while the frequency and runs statistics are in the predicted direction, but not significant ($p < .20$ for both).

While this is the first study to identify the hot hand—and more importantly, at an individual level—an important question is whether this effect is substantial. Figure 2 presents a bar graph which compares RC’s hit rate immediately following a streak of hits with his hit rate immediately

61As discussed in Section 3.2, 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.
62The 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.
63Controlling for multiple comparisons, the composite hit streak statistic, $H$, is highly significant for RC, with $p < .01$. 24
Figure 2: RC has a higher performance when on a streak of 3 or more hits in a row, than any other other shooting situation.

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 shooting situation he may find himself in (p=.002, two-sample one-sided test of proportions).64

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 which 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 see that RC’s conditional hit rate immediately following a run of at least three hits exceeds his conditional hit rate immediately following each of the five other recent shot histories (respective p-values for two-sample one-sided test of proportions: .019, .008, .047, .027, and .002). 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

64 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 and our $H_M$ test statistic. On the other hand, a relative strength of the $H_M$ test procedure over the these tests is: (1) one can separate from cold hand effects when used in conjunction with the symmetrically defined miss streak statistic, and (2) it is not vulnerable to session effects, because the permutation distribution is generated using session strata.
symmetrically defined to his hit streak statistics, are significant; they are not, corroborating that RC has the hot hand, and not the cold hand. By contrast, statistical tests used in previous studies, including conditional probability tests otherwise similar to our own, cannot separate between the hot hand and the cold hand in this way, which leaves them vulnerable to both false positives and false negatives.

Another confound that could be driving what appears to be a hot hand effect in the results of the conditional probability tests associated with Figure 2 is a selection bias at the session level. In selecting shots that are immediately preceded by a streak of hits, one may over-select shots taken on days when a player is shooting well, thus relatively high observed conditional hit rates may be driven by between-session variation in a player’s shooting ability rather than within-session periods of hot hand shooting. To control for this possible confound in Appendix B.1 we estimate a linear model of RC’s hit probability with session fixed effects and the results of the proportion tests reported in Figure 2 are corroborated, suggesting that this is, again, a hot hand effect and not a cold hand effect.

The analysis of RC’s shooting data demonstrates that an individual can have a substantial hot hand effect that systematically re-occurs across time and can be correctly predicted, either on the basis of observed performance in previous shooting sessions, or teammate perception of the shooter’s in-game performance.

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 based on their 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 above the median under the null for 7 out of 8 shooters ($p=.035$, one-sided 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, one-sided binomial test), while the hit streak length statistic is above median levels for 6 out of 8

RC’s significant hit streak statistics presented in Table 1 do not have this issue as they are computed first for each session then averaged across sessions with permutation strata set to the session level.
Evidence from earlier controlled shooting studies

With a substantial individual hot hand effect clearly identified in our data, using the statistical measures we have developed, it is natural to test whether these statistics also detect a similar degree of individual level hot hand shooting in earlier controlled shooting studies. We start with the study which has previously gone uncited within the hot hand literature (Jagacinski et al. 1979; henceforth JNI), as outside of our own study, it provides the richest dataset for individual level testing; each of JNI’s six subjects shoots in nine sessions of 60 shots per session.66

We focus here on one of JNI’s subjects, who we refer to as JNI6, because his hot hand performance is substantial, and systematically re-occurs in almost every one of his nine shooting sessions. Our analysis of JNI6’s shooting data is identical to that performed on RC’s. Figure 4a reports that in all nine of JNI6’s sessions the hit streak frequency statistic is above median levels (p=.002, one-sided binomial test), in eight of nine sessions the hit streak momentum statistic is above median levels (p=.02, one-sided binomial test), in seven of nine sessions the hit streak length statistic is

66The shooters were all collegiate or former collegiate-level players. 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” condition involved turning the lights off in the gym after the player shot the ball, and thus were not comparable to other controlled shooting studies (the JNI study was designed to investigate the role of movement and ball trajectory information in the prediction of shooting performance).
above median levels (p=.09, one-sided binomial test), while the runs statistic displays no particular (directional) pattern (p=.5, one-sided 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. For each hit streak statistic we average its value across all nine of JNI6’s sessions to provide a more representative test in which the magnitude of the effect in each session is weighted equally. The last column of Table 1 reports that each of JNI6’s average hit streak statistics is highly significant (p-values in parenthesis), which constitutes clear evidence that JNI6 has the hot hand. 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 substantial, 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 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 one-sided test of proportions).

For this test, we can also 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 categories of recent shot histories.\footnote{The respective p-values for the two-sample one-sided test of proportions are .107, .006, .001, .015, and .062} We can also perform an additional check in which we test JNI6’s miss streak statistics. We find that they are not significant, which further corroborates that JNI6 has the hot hand, and not the cold hand.

JNI6 is not the only player from the JNI study with significant hot hand effects. A second player has significant hit streak frequency and runs statistics (.05 level, one-sided test). The probability of 2 out of 6 players crossing this significance threshold is .03 (binomial test).

In Section 3 we mentioned that with only a single session of 100 shots per shooter, the individual-level statistical tests for the hot hand effect employed by GVT were severely underpowered. Nevertheless, there is evidence of the hot hand effect in their data, with a magnitude that would be substantial if their results were to maintain in a larger sample. In particular, the conditional hit rate of 8 out of 26 shooters increases by at least 10 percentage points immediately following a sequence of three or more hits. For four of these 8 shooters, the percentage point increase was 22, 25, 28 and 40 respectively—with the increases significant for all four of these shooters (.05 level, one-sided two-sample proportion test).\footnote{These effects are not being driven by between-session variation in performance. As with the player RC, to control for between session variation in performance we estimate a fixed-effects linear probability model of JNI6’s hit rate, which corroborates these results (see Appendix B, Table 3).} 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.\footnote{For the player with a 40 percentage point boost, 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.}

### 4.3 Pooled Analysis: Average Effects

The individual-level analysis reported in Section 6.1 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 always shoot with the same ability, and still others actually under-perform.\footnote{A reason why individual-level tests are even more extremely underpowered in the GVT data is that many of their players had an overall hit rate of 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.}
immediately following a streak of hits.\textsuperscript{71} 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 if there are or are not
individuals with the hot hand in the sample. On the other hand, a pooled analysis \textit{can} answer the
question of whether the hot hand is an average effect of the shooters in a sample.

By using data from controlled shooting studies, we can test for an average hot hand effect, free
of the many identification issues present in game data. In addition, our empirical strategy allows us
to separate hot hand from cold hand effects, which was not possible in previous studies (see Section
3).\textsuperscript{72} We find clear evidence of an average hot hand effect, across all extant controlled shooting
studies. This result is surprising because it indicates that the hot hand is not only a property of a
few rare players, but also a property of the average shooter.\textsuperscript{73}

Table 2 reports the average standardized hit streak statistics for all extant controlled shooting
studies, with GVT’s shooters in column 1, JNI’s shooters in Column 2, the shooters from our panel
in column 3, and all shooters together in column 4.\textsuperscript{74} When considering just the players in our
panel we find a highly significant hot hand effect. All three of the hit streak statistics, as well as
the runs statistic, are highly significant. Moreover, this effect is not driven by any single session; to
the contrary, Figure 3b shows that hot hand performance in Phase One predicted the presence of
hot hand performance in Phase Two (out of sample), six months later. In the first pooled analysis
of GVT’s shooters, we observe significant hit streak frequency and momentum statistics, and JNI’s
shooters exhibit significant hit streak frequency and the runs statistics.\textsuperscript{75} Finally, we pool together

\textsuperscript{71}We find significant individual-level evidence of under-performance after hit streaks for some of our shooters, as well as shooters in earlier studies.

\textsuperscript{72}With the lone exception of Wardrop (1999), who employed a statistic that captures one dimension of hot streak shooting. We used this statistic in our analysis, and called it $H_L$.

\textsuperscript{73}We have mentioned recent pooled analyses of in-game data, which also detect patterns in shooting performance that are consistent with an average hot hand effect (Arkes 2010; Bocskocsky et al. 2014; Yaari and Eisenmann 2011). While these analyses contrast with those in GVT and other earlier work, the detected patterns are also consistent with cold hand shooting, and, due to the potential for selection bias, may be related to neither (see Section 2.1 and Appendix ??).

\textsuperscript{74}As discussed in Section 3, each hit streak statistic in the pooled data is generated by first standardizing the session level statistic for each player, 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).

\textsuperscript{75}Here we address the issue of multiple comparisons the same way as in the individual level analysis; we use a composite hit streak statistic equal to the first principal component of the three hit streak statistics, which is significant for the GVT data ($p = .017$, one-sided).
Table 2: The average value of standardized (session-level) hit streak statistics in pooled data (p-values in parentheses).

<table>
<thead>
<tr>
<th></th>
<th>GVT</th>
<th>JNI</th>
<th>Panel</th>
<th>Pooled76</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit Streak Frequency</td>
<td>.42**</td>
<td>.24**</td>
<td>.51***</td>
<td>.29**</td>
</tr>
<tr>
<td></td>
<td>(.017)</td>
<td>(.038)</td>
<td>(.006)</td>
<td>(.013)</td>
</tr>
<tr>
<td>Hit Streak Momentum</td>
<td>.37**</td>
<td>.06</td>
<td>.48***</td>
<td>.24**</td>
</tr>
<tr>
<td></td>
<td>(.031)</td>
<td>(.341)</td>
<td>(.008)</td>
<td>(.027)</td>
</tr>
<tr>
<td>Hit Streak Length</td>
<td>.24</td>
<td>.13</td>
<td>.43**</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>(.109)</td>
<td>(.164)</td>
<td>(.021)</td>
<td>(.112)</td>
</tr>
<tr>
<td>Total Runs (R)</td>
<td>-.21</td>
<td>-.37***</td>
<td>-.58***</td>
<td>-.21**</td>
</tr>
<tr>
<td></td>
<td>(.144)</td>
<td>(.003)</td>
<td>(.002)</td>
<td>(.046)</td>
</tr>
</tbody>
</table>

p-values in parentheses
50,000 Permutations (session strata)
* p < 0.10, ** p < 0.05, *** p < 0.01 (one-sided)

all available shooting sessions (including ad-hoc sessions conducted with players not included in our panel), revealing an average hot hand effect among players, albeit with a modest .2 to .3 standard deviation increase in performance. This modest average effect size belies the substantial individual effect size we have observed in particular shooters.

Discussion of Our Results
We test for the existence of the hot hand effect in the only basketball shooting environment in which it can be identified or dismissed: a controlled shooting experiment. We find the first evidence of hot hand shooting on the individual level, in the form of two individuals for which the hot hand effect exists, is substantial, and can be predicted out of sample. Further, despite heterogeneity across players, we find an average hot hand effect that can be predicted out of sample in our panel of shooters. Similarly, we find that the hot hand is a property of the average shooter in every extant controlled shooting study, and again when we pool all data from all studies. This is the first evidence of hot hand shooting in a pooled analysis in which the many serious confounds present in-game data are sufficiently controlled for.

76The 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). Our pooled analysis of all studies includes ad-hoc 300-shot sessions that we conducted with players who did not participate in our panel (5 shooters).
5 Discussion

The canonical demonstration of the hot hand fallacy in the domain of basketball has long held as the strongest supporting evidence of the fallacy. Nevertheless, in this domain, we have demonstrated that there has never been evidence that the belief in the hot hand is a fallacy. Further, we have found that the best available evidence indicates that this belief is not a fallacy. In particular, all extant shooting data that can be used to test for the existence of a hot hand effect—including data from the original GVT study, as well as our own—reveals significant hot hand effects, and these effects are substantial for some players. In Section 5.1 we discuss what our results, in conjunction with those of in-game shooting studies, indicate more generally about the hot hand fallacy in basketball games. These results raise questions about the relevance of the hot hand fallacy in decision making more generally. Thus, in Section 5.2 we discuss why the existing literature related to the hot hand fallacy cannot fully answer these questions. Finally, in Section 5.3 we briefly review evidence in support of the hot hand effect in other domains of skilled performance.

5.1 Hot hand effect in games

A number of recent studies have found evidence of a possible hot hand effect in pooled analyses of professional in-game basketball shooting data, which in theory does speak to the hot hand fallacy, because we know that expert practitioners in this domain believe in hot hand shooting, and act accordingly. Arkes (2010) found evidence consistent with hot hand shooting in free-throw (dead ball) data, observing that a player’s success rate is higher following a hit than a miss. Yaari and Eisenmann (2011) then replicated this result with a slightly different analysis, and more data. Bocskocsky et al. (2014) use field goal (live ball) data, with the most extensive list of controls so far possible, and find that, in aggregate, shooting performance (net of expectations) is positively correlated with shooting performance in the previous four shots (net of expectations). These studies are consistent with the possibility of substantial hot hand shooting, and in direct contrast with the original results of GVT. However, in line with concerns that some of these authors themselves point out, and as we argue in Section 2.1, any analysis of in-game shooting, whether it be from the free throw line (dead ball) or field (live ball), will be sufficiently exposed to omitted variable bias, so as to make it difficult to conclude that observed performance clustering is in fact attributable to hot
hand shooting.\textsuperscript{77}

To the extent that hot hand shooting cannot be identified, or dismissed, with game data, but only with controlled shooting data, is at first glance disappointing because what we are primarily interested in is the possibility that, in professional basketball games, the hot hand exists and decision makers are able to detect it and respond to it, as they claim. While we believe we have no choice but to test for the hot hand effect in a sufficiently controlled shooting domain, fortunately, the evidence we present suggests that the hot hand effect likely exists in games and is detectable by decision makers, which is consistent with the results of the in-game studies just summarized. For one, that we find the hot hand effect to be robust across all extant shooting designs, which are all different, suggests that it would be present in other settings that involve the same physical act. Second, in our players’ questionnaire responses, following Phase One of our shooting experiment, they far and away (correctly) rank RC as the hottest shooter, without having observed his session.\textsuperscript{78} Finally, game situations involve coaches and players who know each other well, and therefore, in conjunction with recent shooting performance, have more granular information on which to form beliefs about a player’s current hot state.

5.2 Assessing the Hot Hand Fallacy

Among studies relating to the hot hand fallacy, GVT’s design is unparalleled in its ability to both identify the patterns (or lack thereof) in sequences of outcomes generated by a natural process, and to clearly elicit the corresponding beliefs of experienced observers and decision makers about these patterns (or lack thereof). The authors conducted three separate studies of patterns in basketball shooting performance in which they found that, from the perspective of an uninformed observer, hits arrived from any given player’s sequence of shot attempts as if the hit rate were

\textsuperscript{77}It has been pointed out that a direct effect of the hot hand on shooting performance in games may not be detectable due to strategic adjustments by the defense which lead to few shooting opportunities—but indirect effects of the hot hand may be detectable in improved team performance (Dixit and Nalebuff 1991). Aharoni and Sarig (2011) have found that team performance improves when a player has superior recent performance, suggesting that while the defense is able to adjust and prevent an individual shooter from exploiting his hot hand, this adjustment leaves the defense more vulnerable. It is possible that the defense is reacting to an erroneous hot hand belief and adjusting disadvantageously to the behavior of the offense, but because it has been shown that both offenses and defenses adjust to “hot” shooters (Aharoni and Sarig 2011; Attali 2013; Bocskocsky et al. 2014; Cao 2011; Rao 2009a), if one assumes that these adjustments cancel, this evidence suggests a presence of a hot hand effect in games.

\textsuperscript{78}Instead, their accurate prediction is based on an average 800 hours of previous practice and game experience playing with RC. That they use practice and game experience alone to accurately predict hot hand shooting out-of-sample, in our controlled shooting task, suggests further that there is significant transfer between environments; if a shooter tends to be hot in controlled shooting tasks, he or she probably tends to be hot in practice and games as well.
fixed. In addition, the authors objectively validated that fans and players believe in the hot hand, by conducting a survey with active basketball fans, and an incentivized betting task with basketball players. Furthermore, the studies which followed GVT demonstrated that this instantiation of the hot hand fallacy is unique in its generalizability, as it appeared to be both massive and widespread (Kahneman 2011). The belief in the hot hand has been shown repeatedly to influence high stakes decisions by expert players and coaches, who have strong financial incentives to have correct beliefs, and ample opportunity to eradicate mistaken beliefs through feedback, training, and advice. With the evidence in the crucial basketball domain now reversed, while it is still possible that the hot hand fallacy is massive and widespread in other domains, in our view, the remaining evidence supporting this notion is surprisingly limited.

Field evidence of the hot hand fallacy

In field settings it is difficult to provide as strong of evidence for or against the hot hand fallacy as was seemingly provided by GVT. To do so one must first pin-down the patterns in the outcomes of the data generating process that decision makers have beliefs about, then identify the beliefs that decision makers have about these patterns, and finally, test whether observed patterns and beliefs are in conflict. Moreover, to have evidence on par with that which existed in the basketball domain, one would ideally want to show that these erroneous beliefs are strongly held, by, for example, demonstrating that they are costly to hold, or influence the decision making of experts. These desiderata have not been met by any extant field study relating to the hot hand fallacy.

A large body of field evidence in the domains of casino gambling (Croson and Sundali 2005; Keren and Wagenaar 1985; Sundali and Croson 2006), lottery play (Galbo-Jørgensen, Suetens, and Tyran 2013; Guryan and Kearney 2008), sports wagering (Arkes 2011; Avery and Chevalier 1999; Brown and Sauer 1993; Camerer 1989; Durham, Hertzel, and Martin 2005), and financial markets (DellaVigna 2009; Rabin and Vayanos 2010) identify anomalies that have been attributed to the hot hand fallacy. In this discussion we focus on the domains of casino betting, lottery play, and sports betting, which are each amenable to a test of the hot hand fallacy because their choice alternatives are relatively simple and limited, terminal payoffs are realized in a finite and proximate horizon. Evidence that coaches and players make decisions based on hot hand beliefs can be found in Aharoni and Sarig (2011); Attali (2013); Bocskocsky et al. (2014); Cao (2011); Rao (2009a). With respect to the semi-professional players in our study, their responses to a post-experiment questionnaire reveal that they too believe in the hot hand.
and there is no feedback of individual decisions into the behavior of the data generating process (Sauer 1998; Thaler and Ziemba 1988). While it may be argued that participants in these settings are not representative of the general population, or that an individual’s behavior in these settings is recreational and not representative of his or her behavior in other settings, it has been shown that economic incentives do influence behavior in these settings (Kearney 2005) and economic theory can characterize aggregate behavior (Sauer 1998).

In the domain of casino gambling it is reasonable to assume that there is no predictability in the underlying process, and thus a gambler’s belief that a pattern of recent success signals future success (‘hot hand’) is, by definition, erroneous. In surveys, recreational gamblers have been found to

---

80In the finance literature anomalies such as the fund flow puzzle, as well as momentum and reversals in asset returns, have sometimes been attributed to the hot hand beliefs of investors. It is posited that these investors incorrectly extrapolate past returns, or over-infer from recent returns that there has been a shift in regime (DellaVigna 2009). However, the mechanisms behind these anomalies are difficult to isolate given the relative richness of financial market settings.

81Nevertheless, it is perhaps not entirely unreasonable to believe that predictable patterns may exist in casino game mechanisms. Roulette wheels, for example, are physical devices operated by a human hand, there is a conflict of interest between gamers and the casino, and casinos themselves switch roulette wheels after observing even relatively modest clustering of a single color in recent outcomes (Bar-Hillel and Wagenaar 1991). If a gambler believes the wheel is entirely random (i.i.d.), but does not bet accordingly, so long as the inconsistency is costly, it seems correct to call this a fallacy, but if the gambler has priors that the wheel may not be random, then determining if her beliefs

---
believe in the hot hand (aka “luck”), reporting that they will bet more when they feel they are on a hot streak (Alter and Oppenheimer 2006; Keren and Wagenaar 1985). Nevertheless, it cannot be inferred from survey data the degree to which these beliefs influence behavior. Moreover, the meaning of these beliefs is unclear in games that involve a degree of skill, which players may be uncertain about, and may fluctuate (e.g. blackjack). An actual casino gambling setting appears to be a more fruitful source of data, but in existing studies of gambling behavior, evidence of the hot hand fallacy has been relatively thin due to the limited granularity of available data. The betting changes that have been observed in response to “hot” streaks of success cannot guarantee if a player actually bets more when they are “hot” (i.e. that there are expected costs), and even if one is willing to assume that they do, the betting amount might increase simply because bets are positively correlated with the amount of money available to bet, independent of recent success (Croson and Sundali 2005; Sundali and Croson 2006). More recent evidence indicates gambler’s do not exhibit the hot hand fallacy in their betting patterns following a recent stroke of “luck”, Smith, Levere, and Kurtzman (2009) find that poker players bet more after losing than after winning, Narayanan and Manchanda (2012) find that slot machines players bet as if they follow the gambler’s fallacy for their own performance rather than the hot hand fallacy, finally Xu and Harvey (2014) find that online sports bettors migrate towards bets with more favorable odds after winning and towards bets with longer odds after losing.

In the domain of lottery play, as in casino gambling, a lack of predictability in the underlying process can also be safely assumed, and therefore only the beliefs and behavior of players must be studied. In lottery drawing it has been found that (1) relatively more money is placed on a number under a parimutuel pay-out scheme (implying lower expected payout) the more frequently that number has been drawn recently (Galbo-Jørgensen et al. 2013), and (2) “lucky” stores sell more

---

---
lottery tickets when they have recently sold a winning ticket, implying that players erroneously believe these tickets have a higher chance of winning (Guryan and Kearney 2008). The popularity of hot numbers in parimutuel games does indicate a fallacious belief in hot numbers, though the strength of belief cannot be identified for two reasons: (1) the game is simultaneous move, so conditional on winning, the cost of holding the belief is unknown, and (2) the chance of winning is close to one in one million. The lucky store effect also demonstrates a mistaken belief in the hot hand, but as in the parimutuel setting, the strength of the beliefs are not clearly identified: sales increases are driven by a combination of a substitution away from non-winning stores (unknown cost), and a general increase in sales, which could arise from increased purchases by existing players (costly) or a media and advertising effect that catches the attention of new customers (costly, but not related to hot hand beliefs).

In sports wagering markets the patterns of streak performance that bettors respond to cannot be assumed away as in the casino and lottery context. Therefore hot hand beliefs may be justifiable, so any analysis of these markets must investigate patterns in betting behavior relative to patterns in performance. Using three years of basketball point spread data (betting market “beliefs” about the difference in points scored between two teams in an upcoming game), Camerer (1989) demonstrated that for basketball teams currently on a winning streak, the point spread in the actual game exceeded the final market point spread less than half the time, indicating that the market is biased and believes that “hot” teams will exceed the spread more than they actually do. Nevertheless, Camerer found this bias to be too small to be profitably exploited. Brown and Sauer (1993) conducted a study confirming the analysis of Camerer, but finding that the data do not provide strong evidence of a hot hand fallacy. The authors analyze basketball betting market data in a way that jointly estimates hot hand beliefs and the hot hand effect. They find evidence of modest hot hand beliefs in market aggregates, and that they cannot rule out the possibility that these beliefs are based on a real hot hand effect in team performance. In a data set consisting of seven years of basketball point spreads, Paul and Weinbach (2005) find that for certain minimum win streak

---

85 Using a different approach and a different data set (football betting markets), Avery and Chevalier (1999) find that hot hand beliefs influence movements in the market point spread from the initial spread set by odds-makers to final spread set after their interaction with the bettors. The study was intended to identify bettor sentiment and cannot determine if the hot hand belief has an excessive influence on betting markets, or if odds makers do not believe enough in the hot hand effect when setting the opening betting line. Durham et al. (2005) also analyze movements in point spreads in football betting markets, but find evidence consistent with the gambler’s fallacy rather than the hot hand fallacy.
lengths, a betting strategy, which bets against the teams on a streak, will exceed chance levels and be borderline profitable, over the course of seven years. More recently, Arkes (2011) has found some suggestive evidence of a hot hand effect in team performance, in that realized point spreads appear to be sensitive to variability in recent success, over and above many alternative measures of the strength of the two teams playing.\footnote{Using a seemingly unrelated regression model with the regression equation specifications assumption that the factors determining the market point spread and the realized point spread were identical, Arkes (2011) found the market point spread to be more sensitive to a team’s hot streak than the realized point spread.} It appears that sports bettors are willing to accept odds that are less than fair for bets on teams that are currently on a winning streak, but bettors are correct that teams on a streak are more likely to win, and the perturbation in the odds are too small to be profitably exploited.\footnote{In American college football wagering markets Sinkey and Logan (2013) find that betting houses correctly set point spreads to account for the a hot hand effect in team performance, i.e. point spreads are fair and there is no evidence of a hot hand fallacy in college football betting markets. In American professional football wagering markets, Lee and Smith (2002) document a related regression fallacy, and find that participants overweight past performance against the spread. A betting strategy that bets on each game and places money on the team that in net has performed worse relative to the spread in previous games (and bets in proportion to how much worse this team has performed), will win at greater than chance levels in 5 out of 8 seasons (and overall). Moreover, this strategy is profitable in 4 out of 8 individual seasons. This evidence indicates that either odds-makers overweight observed performance against the spread when setting the line, leaving themselves moderately exposed but not exploited over the 8 seasons, or that the line was set to clear markets, and that bettors overweight this information. Because performance against the spread is a salient and relevant piece of information, and bettors have structural uncertainty, it is difficult to know what its information content is and whether bettors are behaving irrationally.} Further, as the beliefs of bettors are not measured directly, the mechanism that produces this phenomenon may not be a hot hand belief, but may instead be the salience of the usual media coverage and sports commentator opinion that accompanies a team that is performing well.

Finally, in contrast to the test of hot hand fallacy in the basketball domain, these field studies by and large all study the behavior of amateurs rather than experts, with less opportunity to learn from expert feedback, training, and advice. For example, amateur gamblers, bettors, and lottery players are generally less exposed to the type of advice, direction, and feedback that an expert basketball player is exposed to, so even if a persistent bias were observed among them (and were costly), it could not be considered as indicative of a deeply rooted cognitive illusion. Further, evidence in these domains can hardly be said to be representative of decision making more generally, because many of the participants likely view their behavior as recreational, and for those who do not, they are gamblers, and not necessarily representative of the population at large.
Lab evidence of the hot hand fallacy

A large body of early lab experiments documented the systematic biases present in subjects’ perception of sequences of random outcomes. These experiments, however, were subject to several important limitations related to external validity, such as the role of amateur participants, experimenter demand effects, and lack of incentives (Einhorn and Hogarth 1981; Lopes 1982; Morrison and Ordeshook 1975). While GVT’s basketball study addressed many of the limitations present in earlier studies, there is also a body of more recent laboratory studies, which focus on financial applications and theory testing, conduct incentivized experiments, and which can, in principle, provide clearer evidence of the hot hand fallacy than is possible with field data. The designs are varied, and involve subjects determining the probability that one data generating process, rather than another, produced an observed outcome (i.e. signal detection) (Massey and Wu 2005; Offerman and Sonnemans 2004), first observing several periods of a random walk process, and then determining the probability that the next (binary) outcome is a success (Asparouhova, Hertzel, and Lemmon 2009; Bloomfield and Hales 2002), or, similarly, predicting whether the next outcome will be a success, but only after first being given the option to pay for possibly useless prediction advice (Huber, Kirchler, and Stöckl 2010; Powdthavee and Riyanto 2012). In general, the same limitations noted with respect to early laboratory studies also arise in these newer studies. Results are sometimes consistent with the hot hand fallacy and other times not: Massey and Wu (2005) find that subjects generally over-react to signals of possible regime shifts when predicting the next outcome, but under-react to the same signals when giving a probability assessment (p) that a regime switch has already occurred. While Asparouhova et al. (2009) observe a pattern of behavior that is generally consistent with the predictions of a model by Rabin and Vayanos (2010), subjects do not know what the data generating process of observed outcome sequences is (are), so any beliefs are rationalizable (thus none can be said to be fallacious). Huber et al. (2010) find another experiment the authors inform the subjects that the data generating process is a sequence of flips from a fair coin, and in this case subjects' beliefs become more gambler’s fallacy intensive (consistent with Rabin and Vayanos (2010)) but also anticipate positive recency for shorter streak lengths, which is not consistent with the...
that subjects are willing to pay for advice from an “expert,” regarding what the next outcome of a computerized coin flip will be, particularly after the expert’s predictions have been correct for many periods in a row. Because the mechanism behind the computerized coin toss is opaque, and subjects are not informed as to how the expert makes predictions, their behavior can easily be viewed as a rational response to the presence of structural uncertainty, and therefore it is difficult to judge their decisions as fallacious. Further, as subjects observe more predictions and outcomes over time they purchase predictions considerably less often, which suggests that they rationally adapt as they acquire more information about the environment. Powdthavee and Riyanto (2012) run a similar experiment, but make it more transparent to subjects that experts’ advice is of no value. While the experimenters observe the same general pattern of behavior, they do not rule out the possibility that this behavior is caused by an experimenter demand effect also present in earlier studies: the task has subjects make decisions in a spare and repetitive environment, while being provided information that is highly salient, varies, and yet is irrelevant.92

Overall, we evaluate laboratory evidence as rich and informative, but not currently capable of providing as critical a test of the hot hand fallacy as has been performed in the basketball domain. The laboratory has not only produced a lack of relatively clear conflict between hot hand beliefs and performance, but observed effects have been mixed, have not been demonstrated to be robust to learning and feedback, are of questionable external validity, and have been committed by subjects who are amateurs.

5.3 Hot hand effect more generally

Part of the fame of the hot hand fallacy is due to how counter-intuitive the result is; consistent with the beliefs of expert players and coaches, it is natural to expect professionals to sometimes enter into relatively superior performance states. In fact, not only does our finding of a hot hand effect in basketball shooting accord with intuition, but it also agrees with existing literature on human performance more generally. In particular, when considering individual performance, whether in a motor task, or an intellectual task, it seems unsurprising that changes in states, such as expectancy model’s predictions. A field study of college football betting markets Durham et al. (2005) find almost an opposite pattern of results.92 Powdthavee and Riyanto (2012) consider this possibility of demand effects, positing that subjects may think: “I know that predictions contained within these envelopes are useless. But if they are really useless, then why would they be here in the first place?”
and confidence (Bandura 1982; Damisch, Stoberock, and Mussweiler 2010; Nelson and Furst 1972), attention and flow (Csikszentmihalyi 1988; Kahneman 1973), or neural noise (Churchland et al. 2006) can lead to clusters of altered performance. With the detection of the hot hand effect in our study, and in every extant controlled shooting study, it becomes natural to consider the degree to which the hot hand effect is also a general phenomenon in a broader class of performance environments.

In their seminal study, GVT were careful to restrict their conclusion of no hot hand effect to the domain of basketball shooting, in which belief in the effect was already well-established and known to be near-unanimous. The pairing of these results exhibited the strength of the hot hand fallacy, and suggested its more general relevance. (Alter and Oppenheimer 2006; Gilovich et al. 1985; Tversky and Gilovich 1989a, b). Nevertheless, because the possibility that highly motivated professionals have a tendency to exhibit performance clustering is of general interest, independent of the hot hand fallacy, a large and separate related literature has emerged from the original GVT study. While most of these studies pertain to sport (Bar-Eli et al. 2006), other domains are also studied, including financial management (Hendricks et al. 1993). The overall evidence for a hot hand effect has been mixed (Avugos et al. 2013b; Carhart 1997), but several studies have found evidence of possible hot hand performance among hedge fund managers (Jagannathan et al. 2010) and in the sport domains of professional tennis (Klaassen and Magnus 2001), bowling (Dorsey-Palmateer and Smith 2004; Yaari and David 2012), horseshoes (Smith 2003), darts and golf putting (Gilden and Wilson 1995), and more recently, in baseball (Green and Zwiebel 2013). While these studies do share some of the limitations discussed in Section 2.1—e.g. the pooling of data over extended time periods—they are nevertheless instructive, as they find evidence in the direction consistent with the hot hand effect.

That our discovery of the hot hand effect in its canonical domain accords with both intuition, and existing literature from other domains, further dismisses the notion of hot hand belief as necessarily a fallacy. Indeed, coaches of sports teams who must allocate the ball, managers of companies who must allocate resources, and researchers choosing how to allocate their effort across time can now feel more confident that, on occasion, they may be in the midst of, or observing, a

---

93The existence of streaks in professional baseball has been a topic of contention in the statistics literature; see Albert (1993, 2008); Albert and Williamson (2001); Albright (1993a, b); Stern and Morris (1993).

94There is also some concern that variable incentives may drive some of these effects, e.g. contest incentives can induce performance clustering (McFall, Knoeber, and Thurman 2009; Szymanski 2003).
period of unusually superior performance.\textsuperscript{95}

6 Conclusion

We test for the hot hand fallacy in the canonical domain of basketball, where the supporting evidence has long been considered the strongest, and most readily generalizable to other domains. Using a novel empirical strategy to study the best available data—including that from our own field experiment and the original study of Gilovich, Vallone, and Tversky (1985)—we find that contrary to nearly 30 years of research, the belief in the hot hand is not a fallacy. While this result establishes that belief in the hot hand must sometimes be correct, it is still possible (perhaps even likely) that the hot hand is often perceived when it is not in fact present.\textsuperscript{96} We leave for future work the important task of characterizing in which settings, and by which individuals (expert or amateur), hot hand beliefs are well-calibrated, and in which settings they are over (or under) stated.

\textsuperscript{95}There is some recent evidence that momentum affects organizational risk-taking (Lehman and Hahn 2013).
\textsuperscript{96}We suggest two reasons why one should expect people to over-weight recent success beyond what can be justified by hot hand shooting: (1) in a world in which people are unsure if there has been a change in performance state (regime shift), the gambler’s fallacy, a mistake for which the evidence is considerably stronger, implies that people will over-infer from recent success (Rabin 2002; Rabin and Vayanos 2010), (2) in a world in which coaches and players are presented with more information than can be attended to, yet must assess a player’s current physical and mental state without knowing which information is maximally diagnostic for determining if a player is hot, bounded rationality may make it adaptive to pay special attention to salient streaks and recent shooting performance (see Burns (2004) for a discussion of why this behavior can be adaptive, even in the absence of hot hand shooting).
References


Appendix: Experimental Procedures & Instructions

A.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.
Table 3: Linear probability model hit rate over for the player RC and JNI player 6 (with fixed session effects, permutation p-values equal proportion of permuted (session strata) data where coefficient exceeds the realized coefficient (one-sided))

<table>
<thead>
<tr>
<th></th>
<th>The player RC</th>
<th></th>
<th>JNI6</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Main Categories</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>constant</td>
<td>.609</td>
<td>.678</td>
<td>.585</td>
<td>.706</td>
</tr>
<tr>
<td></td>
<td>(.991)</td>
<td>(.996)</td>
<td>(.971)</td>
<td>(1.000)</td>
</tr>
<tr>
<td>Hit 3+</td>
<td>.069***</td>
<td></td>
<td>.120***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.005)</td>
<td></td>
<td>(.001)</td>
<td></td>
</tr>
<tr>
<td>Hit 2</td>
<td>−.109***</td>
<td>−.068</td>
<td>(.004)</td>
<td>(.113)</td>
</tr>
<tr>
<td></td>
<td>(.048)</td>
<td></td>
<td>(.007)</td>
<td></td>
</tr>
<tr>
<td>Hit 1</td>
<td>−.055**</td>
<td>−.125***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.048)</td>
<td></td>
<td>(.007)</td>
<td></td>
</tr>
<tr>
<td>Missed 1</td>
<td>−.046*</td>
<td>−.160***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.083)</td>
<td></td>
<td>(.002)</td>
<td></td>
</tr>
<tr>
<td>Missed 2</td>
<td>−.098**</td>
<td>−.121**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.021)</td>
<td></td>
<td>(.036)</td>
<td></td>
</tr>
<tr>
<td>Missed 3+</td>
<td>−.078*</td>
<td>−.082*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.059)</td>
<td></td>
<td>(.068)</td>
<td></td>
</tr>
</tbody>
</table>

50,500 Permutations (session strata)

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

B Supplementary tables and figures

B.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 RCs 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 3, we present the coefficient corresponding to the marginal effect of hitting three or more shots in a row (\textit{Hit} 3+) on the probability of hitting the next shot (p-value in parenthesis).\footnote{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 than the approximation of the exact p-values under the null reported.} We can see immediately that the
marginal effects in the regression corroborate the estimates presented in Figure 2, as well as the associated two-sample proportion test on their differences.\textsuperscript{98} In column two of Table 3 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 (\textit{Hit} 2, \textit{Hit} 1, \textit{Miss} 1, \textit{Miss} 2, \textit{Miss} 3+).\textsuperscript{99} 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 \textit{Hit} 2 and \textit{Hit} 1, suggesting that this is a hot hand effect and not a cold hand effect.

B.2 Supplementary Graphs

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{example_graph.png}
\caption{Observed vs. median hit streak statistics for the player RC, where median is based on the exchangeability assumption (the session labeled was a doublesession)}
\end{figure}

\textsuperscript{98}The marginal effects and their significance do not differ substantively under a logit model and we should not expect that they would (see discussion in Angrist and Pischke (2008), p. 103).

\textsuperscript{99}Hitting three or more shots in a row serves as the base category.
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))**

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

p-values in parentheses
50,000 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))**

<table>
<thead>
<tr>
<th></th>
<th>Main Effect</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RC</td>
<td>JN6</td>
</tr>
<tr>
<td>constant</td>
<td>.609</td>
<td>.585</td>
</tr>
<tr>
<td></td>
<td>(.991)</td>
<td>(.971)</td>
</tr>
<tr>
<td>Hit 3+</td>
<td>.069***</td>
<td>.120***</td>
</tr>
<tr>
<td></td>
<td>(.005)</td>
<td>(.001)</td>
</tr>
<tr>
<td>Hit 2</td>
<td>−.109***</td>
<td>−.068</td>
</tr>
<tr>
<td></td>
<td>(.004)</td>
<td>(.113)</td>
</tr>
<tr>
<td>Hit 1</td>
<td>−.055**</td>
<td>−.125***</td>
</tr>
<tr>
<td></td>
<td>(.048)</td>
<td>(.007)</td>
</tr>
<tr>
<td>Missed 1</td>
<td>−.046*</td>
<td>−.160***</td>
</tr>
<tr>
<td></td>
<td>(.083)</td>
<td>(.002)</td>
</tr>
<tr>
<td>Missed 2</td>
<td>−.098**</td>
<td>−.121**</td>
</tr>
<tr>
<td></td>
<td>(.021)</td>
<td>(.036)</td>
</tr>
<tr>
<td>Missed 3+</td>
<td>−.078*</td>
<td>−.082*</td>
</tr>
<tr>
<td></td>
<td>(.059)</td>
<td>(.068)</td>
</tr>
</tbody>
</table>

p-values in parentheses
50,000 Permutations (session strata)
* p < 0.10, ** p < 0.05, *** p < 0.01 (one-sided, right)