Master Thesis

William Möllestam

2022

He’s on Fire?

Strategic Decisions and Allocation Adjustments Under the Hot Hand Fallacy

Lund University

School of Economics and Management

Department of Economics

Supervisor: Erik Wengström

Abstract

This paper empirically explores 10 years of play-by-play data from the seasons 2011-2021 from National Basketball Association with a comprehensive dataset with 2.7 million shots to test for a hot hand, (i.e., predictability in future outcomes) using three separate models. I find proof for its existence in all short-term categories using panel data as well as when adjusting for shot difficulty using a linear regression prediction model. The measured values are significant where a hot hand equates to around a 0.70-0.77 standard deviation of player ability distribution. The findings are in contrast to most of the hot-hand research, which has typically reported either no hand or a very weak hand in sports when controlled shooting designs are used. This difference, I argue, is due to the lack of large datasets and the absence of the athletes' natural environment, both of which are necessary for creating a realistic setting for evaluating a hot hand. In my regression discontinuity design, I found that hot players remain rational by not taking more difficult shots. There does not seem to exist any endogenous defensive response, i.e., the closest defender distance remains unchanged, and the coach begins to modify strategy first after longer streaks (a sequence of seven hits or longer).

Keywords: Hot hand fallacy, law of small numbers, sequential decision making, selection bias

Contents

LIST OF FIGURES **IV**
LIST OF TABLES **V**
LIST OF ABBREVIATIONS **VI**

INTRODUCTION **1**
THEORETICAL / LITERATURE REVIEW **4**
HOT HAND FALLACY 4
FOUNDATION OF THE HOT HAND 4
ECONOMIC AND FINANCIAL CONTEXTS 5
SPORT CONTEXTS 6
RELATED LITERATURE 7
BASKETBALL 9
DATA **12**
DATASET 12
PRE-REGISTRATION 13
THE MODELS **14**
BAYES' THEOREM & CONDITIONAL PROBABILITY 14
Bias correction in the Hot Hand *15*
WEIGHTED LEAST SQUARES REGRESSION 18
LINEAR REGRESSION 20
REGRESSION DISCONTINUITY DESIGN 22
EMPIRICAL FINDINGS AND ANALYSIS **25**
WEIGHTED LEAST SQUARES ESTIMATES 25
LINEAR REGRESSION ESTIMATES. 26
DOES THE ENVIRONMENT DETECT A HOT HAND? 28
The hot player *28*
Teammates & opponent *29*
Coaches *30*
DISCUSSION **31**
CONCLUSION **32**
REFERENCES **33**
APPENDIX **37**
APPENDIX A 37
APPENDIX B 38
APPENDIX C 41
APPENDIX D 42

List of Figures

FIGURE 1: ILLUSTRATING DIFFERENCES IN SHOT ALLOCATION 11
FIGURE 2: REGRESSION DISCONTINUITY DESIGN.. 24
FIGURE 3: BIAS CORRECTED DIFFERENCE……………………… 25
FIGURE 4: DISTRIBUTION: MADE/MISS STREAK…………… 26
FIGURE 5 NUMBER OF SHOTS……………………………………………………28
FIGURE 6: SHOT DISTANCE 28
FIGURE 7: NUMBER OF TOUCHES…...29
FIGURE 8: CLOSEST DEFENDER……. 29
FIGURE 9: MINUTES PLAYED.. 30
FIGURE 10: POINTS PER GAME……………………………………………………..30
FIGURE 11: TOTAL SHOOTING PERCENTAG…………………...………………...41

List of Tables

TABLE 1: SHOOTING VARIABLE SUMMARY STATISTICS....13
TABLE 2: THE BIAS IN THE CASE OF COIN FLIPS....18
TABLE 3: SHOOTING PERFORMANCE FOR K=3....27
TABLE 4: PLAY-BY-PLAY SUMMARY STATISTICS....37
TABLE 5: TOTAL SHOOTING PERCENTAGE SUMMARY STATISTICS....37
TABLE 6: DO PLAYERS HAVE A HOT HAND K=1....38
TABLE 7: DO PLAYERS HAVE A HOT HAND K=2....39
TABLE 8: DO PLAYERS HAVE A HOT HAND K=3....40
TABLE 9: HIGHEST AUTOCORRELATION IN THE WLS....41
TABLE 10: NUMBER OF SHOTS (rdd)....42
TABLE 11: SHOT DISTANCE (rdd)....42
TABLE 12: PLAYERS TOUCH (rdd)....42
TABLE 13: CLOSEST DEFENDER DISTANCE (rdd)....43
TABLE 14: MINUTES PLAYED (rdd)....43

List of Abbreviations

FTA	Free Throws Attempted
FGA	Field Goals Attempted and
FG%	Field Goals percentage
NBA	National Basketball Association
OLS	Ordinary least squares
RDD	Regression discontinuity design
TS%	True Shooting percentage
WLS	Weighted least squares

Introduction*

Academics and policymakers are increasingly interested in the *hot hand fallacy*, whereas decision-makers tend to overlook randomness, assuming that the existing strategic decision will be extended in the near future owing to the systems' previously observed streak of successful achievement[1]. Several important events in recent decades have been blamed on poor strategic decisions. For instance, the primary cause of the global housing crisis in 2008, has been argued to be psychological in nature and that the fallacious concept *hot hand* strongly influenced the judgment and the decisions of elected officials, government regulators, financial firms, rating agencies and institutional investors. With an estimated growth of $150 billion in sports betting in U.S 2018[2], investment methods based on a hot hand may emerge where similar methods are already established in almost all software programs in the financial market, where it is a well-known axiom among Wall Street investors to buy stocks in the up-trending market.

The term *hot hand fallacy* was introduced in a seminal paper by Gilovich, Vallone & Tversky *(*1985*)* after the discovery that the hitting and missing streaks of Philadelphia 76ers players occur in random order when shot sequences were analyzed. Insecurity and chance are not natural human attributes. Consequently, when faced with uncertainty, humans seek order and symmetry and thus have a propensity to draw conclusions based on a few observations — "the rule of small numbers" (Kahneman & Tversky, 1971, p106). For over four decades, researchers have considered the hot hand as a fallacy in a variety of domains, including presidential primaries, stock trading, betting, and sports, and it is one of the most well-known theories in behavioral economics. In recent years, some researchers have questioned the hot hand effect. However, the original definition introduced by Gilovich, Vallone & Tversky (1985) more than 30 years ago, has never been fully challenged (Morgulev & Avugos, 2020). Miller & Sanjurjo (2015) observed however, that the true base rate of hits followed by another hit is lower than previous literature estimates, concluding that the concept has been misguided and falsely claimed, to be a fallacy.

* I would like to thank my supervisor, Erik Wengström, whose expertise was invaluable and brought my work to a higher level. I would also like to thank Tobias König and Thomas Giebe at Linnaeus University, for their valuable guidance throughout my studies. Lastly, a special thank you to Björn Johansson for your insightful feedback.

[1] One indicator of academic interest is the publication of the book *The Hot Hand: The Mystery and Science of Streaks* (Cohen, 2020).

[2] In May 2018, the supreme Court struck down a 1992 federal law to legalizing the estimated $150 billion in illegal wagers that Americans make every year, providing the leaders of America with a rare chance to build smart public policies from the ground.

With only a few exceptions in Basketball, the hot hand is typically evaluated based at the Three Point Contest of the NBA or in a controlled shooting design, where researchers often advocate that it be broadly adopted. However, this ignores additional physiological knowledge of the game (Mack & Stephens, 2000). For instance, Wood and Stanton (2012) argue that athletes in their natural environment place a higher importance on success than in a controlled design. Similarly, Morgulev et al. (2020) suggest that in the absence of face-to-face competition, it is unlikely to provide a realistic environment for examining a hot hand.

In this paper, I explore the hot hand fallacy in the true game environment with 2.715.094 shots from 10 NBA seasons with over 6.500.000 play-by-play data obtained by BigDataBall to capture every detail on all aspects of the game. A detailed analysis of game data may result in a more specific explanation of the concept, as well as a better understanding of how individuals or groups respond in the presence of a potential hot hand. My data set has two features that enable me to conduct a hot hand analysis of game data. Firstly, it provides shot location and the specific shot coordinates for every shot that has been taken which allows me to (1) use the total shooting percentage, that to my knowledge, no previous studies have used, and (2) account for variance between players, which means that if two players try the same shot, their chances of making it varies and an adjustment of the estimations can be made. Secondly, each in-game movement is tracked using a unique ID that contains critical information each time a player receives the ball and is active on the court. This enables the analysis of any allocation adjustments and changes in the teams' strategic decisions.

In simple terms, my approach is to use game data to challenge both the autocorrelation and the nonstationary claim separately. As noted before (Warpdrop, 1995, 1999), sequential performance should be evaluated using a variety of different statistical methodologies, depending on the nature of the claim, as they are often mixed up. I analyze the assumption of autocorrelation using panel data and a weighted least squares approach. To conduct an examination of the nonstationary assumption, I replicate the work of Gilovich, Vallone & Tversky (1985) by developing a model that predicts the shooting difficulty of each shot using linear regression in order to adjust the difficulty of each shot. Finally, I use multiple regression discontinuity methods to examine the environment around the hot hand. I prove that, on average, a hot player deviates around 0.70 and

0.77 standard deviations, causing a 50th percentile shooter to shoot like a 77th percentile shooter. Additionally, I observe that the hypothesis of endogenous responses in game situations is wrongly stated, as are the long-standing implications arising from this assumption. While the environment does recognize indicators of a hot hand, my results imply that is not to the extent previous researchers have argued.

The structure of this paper is as follows. In the next section, the conceptual foundation of the hot hand is laid forth, followed by a review of the most relevant literature on the subject. The section ends with a review of how basketball is played and a discussion of some of the criticism aimed at game statistics. Section 3 presents the data that has been used. Models that were applied can be seen in section 4. In section 5, I provide a discussion that links the results to economics, the results are presented in Section 6, as well as contrasted to the prior literature, and lastly, section 7 concludes.

Theoretical / Literature Review

This section aims to present the literature and theoretical review, covering where the Hot Hand fallacy originates from and the basics of the concept, as well as try to place it in the current debate. This is followed by the related literature and previous empirical findings, before finally relating the target to the specific case of basketball where I discuss in detail how the game is played for those that are unfamiliar with it, and address some of the criticisms aimed at game data analysis.

Hot Hand Fallacy

The *Hot hand fallacy* is the perception that if an individual or organization has had a previous streak of success, the tendency of success is anticipated to continue (i.e., recent past outcomes are indicators of future outcome, despite the independency of the two events). For instance, if one predicts a head three times in a row in a coin toss, a hot hand may be claimed. When this occurs, the individual believes their possibilities of properly predicting which side the coin will fall on next are greater than 50 percent.

Foundation of the Hot Hand

Alderman (1974) outlines the phenomena of momentum in his early 1970s book 'Psychological Behavior in Sport. Adler & Adler (1978) develop the concept of momentum further where they establish a link between a variety of psychological variables (motivation, effort, determination, confidence) and the subjective sensation of momentum ('being in the zone'). These psychological mediators are altered by competitive achievements and are hypothesized to improve the possibility of future success – especially those that are unexpected. Iso-Ahola & Mobily (1980) expand on the relationship between success and the psychological state of the performer where they coined the term psychological momentum, which proposes that recent success may alter a person's self-confidence, boosting mental performance (focus, thinking and attention), and thus increasing the amount of mental and physical effort. Psychological momentum has been used by academics across a wide range of fields to represent non-random distribution of events in a continuous process to detect short-term sequential dependencies, e.g., weather events, betting, and financial markets. Thus, if asked what momentum is, most individuals would reply in psychological terms rather than Newtonian physics terms. (Iso-Ahola, 2014). The hot hand and psychological momentum have similar mechanisms of action and have been thought to be the same phenomena in a large number of studies (Briki, 2017). Despite this, the hot hand has been extensively examined and debated in hundreds of articles, often without reference to the idea of psychological momentum.

Economic and Financial Contexts

The hot hand fallacy is fundamental to understanding how the economy operates and whether individuals make rational decisions[3]. Neoclassical economists believe markets are efficient and that products are correctly priced at any given period. This is based on the assumption of perfect information and that all individuals are rational. Behavioral economists, however, have challenged this notion where they throughout time have demonstrated that individuals often make irrational financial decisions when taking risks (Kahneman & Tversky, 1979). For instance, gamblers' tendency to exchange lottery tickets for additional tickets rather than cash is consistent with the hot hand phenomenon, while the same tendency can be observed in betting, where the better is more likely to place a second bet with the price money rather than withdraw the money he has already won (Croson & Sundali, 2005).

The cognitive bias is an important factor influencing investing choices in financial markets where investors frequently overshoot previous success, owing to the fact that investors and traders make decisions based on emotions. As was evident during the dot.com era, investors followed the markets previous success where they continued to buy high and sell low when the market crashed, whereas rational investors did the opposite (i.e., bought low and sold high) (Kahneman, 2011). In the financial market, investors increase their fund purchases or diversify their portfolios in expectation of increased returns (Capon et al., 1996). This is a common occurrence among investors who delegate decision-making authority to fund managers. Due to the hot hand effect, investors often purchase funds that have performed well in the past, anticipating that the

[3] In the financial markets, casino gambling, lotteries, and sports betting, the hot hand fallacy has been extensively studied as a potential explanation for a variety of problems and behavioral changes (Arkes (2010), Brown & Sauer (1993), Galbo-Jörgensen et al. (2016), Kahneman & Riepe (1998), Lee & Smith (2002), Shefrin (2009), Sinkey & Logan (2013), Xu & Harvey (2014) and Yuan et al. (2014)).

managers' performance will continue. However, given the inconsistency of fund performance, the hot hand may result in a biased decision (Goetzmann, & Peles, 1997). The fallacy assumes that a certain manager is hot, not a specific outcome. For instance, investors have a tendency to overestimate the importance that if a professional manager has picked lucrative funds in the past, the funds they choose will likely continue to be profitable in the future (Arkes, 2010).

Sport Contexts

In 1993, NBA JAM was released as an arcade game all over America. The game gave the words "He's on fire" a special meaning when the electronic voice of the announcer screamed it out when a player had hit three shots in a row. From that moment on, until the other team scored, that player's shots were never missed, literally leaving a trail of fire and smoke in their wake.

The phrase *hot hand*, however, was first used in a scientific debate by Gilovich, Vallone & Tversky (1985) a few years earlier, where they state that the hot hand is a myth and a cognitive illusion. The conclusion that current success is independent of past performance was based on Philadelphia 76ers chance of making a field goal, on the players' prior success or failure in earlier shots. Additionally, Gilovich, Vallone & Tversky study nine other professional players and their chance of making a free throw after a series of hits or misses. Due to the lack of substantial serial correlations for individual players, they establish that a hot hand did not exist. The remarkable finding that players' hitting, and missing streaks occur in random order have prompted a large body of research, but due to the rarity and complexity of the hot hand, (Hamberger & Iso-Ahola, 2004), some researchers argue that it is impossible to detect until big data are highly mathematically and statistically processed, and that serial reliance needs a microscopic analysis of its existence and effects (Doron & Gaudreau, 2014).

The original definition of the hot hand given by Gilovich, Vallone & Tversky (1985) can be interpreted in one of two ways: autocorrelation or nonstationary. While the first interpretation refers to the sequential dependency claim, implying that the probability of success increases with the shot numbers, the second implies that after a few consecutive hits, a hot players probability of success increases for the next trials before return again to its base rate – implying that it is a temporary and short-lived event. The two interpretations of sequential performance should however be evaluated using different statistical methodologies, depending on the nature of the claim, although they are often conflated, misinterpreted, and combined (Avugos & Bar-Eli, 2015).

Related Literature

Hot hand is by no means a new phenomenon, with its modern foundational roots often ascribed to Kahneman & Tversky (1971) and Gilovich, Vallone & Tversky (1985). Some however, look even further back for its foundation. Gigerenzer (2000) notes that Laplace (1951*)* used the same theory to explain the *Monte Carlo fallacy*[4], describing the fallacy with the same fundamental problems as the hot hand. Both of these paradoxical events, however, have been linked to the "representativeness" heuristic.

Literature on the subject during the 90's, and early 2000's is sparse, with some notable exceptions in Albert (1993), Frame et al. (2003), and Miyoshi (2000). Albert (1993) suggests a Markov model in which a player switches between two states (hot and cold), within or/and across games. This model, however, did not demonstrate that any streakiness exists. Miyoshi (2000) demonstrates through simulations that the original paper by Gilovich, Vallone & Tversky (*1985*) could provide a hot hand if the four parameters they used were adjusted realistically; however, when the author changed the variables that were used, the test could detect, on average, only 12 percent of all hot hands. Hence, he concludes that one could not reject the existence of a hot hand, only based on the findings of Gilovich, Vallone & Tversky (1985). Finally, Frame et al. (2003) prove that typical hot hand tests are incapable of detecting non-stationarity and changes in the chance of success. The authors propose a model in which a player had a set probability of switching between hot and cold regimes. In this model, there was a weak signal of a hot hand. While numerous others may have established the notion theoretically, these authors contribute with new tools for empirically examining the hot hand. Although these studies are empirical, they tend to favor the often desirable controlled experimental setting, but more modern approaches have subsequently been applied in the literature since the criticism on the original paper by Miller & Sanjurjo (2014).

[4] Also known as the Gambler's fallacy; If an event occurred more often than usual in the past, it is less likely to occur again in the future (or vice versa)

One of the first such studies is by Bocskocsky et al. (2014), that investigate the hot hand with game data from NBA from the 2012-2013 season. They employ a model for which they try to control different settings during a game (e.g., defensive controls, shot controls and game conditions). They show that a hot player is more likely to take the next shot of the team, as a result of players' optimal strategies, and that the common view that the hot hand is a fallacy is a fallacy itself.

Goldberg et al. (2018) similarly study game data over the 2016-2017 season where they add a string of 1's and 0's to the Warriors where they specifically follow the three superstars (i.e., Curry, Durant & Thompson). The authors conduct a 10,000-fold permutation test on all the strings that represent the shot pattern and discover no statistically significant pattern, and so, no hot hand, even though Thompson produced one of the greatest quarters of all time, scoring 60 points in 12 minutes on 31 of 44 attempts.

Leaving basketball, Green & Zwiebel (2018) investigate the topic on Major League Baseball, with approximately 2 million observations from 12 seasons. They further distinguish themselves in the choice of the empirical model, for which they use panel data regression on different lengths in their data set to account for player ability, and where the results are in contrast to the majority of the literature, where they find strong evidence of being hot. Additionally, the research indicates that teams react fast to a hot hand but tend to overreact to previous performance. They suggest that if one accounts for endogenous defense mechanisms, similar techniques may be beneficial in other sports.

In the field of nonverbal communication, there is a large body of research where Raab et al. (2012), find that in volleyball, players and coaches could identify and use the hot hand in their tactical and strategic decisions where teammates allocate the ball more often to players in the middle of a streak rather than to a player with the higher average base rate. These behavioral alterations indicate that both performers and viewers are good at detecting and responding to a hot hand.

Finally, Lantis & Nesson (2021) compare the percentage of successes occurring immediately after *k* consecutive successes to either the overall percentage of successes or the percentage of successes occurring immediately after *k* consecutive failures. They discover that large data sets are required to obtain a critical evaluation of basketball shooting deviation and their permutation tests of the null hypothesis of randomness provide a mathematical and statistical foundation for designing and validating experiments that directly compare deviations from randomness to human beliefs about deviations from randomness – thereby constitute a direct test of a hot hand.

Basketball

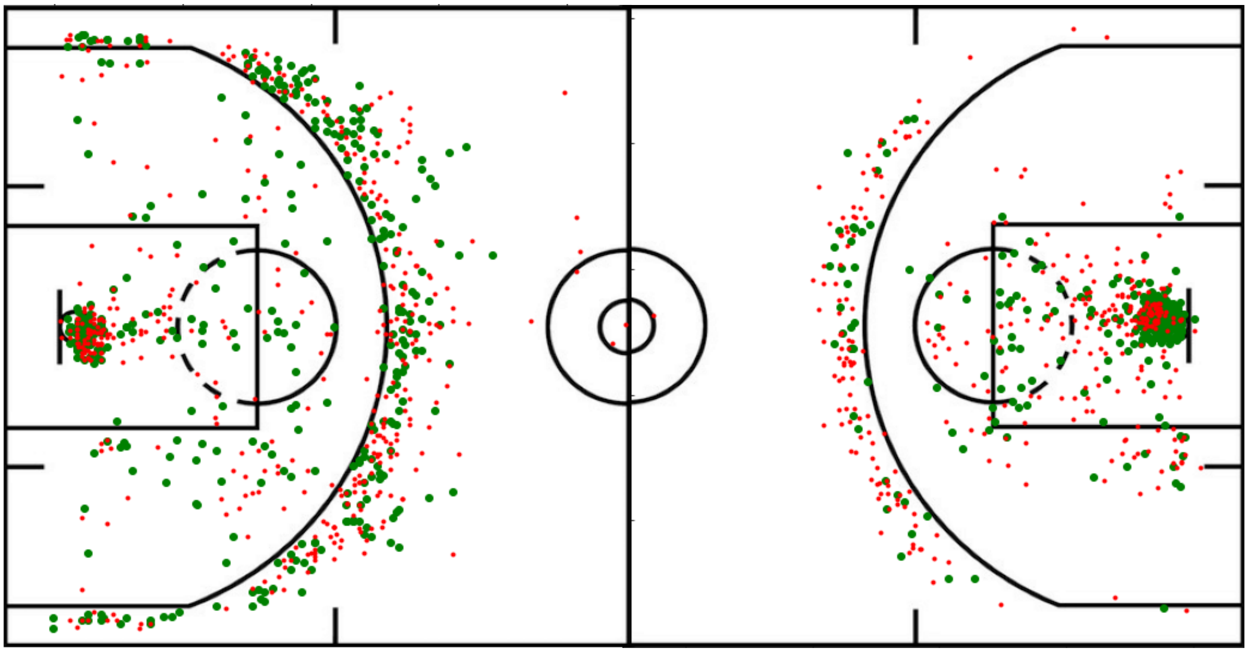
Basketball is a team sport in which all active players are on the court at the same time (5v5) and defenses are highly dynamic. On each possession, the offensive team distributes the ball among players in search of the optimal shooting chance, while the defense distributes defensive resources across offensive players. Offenses and defenses are always changing, with constant adjustments made to create and defend against shooting possibilities. These adjustments are frequent, comprehensive, and highly dependent on the circumstances. A simplified generalized model of this kind of play is that offenses attempt to maximize the value of a shot across players and shooting possibilities, while defenders attempt to minimax the value of a shot throughout this offensive selection.

Previous assumptions of game data have been that in a Nash Equilibrium, endogenous responses are likely to equate margins across offensive players – offenses (defenses) are responding to opposing defenses (offenses) optimally (Green & Zwiebel, 2018) .That is, if player A has a larger number of marginal shots than player B, player A should shoot more, and player B should shoot less, and defenses should cover player A more and player B less until this disparity disappears. Hence, hot players, should not have better marginal shots; both defensive and offensive adjustments will attempt to correct for such margins. This straightforward assumption, however, contains some qualifications and limitations. One such limitation is based on the unique characteristics of the game – for example, three-point shots are worth more than two-point shots, and the shots of inside players are harder to set up than attempts of outside players. A difference should thereby be made between the value of the shot and the chance of making the shot. These points should not affect the core argument that defenders adjust to reduce the best shot that offenses may take, but it is important to emphasize that the highest value shot for an attack will not always be the highest percentage shot for an offense for a variety of reasons. To accurately account for the worth of a shot, a variety of other criteria must be considered, for example, long shots are worth more, the possibility of getting additional scoring opportunities (rebounds and the likelihood that the shooter being fouled), and the difficulty in setting up a shot. Hence, for these and related reasons, the assumption that marginal shooting percentage across offensive players should be equated is not entirely accurate (Aharoni & Sarig, 2012).

Huizing & Weil (*2009*) demonstrate that defensive adjustments are constrained since a shooter who has recently completed a sequence of shots continues to shoot at the same percentage rate as before. A few previous studies using game data with field goal percentage as a metrics have found that there may be a defensive reaction when examining hot players in NBA (Bocskocsky et al., 2014, Csapo & Raab, 2015). Generally, field goal percentage is a poor indicator of the shooting skill of an athlete, since chance variation may conceal actual distinctions between players[5]. This is particularly true when extra contextual information such as shot location or shot type is included (see Figure 1). For instance, Franks et al. (2016) demonstrate that the most of reported variations in three-point percent are attributable to sample variability rather than genuine ability differences, making it an ineffective statistic for player discriminating. They illustrate how to reduce these difficulties by using hierarchical models that reduce empirical estimates to more realistic prior means such as the true shooting percentage. These compressed estimates are more robust.

[5] Example, James Harden is one of the best players in the league, with an FG% of 0.444 in 2019-2020, which is below the league average of 0.460. Closer inspection reveals that Harden makes 55 percent of his shot from the three-point range, which is not included in his FG%. Harden had a TS% of 0.627 that same year, second best in his position and distinguishable above the league average of 0.500.

FIGURE 1 - ILLUSTRATING DIFFERENCES IN SHOT ALLOCATION



Notes: This figure illustrates outcomes of shots from two all-star players in the season 2020-2021: Stephen Curry (left court) and Giannis Antetokounmpo (right court). The green dots represent all shots that are hits and the red dots are shots that are either missed or blocked[6]. There is a substantial difference in the locations of the players' shots. This, however, will only be represented in the TS%.

[6] There is a 0.1-inch (0.25cm) deviation of some dots to provide a more visual representation of the shots.

Data

This section provides a description of the data obtained by BigDataBall and the modified data frame with its key characteristics.

Dataset

I use game data from NBA of the ten-season from 2011-2021, which I obtain from BigDataBall.com. Each observation in the BigDataBall data set represents an event, which is defined as any occurrence that alters the state of the game. Additionally, each observation includes 45 variables that characterize the status of the game before and during the event (e.g., player that takes the shot, closest defender, type of shot, home/away game, date, quarter, location of the shot, etc.). There are around 6.5 million observations where 2.7 million are shots that have been taken from 1.333 different players in 12.581 games played in total.

Table 1 summarizes the shot log I built from the play-by-play data from BigDataBall to create a robust analysis of each shot. I made every effort to acquire the best data possible, where the total number of shots used is *at least* one order of magnitude higher than previous analysis of game data that I am aware of, in which they have used at most 10^5 number of observed shots.

The shot location is a robust characterization of the two variables X and Y coordinates to make sure that each shot has the correct measurement. The X and Y coordinates represent the dimensions of a basketball court (in feet), with Y denoting the width of the court $(0 - 50)$ and X denoting the length of the court $(0 - 94)$. Games in which a player takes less than six shots were removed owing to a Markov Chain estimate of the number of shots required to have a streak of at least two consecutive hits[7,8]. Thirty-two players were dropped during this sorting as they did not have at least one game with a sequence of six or more shots (see Appendix A for more data).

[7] With an exception in the regression discontinuity design where no player is removed due to the model and the methods that are used.

[8] $A=(1-p)(1+A)+p(1-p)(2+A)+p^2 \cdot 2 = A=\frac{1+p}{p^2}$, Set p = 0.5, average tries for 2 in a row = 6 tries

TABLE 1 - SHOOTING VARIABLE SUMMARY STATISTICS

Variable	Non-Missing Observations	Mean	SD	Min	Max
Make Streak	1.005.432	0.793	1.134	0	14
Miss Streak	858.266	1.031	1.341	0	16
Unique Shot Series	254.444	13.295	5.333	0	50
Shot Number	2.715.094	25.074	10.900	0	50
Shot Made	1.248.943	0.460	0.050	0	1
Shot Missed	1.466.150	0.540	0.043	0	1
TS%	1.333	0.500	0.046	0	1
Shot Distance	2.364.412	13.317	10.163	0	89
X Coordinate	2.364.412	13.318	10.163	0	50
Y Coordinate	2.364.412	47.091	34.408	-0.200	94.2
Players Touch	6.204.111	42.175	18.537	1	112
Played Minutes	6.596.414	17.450	7.1898	0.2	51.368
Closest Defender	2.364.412	4.129	2.759	0	53.2

Notes: This table summarizes the shooting variables. All variables are from BigDataBall. The TS% is calculated in the following way: $TS\% = PTS \cdot 100/2 \cdot (FGA + (0.44 \cdot FTA))$. Players touch represent each ball touch player *i* does throughout one game whereas played minutes is the total number of minutes player *i* played during a game and finally, closest defender is the distance between the shooter and the defender relative to a straight line between the two of them. Make streak is the number of shots that are made with at least two hits in a row. Miss streaks represent the opposite (i.e., number of shots that are missed with at least two misses in a row. Shot number is a numeric measure to see the sequence of shot each player has taken during each game (i.e., if a player takes 10 shots, shot number gives each shot a unique value from 1 – 10 in the order the shot was taken). The unique shot series is a numerical value calculated by grouping all players' shot numbers throughout a game by the actual number of games in which the shot number was taken.

Pre-registration

This study was pre-registered at AsPredicted (reference number 83508; date 12/17/2021). The pre-registration is available at https://aspredicted.org/6q3i5.pdf.

The Models

This section aims to provide a detailed description of the models used in the paper and their key components.

Bayes' Theorem & Conditional Probability

I rely mainly on Cunningham (2020) for the fundamentals of Bayes' theorem, where Cunningham (2020) is referred to for a more complete evaluation. The essential principle of Bayesian techniques is to use probability distributions to try to express the parameter vector's uncertainty. As such, it treats the parameter vector as a random variable, in contrast to conventional estimation techniques, which assume the presence of a fixed true variable value and so estimate it. The technique begins with the specification of priors, which refer to a priori knowledge or assumptions about the model parameters, that is, prior to accounting for the information included in the data. By combining the prior probability, which represents the distribution of the data given the model, and the posterior probability, which I need to prove causation between the two probabilities, the following equation is obtained:

$$State_{i,t} = P\left(\gamma_S \middle| Y_{i,t-k}, S\right) = \frac{P(Y_{i,t-k}|\gamma_S)P(\gamma_S)}{P(Y_{i,t-k}|\gamma_S)P(\gamma_S)+P(Y_{i,t-k}|\sim\gamma_S)P(\sim\gamma_S)} \tag{1}$$

Here $P(\cdot)$ is the probability function, γ_S is the key predictor parameters of the model, S is used to note the specific sequence of a player, and Y is the data up to and including period $t-k$ for player *i*. The denominator is the marginal probability of the data, conditional on the model (i.e., making previous shot), and the numerator corresponds to the posterior probability – making consecutive shot(s). The model may be rewritten using the naive Bayes' algorithm for simplified interpretation:

$$P\left(\gamma_S \middle| Y_{i,t-k}, S\right) = \frac{P\left(Y_{i,t-k}|\gamma_S\right)P(\gamma_S)}{P(Y_{i,t-k})} \tag{2}$$

A naive Bayes' classifier assumes that the existence of one feature in a class has no correlation with the presence of any other feature with an assumption of independence among predictors. The naive Bayes' model is particularly beneficial when dealing with extremely big data sets. Along with its simplicity, naive Bayes' algorithm has been shown to outperform even the most advanced classification techniques (Kotu & Deshpande, 2019). The two events, γ_S & $Y_{i,t-k}$ are independent if and only if:

$$P\left(\gamma_S \middle| Y_{i,t-k}, S\right) = P(\gamma_S) \quad (3)$$

When it is thought that they actually affect one another, a logical fallacy known as post hoc ergo propter hoc occurs. Post hoc ergo propter hoc is Latin for "after this, therefore because of this." This fallacy admits that the chronological ordering of events is insufficient to establish causality between events (e.g., hot hand fallacy).

Bias correction in the Hot Hand

Miller & Sanjurjo (2018) question the findings of Gilovich, Vallone & Tversky (1985) that have survived three decades of often critical research[9]. It is noteworthy, as the misguided methodology of Gilovich, Vallone & Tversky, underpins several lines of research on theoretically unstable occurrences[10]. The finding corrects the paper of Gilovich, Vallone & Tversky and observes that the Hot Hand fallacy itself is a fallacy. As a result, I provide the following brief statement to explain and acknowledge their reasonable criticism.

In basketball, a player shooting for the basket n times where it is assumed that the probability of a hit is constant throughout the sequence, like a coin toss, it is also assumed that all type of conditional probabilities follow from stationarity (independent of the order in which they occur) and independency. However, if the probability of a streak of one (or more) hits being followed by another hit in the next shot is greater than $P(H)$, independence does not hold. A positive dependence is usually referred to as the phenomenon a hot hand.

[9] See Avugos et al. (2013Aa) and Koehler & Conley (2003).
[10] See Rabin & Vayanos (2010).

The conditional probability that a streak of k hits will be followed by another hit is denoted by $P(hit|k\ hits)$. Miller & Sanjurjo point out here, that these relative frequencies are not the true conditional probabilities; rather, they are estimators of the unknown conditional probabilities and are, on average, biased downward under the assumption of independence. In addition, because the conditional probabilities of hits and misses are complements, the estimators of the probability of a streak ending on the next shot are biased upward. I will provide the derivation for the condition $k = 1$, that confirms the Miller & Sanjurjo (2018) calculations, demonstrating that the bias grows as the conditioning streak length increases[11].

Proposition 2.1 *Let* $p = P(H),\ q = P(T) = 1 - p,\ m = n - 1$*; with the assumption that coin tosses are independent*[12]*. S will be the sequence that the number of H's is among the first n-1, is nonzero and the estimator* $\hat{P}(H|H)$ *is well defined. Then*

(a) The conditional expectation of the estimator $\hat{P}(H|H)$ *given the event S, satisfies for any* $0 < p < 1$,

$$E(\hat{P}\ (H\,|\,kH)\,|\,S) = \frac{(1-q)}{1-q^m} - \frac{q}{m} \tag{5}$$

(b) The bias may be stated in the following way:

$$P(H) - E(\hat{P}(H|\ kH)|\ S) = q(1-q)[1 + q + \cdots + q^{(m-1)} - mq^{(m-1)}]/[m(1-q^m)]. \tag{6}$$

(c)

$$E(\hat{P}(H\,|\,kH)\,|\,S) < P(H)$$

which states that the estimator $\hat{P}(H|kH)$ *has a lower conditional expectation than the conditional probability given the event F under independence.*

Proof. If the first sequence of the first $m = n - 1$ is all T's, the S would not hold. Hence, the first *m* tosses can be one of $2^m - 1$ sequences with the probability of event $S = 1 - q^m$. The probability that a random sequence has *i* H's in the first m tosses and $m - i$ T's can be shown as:

$$\binom{m}{i} p^i q^{m-i} \tag{7}$$

[11] Without losing continuity, a reader who desires to disregard technical details might skip the proposition and its proof.
[12] Miller and Sanjurjos calculations are based on coin flips. Hence, the notation H (heads) & T (tails).

By choosing one of the *i* H's and placing it at random at one of the first *m* positions, it will have the probability $\frac{1}{m}$ and be the last *i* H (that is, *m*th). The following H will then have the probability *p* which will not generate any sequence. However, with the probability $\frac{m-1}{m}$, a random chosen H will not be the last. In this case, the following H will a probability of $\frac{t-1}{m-1}$. By combining this information, we can determine the probability of the event A conditioned on S using the formula for total probability.

$$P(A|F) = \frac{P(A\cap F)}{P(F)} = \sum_{i=1}^{m} \binom{m}{i} p^i q^{m-i} \left[\frac{1}{m}p + \frac{m-1}{m} \cdot \frac{t-1}{m-1}\right] / (1 - q^m) \tag{8}$$

By using the binomial sum and the expectation formulas*:* $\sum_{1=1}^{m} \binom{m}{i} p^i q^{m-i} = 1 - q^m$ and $\sum_{i=1}^{m} i \binom{m}{i} p^i q^{m-i} = mp$ equation (2) is obtained as follows:

$$P(A|F) = \frac{p}{m} + \frac{p}{1-q^m} - \frac{1}{m} = \frac{p}{1-q^m} - \frac{1-p}{m} = \frac{1-q}{1-q^m} - \frac{q}{m} \tag{9}$$

And by the fact that $0 < q < 1, 1 + q + \cdots + q^{m-1} > mq^{m-1}$, *implies eq (c).*
It's essential to know how bias might lead a real hot hand to look erroneously as independent.

$P(H|kH) = P(H|kT)$ implies independence, while hot hand implies $P(H|kH) > P(H|kT)$. Thus, the aforementioned biases suggest that under independence, the expectancies $E\hat{P}(H \mid kH) < E\hat{P}(H|kT)$. This indicates that, in the absence of independence, one will more often than not observe $\hat{P}(H \mid kH) < \hat{P}(H \mid kT)$. Thus, small positive, negative, or zero values of $\hat{P}(H \mid kH) - \hat{P}(H \mid kT)$ may actually indicate that independence has been broken, implying a hot hand, since the bias causes $\hat{P}(H \mid H) - \hat{P}(H \mid T)$ to be larger than predicted.

Based on a Monte Carlo simulation I ran, Miller & Sanjurjos corrections are accurate. In a coin toss with a constant probability of 0.5, the true $P(H|3H)$ should be expected to be 0.5. The expected value of the observed conditional probability $\hat{P}(H|3H)$ is surprisingly not 0.5 but rather

0.405. Thus, although comparing observed conditional probability to 0.5 appears appealing, it is highly misleading and biased[13].

TABLE 2 - THE BIAS IN THE CASE OF COIN FLIPS

$P(H \mid k\ H)$ Flip sequence	P(H) The bias
k = 0	.500
k = 1	.482
k = 2	.440
k = 3	.405
k = 4	.347
k = 5	.320

Notes: Left column lists the six different sequences from the first to the fifth toss if the coin continues to land on head. The bias of the flips that immediately follow a previous sequence of heads is reported in the right column. For each toss, the bias grows substantially where a probability of .500 is misleading.

Weighted Least Squares regression

The model employed for the autocorrelation claim is a regression with a strategy similar to what was previously described by Green & Zwiebel (2018). The main aspect in which this paper parts from their study is in the sports that was analyzed. While they looked at baseball, I will here aim to fit the model to the case of basketball. I further also extend the model with a weighted least squares regression instead of an ordinary least square regression where I will weight all shots equally. This model adjustment is motivated by the fact that when the predictor variable is measured with error in a linear regression, attenuation of the regression coefficient tends to skew the slope towards zero – an underestimation of its absolute value – due to the independent variable's errors (Wayne, 1978). For example, if player *i* makes 60 percent of shots when hot and 40 percent when cold, the P(hit|1 hits) is .60(.60)+.40(.40)=.52, while the P(hit|1 miss) is .40(.60)+.60(.40)=.48. However, by using a weighted average rather than an unweighted average, I can mitigate this tendency, as weighting all shots equally creates another form of

[13] See Miller & Sanjurjo (2018) for further details.

"Simpson's Paradox"[14], which is a bias toward finding the hot hand (selecting a better player). However, due to the extensive data set, this bias will be exceedingly small, if at all, when adjusting for attenuation[15].

I estimate the weighted least squares for 12 different outcome statistics of particular interest, namely, the state of the player after *n* (1, 2, 3) consecutive hits. I evaluate the state of the player using his recent success rate in his last *k* (full season, half season, ten most recent games, and five most recent games) games were, Y_{it} denote the prediction error between the outcome of the shot and the true percentage rate of the player in player *i*'s attempt at time *t*.

$$Y_{i,t-k} = \propto + \gamma State_{it} + \beta X_{it} + \epsilon_{it} \quad (10)$$

Where $\epsilon_{it} \sim N(0, \sigma^2/w_i)\ for\ known\ constant\ w_1, \ldots, w_n$.

I will use $w_i = 1/x_i$, where x_i= number of shots per game as I expect an increasing relationship between $\mathrm{Var}(\gamma | Y_{i,t-k})$. The weighted least squares estimate of $\propto + \gamma$ minimize the quantity[16]

$$S_w(\propto, \gamma) = \sum_{i=1}^{n} w_i (y_i - \propto - \gamma State_{it} + \beta X_{it})^2 \quad (11)$$

For each outcome, I estimate the model in which *i* refers to a player in a given year. X_{it} denotes a vector of control variables, and the coefficient of interest, γ, indicates whether previous shot predicts the outcome of current shot. If it does not, then one would expect γ = 0. While the standard test for a hot hand in the literature is formulated around whether *gamma* is positive or not, a positive coefficient could indicate streakiness, hence being hot. In contrast, a negative coefficient would be indicative of reversals in outcome. Stone (2012) describes this as a form of "measurement error" on hot hand estimates, arguing that it is more appropriate to condition on the probabilities of previous makes, $E\left[\gamma | E[Y_{i,t-1}], \ldots, E\left[Y_{i,t-k}\right], X_t\right]$ rather than observed makes and misses themselves. Lastly, I will extract the estimated autocorrelation coefficient for each

[14] Simpson's Paradox:
Unweighted average of 14 sequences= $[((6 \cdot 1/6) + (6 \cdot 1/2) + (2 \cdot 5/6)]/14 = [17/3]/14 = .405$
Weighted average of 14 sequences= $[((1)(6 \cdot 1/6) + (2)(6 \cdot 1/2) + (3)(2 \cdot 5/6)]/(1 \cdot 6) + (2 \cdot 6) + (3 \cdot 2)] = 12/24 = .50$
[15] I conducted a robustness check with an ordinary least squares regression which produced the same coefficients and parameters as the weighted least squares regression. Hence, no coefficient is given for the OLS.
[16] Weights are inversely proportional to the corresponding variance in the weighted sum of squares; players with low variance will have higher weights, while players with high variance will have lower weights.

player in the regression in order to identify players with the largest autocorrelation between the two measurement errors. This will be performed by doing separate regressions on each player in the dataset for each season in which they appear. The following trade-off arises when selecting L:

$$State_{i,t} = R_{it}^{L} = \frac{1}{L}\sum_{k=1}^{L} \gamma_{i,t-k} \tag{12}$$

If the state is constant over time, a longer period will result in a more exact identification as a player that has made three shots in a row might as well be lucky. To add additional shots and games will help to distinguish between state and player ability. For instance, if the state changes significantly over time, the longer time periods are less useful for describing the present state and are more important for evaluating the long-term ability of a player.

Linear Regression

I am not the first to recognize that the Gilovich, Vallone & Tversky data are underpowered for the Markov chain alternatives. Miller and Sanjurjo (2018), Miyoshi (2000), and Wardrop (1999), for example, evaluate the power of individual tests against particular parameterizations of related models via simulation. Korb & Stillwell (2003) and Stone (2012) evaluate power against certain non-stationary alternatives. I contribute to these studies by applying a linear regression to construct a model that predicts the shooting complexity of each shot for player *i* taking shot *s* in each game to provide robust characterization to be able to adjust the difficulty of each shot. The technique is influenced by Bocskocsky et al. (2014) where the shot difficulty is anticipated based on four main categories of factors of shot difficulty:

$$\hat{P}_{is} = \propto + \beta \cdot (Game\ Condition_{is}) + \gamma \cdot (Shot\ Controls_{is}) + \delta \cdot (Defensive\ Controls_{is}) + \theta \cdot (Player\ Fixed\ Effects_{is}) \tag{13}$$

Game condition proxy for the differences in effort across shots, player pressure and player fatigue where location of game (Home/Away), time remaining and the marginal score between the teams are used.

Shot Controls evaluate the difficulty of a shot where the variables; Shot clock, point type and the exact coordinates of each shot with its play-by-play categorization is used.

Defensive Controls assess defensive intensity, such as the absolute distance between the shooter and the closest defender relative to a straight line between the shooter and the basket.

Player Fixed Effect uses the TS% to account for variances between players, which means that when two players shooting from the same coordinates, their chances of making it differ.

There are two main differences between my model and the Bocskocsky et al. (2014) model. I calculate the player fixed effects with the true shooting percentage rather than the field goal percentage and further, use the exact X and Y coordinates of each shot rather than the shot distance from the basket. In the case where the model predicts a value greater than one or less than zero, I replace those values with 0.99 and 0.01 as no shot in basketball has a hundred or zero percent chance of going in. To test the accuracy of the model, I run it on a randomized training set consisting of half of the data where I then predict all the P for the validation set, creating a pseudo-out of sample test[17]. Due to the non-linear relationship between shot efficiency and shot coordinates, I divide the shot coordinates into 30 categorical groups of distance and include this set of mutually exclusive and exhaustive dummy variables in the shot difficulty model. This allows for a non-parametric specification, where each categorical group coefficient is allowed to vary.

To visualize the predicted P(hit), I group all the P and bin all shots whose values are within the same hundredth increment of the predicted P(hit) and then calculate the actual P(hit) based on the hit and miss status of all shots within the same bin. If the model is accurate, the P bins should correspond closely with actual made percentage. Figure 3 presents the scatter plot of the data and illustrates that the model generally fits the data very well[18].

[17] Due to the high number of variables and interactions, as well as the highly non-parametric specification, the majority of the hundreds of coefficients do not have an easy interpretation as a result, no coefficients are provided.

[18] Codes are given upon request

P(hit) then serves as a proxy for shot difficulty when calculating adjusted conditional probability. The adjusted probability is determined in a manner similar to how a biased coin is corrected[19]. Adjusted hits= $\sum\left[\left(\frac{1}{p}\right)\right.$ for *prob* in hits] and adjusted misses=[$\left(\frac{1}{1-p}\right)$ for *prob* in misses]. I then create a metrics to evaluate all players performance where I pool out the ten best players for analysis, that is picked based on the outcome of the matrices as they in general should score more, play more games over the season, and shoot more each game. This enables the study of more series and longer series per player (Appendix C).

Regression Discontinuity Design

To estimate strategic decisions and allocation adjustments (i.e., offensive adjustments, strategic decisions, and defensive allocation), I will conduct the RD design where the fundamental concept of the model is to define a threshold (three consecutive hits), over which highly identical players may have significantly different outcomes. Assume that D_i is a discontinuous function of a continuous underlying variable x_i. When x_i exceeds a given threshold x_0, exogenous reasons cause a discrete "treatment" D_i to take effect.

$$D_i = \begin{cases} 1 \; if \; x_i \geq x_0 \\ 0 \; if \; x_i < x_0 \end{cases} \tag{14}$$

By assuming a linear function of x_i, the regression model will take the following form:

$$y_i = \propto + \beta x_i + \rho D_i + \eta_i \tag{15}$$

Where D_i is a deterministic function of x_i. However, to allow for reasonably smooth flexibility in the model, it can be re-written as:

$$y_i = \propto + f(x_i) + \rho D_i + \eta_i \tag{16}$$

[19] Consider a biased coin that give head 30% of the time and tail 70% of the time. Each time I get a head, I will count it as 0.7 and each tail as 0.3 for this coin, the adjusted expected value of obtaining a head or a tail would be approximately 0.5, as if the coin were fair.

If $f(x_i)$ is continuous in the neighborhood of c_0, it is possible to estimate this model, even with a flexible function form for $f(x_i)$. I perform a non-parametric RD on data that is close to the discontinuity. The assumption is that players who are slightly above or below the threshold would behave similarly, with the expectation that some will be "treated," (i.e., subjected to the threshold (make the fourth shot). I can thus look at data in a neighborhood around the discontinuity: $[c_0 - \delta, c_0 + \delta]$ for some small number δ:

$$E[\gamma_i | x_0 - \delta < x_i < x_o] \cong E[y_{0i} | x_i = x_0] \quad (17)$$

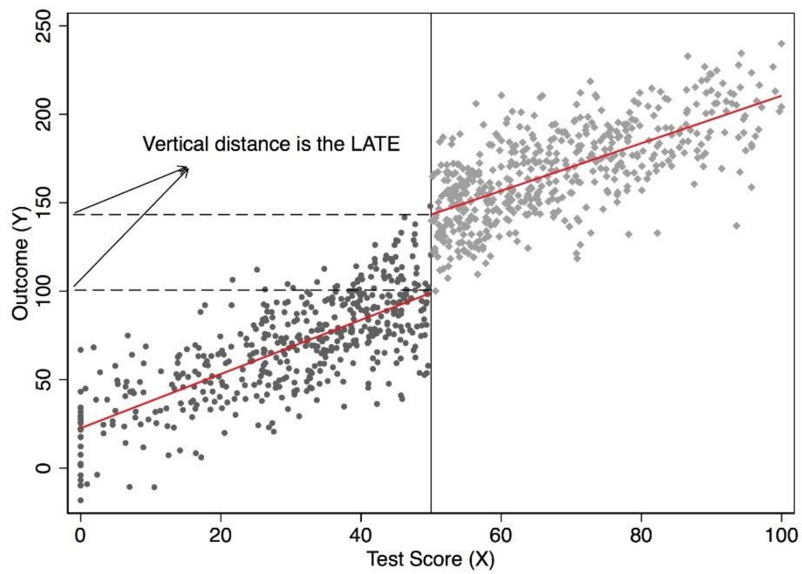
$$E[\gamma_i | x_0 < x_i < x_0 + \delta] \cong E[y_{1i} | x_i = x_0] \quad (18)$$

The non-parametric RD-estimate can then be written as:

$$\lim_{\delta \to 0} E[\gamma_i | x_0 < x_i < x_0 + \delta] - E[\gamma_i | x_0 - \delta < x_i < x_o] = E[\gamma_{1i} - \gamma_{0i} | x_i = x_0] \quad (19)$$

Where, $E[\gamma_{1i} - \gamma_{y0i} | x_i = x_0]$ represents a comparison of the environment for a player that made the fourth shot and one that missed it (i.e., more passes, defensive attention, minutes in the game or a change in the shot distance of the player) (see Figure 2). Control variables other than the driving variable x_i are not required in the RD technique. The reason for that is that large difference in other factors should not exist around the cutoff point x_0.

FIGURE 2 - REGRESSION DISCONTINUITY DESIGN



Notes: The gap difference $E[\gamma_{1i} - \gamma_{y0i}|x_i = x_0]$ is denotes as LATE in figure 2. The difference in the two intercepts represents the different response the environment has on a hit in the fourth shot compared to a miss in the fourth shot. Source: Cunningham (2020)

To be able to demonstrate the features of the model, I will aggregate all of the data for the 200 best players who will be chosen from the players metrices in Appendix C. Additionally, as most of the literature evaluate a hot hand on what happens after three consecutive hits. I will set the threshold to 3.5 to observe how the environment responds to a successful shot vs a missed shot on the fourth trial. Assuming that a player could have multiple streaks throughout a game, the longest streak a player has during the game will decide his grouping in that game. As described in the Data section, I will not drop games if the players have less than 6 shots. Thus, if a player appears for half a minute or makes a single shot throughout the course of the game, this will be noticeable. However, although I am only interested in what happened after the third strike, this has no impact on how the proposed threshold is interpreted.

Empirical Findings and Analysis

This section aims to provide a descriptive presentation and analysis of the results and try to bridge my findings to the previously established research.

Weighted Least Squares Estimates

Appendix B contains regression estimates. Table 6 – Table 8 include estimates for each of my 12 statistics, as well as for the main considerations' specifications discussed in section Bayes' Theorem & Conditional Probability.

The outcomes are rather remarkable. For each of the twelve statistics considered, the estimates for γ, the coefficient for players R_{it}^{L} (state), are statistically significant at a 0.1 percent confidence level. In more than half of the estimates, I find evidence supporting the hot hand (58.3 percent). Even if the coefficients on all variables from recent history are statistically significant at the 0.1 percent level for all 12 statistics and three specifications shown in Table 6 – Table 8, columns (3)–(4) remains positively correlated, whereas column (2) fluctuates depending on the specific streak length. column (1) shows a negative connection in each of the three tables where the law of small numbers is applicable (Kahneman & Tversky, 1971). Throughout longer amounts of time (over an entire season), the regression is more likely to detect a player ability rather than a hot hand where the variation of a player goes to zero.

For each of the Table 6 – Table 8, row 1 contains the coefficient on state, which is the increase in the probability of an event (in percentage points) caused by a one percentage point increase in R_{it}^{L}. For instance, column (4) of Table 6 illustrates that a 10 percent point improvement in the previous shooting percentage of player *i*, results in a 0.12 percentage point increase in the likelihood of obtaining the streak on the current shot. Thus, the difference between a hit percentage of 0.400 and 0.600 (indicating a recent cold run and a recent hot run, respectively) translates into a (.2 * .12)*$100 = 2.4$ percent increased likelihood of scoring in the next shot. Similar, column (4) of Table 8 gives the following percentage increase: (.2 $*$ .07)*$100 = 1.4$. This magnitude is significant from a strategic perspective. In Table 5, the standard deviation of .046 in shooting percentage across all players throughout the sample corresponds to approximately a $.70 - .77$ standard deviation in player ability given a $1.4 - 2.4$ percentage increased likelihood. In other terms, the difference between a player who is hot and a player who is cool is about equivalent to the difference between a 50th percentile shooter and a 70th percentile shooter.

Linear Regression Estimates.

FIGURE 3 – BIAS CORRECTED DIFFERENCE

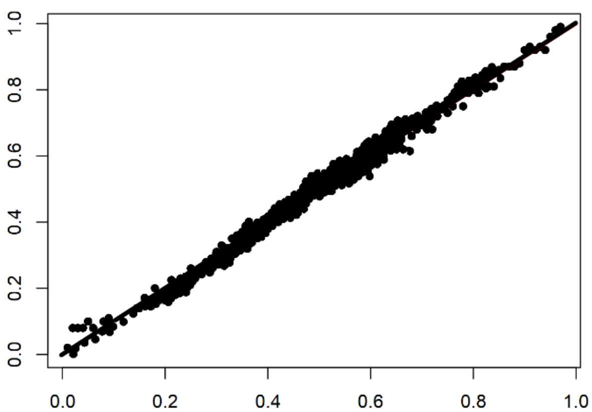


FIGURE 4 – DISTRIBUTION: MADE/MISS STREAK

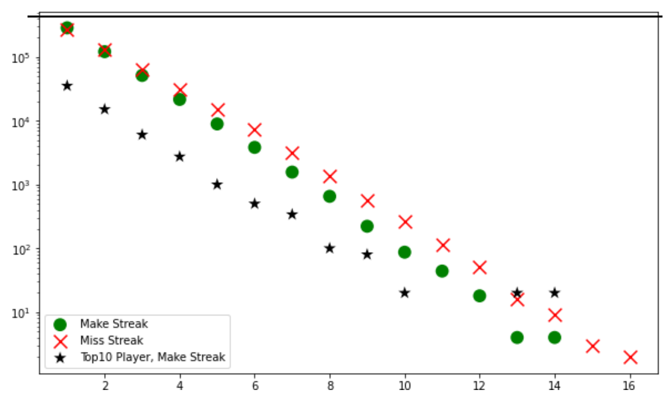


Figure 3 presents the predicted $\hat{P}(Hit)$ for which it agrees really well with the actual hit percentage. The predicted P(Hit) between 0.2 to 0.8 has the majority of the shots since most shot fall within this range. Overall, the predicted P(Hit) of the data is reliable. Figure 4 presents the distribution of made/miss streaks. Where the black stars represent the top ten best players make streak. As can be seen, the best players perform similarly to the rest of the players, except when the sequence becomes longer – a sequence of 13 or 14 streaks in a row.

Following the shot difficulty adjustment, in which I account for shot selection dependence, the P(Hit) of the majority of players should be around 0.5, as seen in Table 3, column (3) where the average adjusted P(hit) of the best players is precisely 0.5. Column (6) is Gilovich, Vallone & Tversky's calculation of a hot hand where they simply see the difference in column (4) and (5): $\hat{P}(hit|3\ hits)$-$\hat{P}(hit|3\ misses)$. As can be seen in column (7), the bias adjustment that Miller & Sanjurjo (2015) found, changes the outcome of the table from a negative correlation to a strong positive correlation with a difference of 0.2 percentage points.

TABLE 3 - SHOOTING PERFORMANCE FOR K=3

Player	(1) # Shots	(2) $\widehat{\text{P}}$(hit)	(3) $\widehat{\text{P}}$(hit) adj.	(4) $\hat{P}(hit\|3\ hits)$	(5) $\hat{P}(hit\|3\ misses)$	(6) GVT est. ($\widehat{D}_3$)	(7) bias adj. ($\hat{A}_3$)
Stephen Curry	13.258	.47	.50	.44 (1006)	.47 (1281)	-.03	-.01
LeBron James	16.338	.53	.51	.47 (1704)	.53 (1473)	-.06	-.02
Kevin Durant	13.167	.51	.51	.48 (808)	.48 (1254)	.01	.02
James Harden	15.435	.44	.50	.50 (1146)	.48 (2079)	.03	.05
Kyrie Irving	11.708	.47	.48	.49 (2179)	.51 (1200)	-.02	.03
Klay Thompson	11.759	.46	.47	.48 (742)	.51 (1238)	-.03	.06
Joel Embiid	4.947	.48	.49	.52 (475)	.51 (603)	.02	.03
Nikola Jokic	6.713	.53	.53	.50 (788)	.48 (602)	.02	.04
Zach LaVine	6.039	.46	.47	.48 (533)	.50 (690)	-.02	-.01
Gianni Antetokounmpo	9.834	.53	.53	.53 (1053)	.49 (843)	.04	.05
Average		**.49**	**.50**	**.49**	**.50**	**-.04**	**.24**

Note: Table 4 replicates in part Table 4 of Gilovich, Vallone & Tversky (1985) and also in part Table II of Miller & Sanjurjo (2018). The number of shots in each category is put in parentheses. The Gilovich, Vallone & Tversky estimator can be found in column (6), which is just the difference of columns (4) and (5). Column (7) is a correction of Gilovich, Vallone & Tversky estimate for the corresponding bias which Miller and Sanjurjo (2018b) used. This result is statistically significant ($p<.01$), showing that on a group level, these players scored higher after a sequence of three hits than following a string of three misses.

Remarkably, as seen in Table 9, each of the 10 players chosen by the matrix have a much higher probability of scoring after k=1, where Klay Thompson increases his likelihood of making a second shot by 17.8 percent after he has previously made one. Kyrie Irving, on the other hand, raises his probability by 9.2 percent. Thus, certain athletes seem to have a tendency towards becoming hot[20].

[20] When randomly rearranging the players shot strings of make and miss patterns to allow for quantitative assessment of its rarity, no player had a rare shooting pattern.

Does the environment detect a hot hand?[21][22]

The hot player

FIGURE 5 - NUMBER OF SHOTS

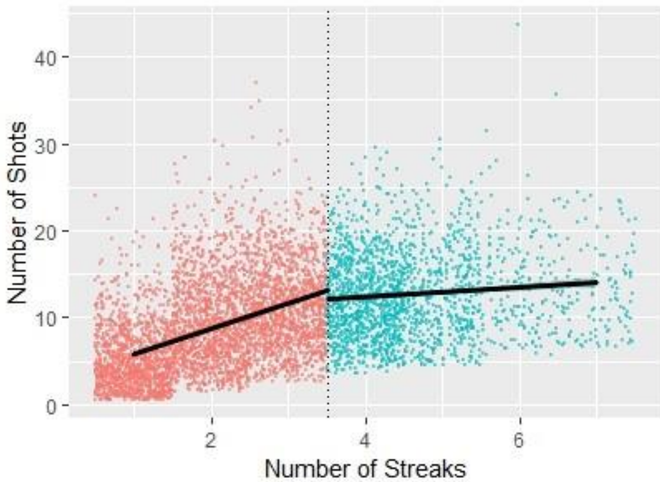
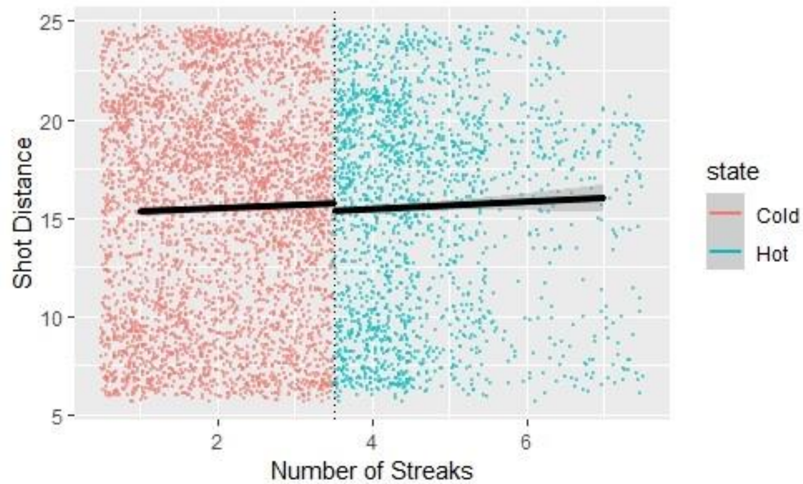


FIGURE 6 - SHOT DISTANCE



As seen in figure 5, players who have a higher sequence of recent shots or have performed better than expected on these shots are more selective in their shooting, whereas a hot player typically shoots -0.9 to -2.6 (see Table 10, column (4)) less shots after a three-hit streak. Hence, players are more cautious with their shooting and might even break the assumption of independence after a hit on the fourth shot. Around the threshold, the player shoots on average -0.9 less shots compared to what he would have shot throughout the game if he had missed his fourth shot.

Figure 6 and column (4) of Table 11 illustrate that the shot distance decreases when a player is getting hot. The size of the effect is -0.254 close to the threshold. As a result, if a player makes one more of his previous four shots rather than misses one, his predicted shot distance decreases by three inches (7.62cm). In comparison to the average shot distance 13.317 feet (4.06 meters), this is a drop of around 1.7 percent. This indicates that players stay rational throughout the majority of games, which contradicts what the literature on game statistics suggests. For instance, Bocskocsky et al. (2014) argue that players attempt more difficult shots after a longer series, with

[21] To improve the visualization of the graph, sequences more than 8 are omitted. This is done in order to have a better understanding of what is occurring at the threshold. They are, however, in regression.

[22] The figures presented in this section have been modified by jittering the observations with width=.5, height=.5 for a more appealing and a better understanding of the figure. This does not change the result, only the presentation of the figure.

the hot player shooting from an increased distance of 6 inches on average. However, they use the shot distance to the basket in their model, whereas I use the precise X and Y coordinates. As a result, some changes in distance may occur; nevertheless, the interpretation of the results is in direct opposition of what previous research has found. A professional basketball player remains rational and stays in a location of the field that he is used to being in.

Teammates & opponent

FIGURE 7 - NUMBER OF TOUCHES

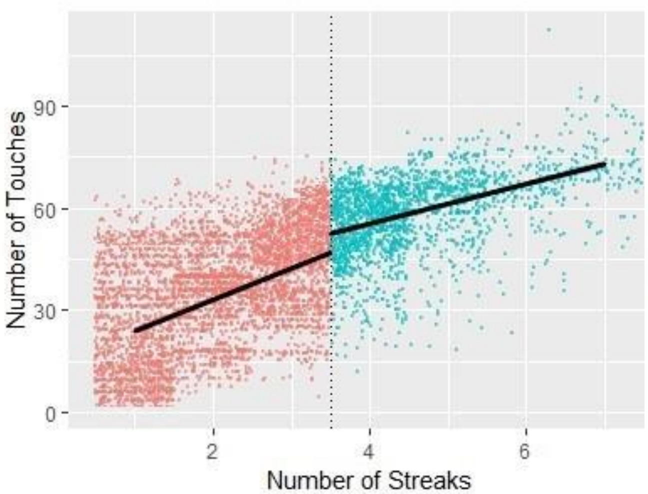
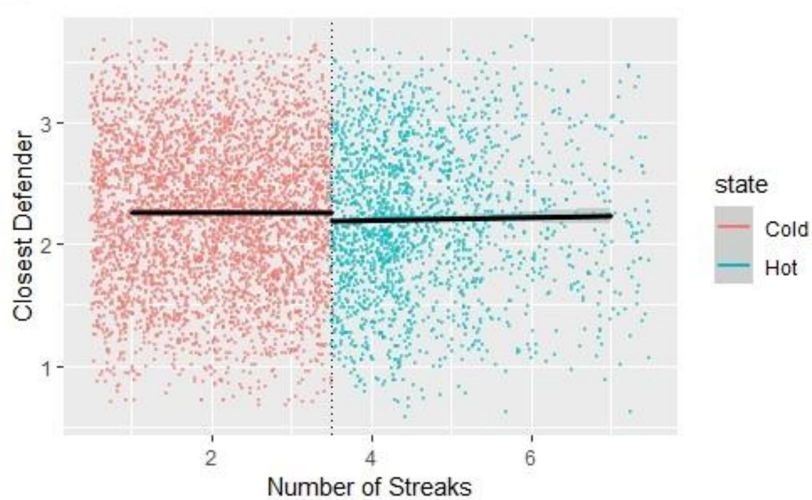


FIGURE 8 – CLOSEST DEFENDER

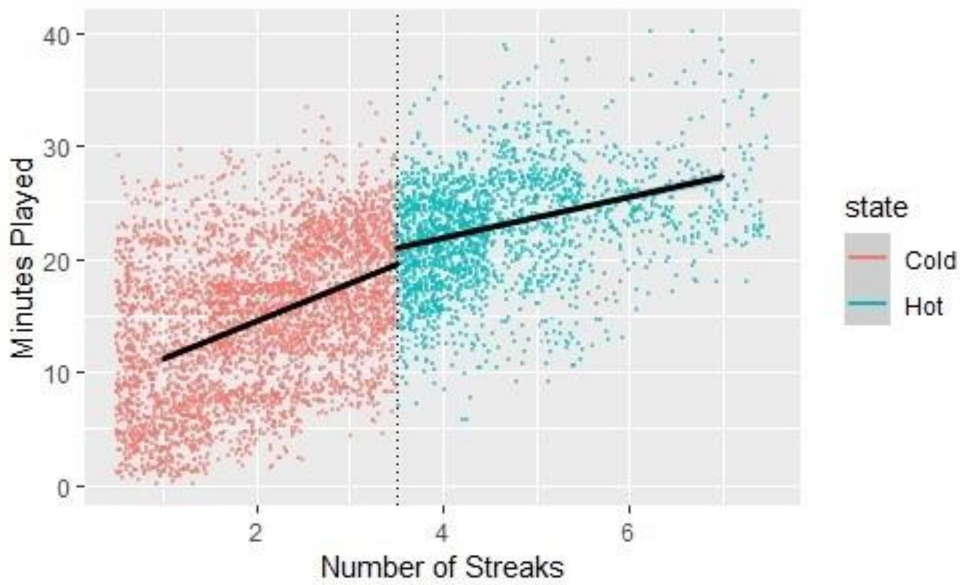


As earlier studies argue[23], basketball data contains endogenous reactions that are likely to equal margins between offensive players and that defensive adjustments would seek to compensate for such margins. However, as could be seen in column (4) of Table 13, when the offensive player hits his fourth consecutive shot, his defender reduces the gap between them by a half inch (1.5cm), nonetheless, this response is insignificant. Hence, game data is an ideal setting for analyzing the hot hand. Furthermore, the increase in a player's touch (i.e., every time the player touches the ball) increases when a player gets a longer sequence, teammates are fast to pick up signals on a hot hand and are willing to give the ball to a player in the middle of a streak. This goes in line with Raab et al. (2012) study of allocation decision in volleyball during a hot hand. A player that has made four consecutive hits, compared to a player that has had a streak of 3 hits, has on average 6.935 more touches during the game (see Table 12).

[23] See Green & Zwiebel (2018).

Coaches

FIGURE 9 - MINUTES PLAYED



I found no significant differences in how managers adjust their strategies and decisions when the streak is short. Whereas a streak of four compared to a streak of three provided the streakier player on average 20 seconds more playtime. However, as seen in column (4) Table 14, after a streak of seven or more, the coach increased the hot player's playtime by about two minutes, which is a significant amount of time even at ($p<.01$). This corresponds to a 4 percent increase of game time per quarter (see Figure 9).

Discussion

This study demonstrates a significant increase in players' probability of hitting the last shot in a two-, three-, and four-shooting sequence. I provide strong evidence that in a set of three consecutive shots, the probability of hitting the fourth shot is greater following a hit series than following a miss series on the previous shots for the top players (nonstationary claim). In general, the standard deviation of a player when becoming hot is approximately 0.70 – 0.77, which gives the player a 1.4 – 2.4 increased probability of hitting the target (autocorrelation claim). Furthermore, my findings illustrate that a professional basketball player remains rational throughout the sequence of hits whereas teammates react quick when detecting a pattern and where they actually act on it, however, my result show that these actions are moderate. Similar patterns have been seen where behaviorists have argued in the field of finance that investors often get lured into bad decisions by seeing patterns that are not real (Morgulev & Avugos, 2020). Hence, they make cognitive mistakes that might range from overconfidence in their own ability to a tendency to over-react to others (Goetzmann, & Peles, 1997). However, investment groups are less prone to the hot hand fallacy and tend to decide more optimally than individuals in strategic and non-strategic setting (Kahneman & Riepe, 1998), which may explain why coaches did not respond to a four-hit series but began to adjust during a longer sequence. Behaviorists use the claimed fallacy to reach erroneous findings[24]. The issue with believing that the hot hand is a cognitive mistake, could lead to a misinterpretation of an issue, as for example the mortgage crash in 2008, which might result in the application of an incorrect decision strategy. For instance, investment bankers have been criticized for poor decision-making and for underestimating the risks associated with subprime mortgages and for their irrational optimism (Shefrin, 2009). However, it is plausible that irrational risk-taking was encouraged or even subsidized by regulatory policies, such as those that protected banks considered *too big to fail*. The interpretation of these two financial crises may lead to two distinctly different policy strategies. Thus, the financial crisis can be represented as a result of a hot hand with a series of cognitive mistakes, or as the result of risky choices that were made based on a response to the incentives and benefits of the system.

[24] See Gilovich, Vallone & Tversky *(1985)*.

Conclusion

This paper looks into the hot hand fallacy in the case of basketball, by employing four different models estimated on data for the period 2011-2021. By looking at the performance of individual players in the case of play-by-play data, it aims to provide further insights into the performance of the phenomena under differing empirical models, as no empirical hot hand fallacy study of basketball has previously been performed on, to my knowledge, (1) a larger dataset and (2) with the empirical models that I used in this paper (i.e., RDD design and linear regression to adjust and recreate Gilovich, Vallone & Tversky (1985) original data but with game data). If controlling for the dependent shot selection using a regression framework to adjust and control for past expectations, basketball game data is optimal for analyzing hot hand.

It is reasonable, however, to assume that a player or a coach can identify a hot streak better than an econometrician who only looks at outcomes. Hot streaks may be identified by the player/coach but not by the econometrician. Players and coaches may also evaluate if a hit was solid or accidental. A more accurate state identification would accentuate the performance gap. As a result, future study will need the collection of information on other critical components of the game. For instance, on defense, the poor pass you force the other team to make, or the excellent long pass you make that sets up another pass that results in a score. These are some of the areas in which an individual may excel that are not captured in modern statistics. Finally, analyzing the duration of the streak will provide a more specific indication of whether the player is hot or not, since the period between shots may be minutes, indicating that the player is shooting from a normal state rather than a hot one. Lastly, timing is critical when it comes to profiting on the hot hand effect, whether on basketball courts or Wall Street.

References

Adler, P. and Adler, P., (1978). The Role of Momentum in Sport. Urban Life, 7(2), pp.153-175.

Aharoni, G. and Sarig, O., (2012). Hot hands and equilibrium. Applied Economics, 44(18), pp.2309-2320.

Albert, J., (1993). A Statistical Analysis of Hitting Streaks in Baseball: Comment. Journal of the American Statistical Association, 88(424), pp.1184.

Alderman, R., (1974). Psychological behavior in sport. Philadelphia: W.B. Saunders Co.

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

Avugos, S., Bar-Eli, M., Ritov, I. and Sher, E., (2013a). The elusive reality of efficacy–performance cycles in basketball shooting: An analysis of players' performance under invariant conditions. International Journal of Sport and Exercise Psychology, 11(2), pp.184-202.

Avugos, S. and Bar-Eli, M., (2015). A Second Thought on the Success-Breeds-Success Model: Comment on. Review of General Psychology, 19(1), pp.106-111.

Bocskocsky, A., Ezekowitz, J. and Stein, C., (2014). Heat Check: New Evidence on the Hot Hand in Basketball. SSRN Electronic Journal,. pp. 1-10.

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

Briki, W., (2017). Rethinking the relationship between momentum and sport performance: Toward an integrative perspective. Psychology of Sport and Exercise, 30, pp.38–44.

Capon, N., Fitzsimons, G. H., and Prince, R. A., (1996). An individual level analysis of the mutual fund investment decision. Journal of Financial Services Research, 10(1), pp.59-82.

Croson, r., and J. Sundali., (2005). "the gambler's fallacy and the hot hand: Empirical Data From Casinos," Journal of Risk and Uncertainty, 30, pp.195-209.

Csapo P., and Raab M., (2015) Correction. "Hand down, Man down." Analysis of Defensive Adjustments in Response to the Hot Hand in Basketball Using Novel Defense Metrics. Plos One 10(4).

Cunningham, S., (2020). Causal inference. isbn:9780300251685.

Daks, a., Desai, N. and Goldberg, L.R., (2018). Do the Golden State Warriors Have Hot Hands?. Math Intelligencer 40, pp.1-5.

Doron, J., & Gaudreau, P., (2014). A point-by-point analysis of performance in a fencing match: Psychological processes associated with winning and losing streaks. Journal of Sport and Exercise Psychology, 36(1), 3–13.

Frame, D., Hughson, E., & Leach, J. C., (2003). Runs, regimes, and rationality: The hot hand strikes back. Working paper.

Franks AM, D'Amour A, Cervone D. and Bornn L., (2016). Meta-analytics: tools for understanding the statistical properties of sports metrics. Journal of Quantitative Analysis in Sports 12:151–165.

Galbo-jörgensen, C. B., Suetens, S. and Tyran, J-R., (2016). "Predicting Lotto Numbers a Natural Experiment on the Gambler's Fallacy and the Hot Hand Fallacy," Journal of the European Economic Association, 14.

Gigerenzer, G., (2000). Adaptive thinking. Rationality in the Real World. Oxford: Oxford University Press. Acta Psychologica, 111(3), pp.351-354.

Gilovich, T., Vallone, R. and Tversky, A., (1985). The hit hand in basketball: On the Misperception of Random Sequences. Cognitive Psychology, 17(3), pp.295-314.

Goetzmann, W. N., and Peles, N., (1997). Cognitive dissonance and mutual fund investors. Journal of financial Research, 20(2), pp.145-158.

Green, B. and Zwiebel, J., (2018). The Hot-hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments? Evidence from Major League Baseball. Management Science, 64(11), pp.5315-5348.

Hamberger, M., and Iso-Ahola, S. E., (2004). Psychological momentum and athletic performance: A critical review of research. Journal of research. Journal of Contemporary Athletics, 1, pp.207-226.

Huizinga, J. and Weil, S., (2009). "Hot Hand or Hot Head? The True About Heat Checks in the NBA", MIT Sloan Sports Analyrics Conference.

Iso-Ahola, S. and Dotson, C., (2014). Psychological Momentum: Why Success Breeds Success. Review of General Psychology, 18(1), pp.19-33

Iso-Ahola, S. and Mobily, K., (1980). "Psychological Momentum": A Phenomenon and an Empirical (Unobtrusive) Validation of its Influence in a Competitive Sport Tournament. Psychological Reports, 46(2), pp.391-401.

Kahneman, D. (2011). Thinking fast and slow. New York, NY: Farrer, Straus, and Girox.

Kahneman, D., and Riepe, M. W., (1998). "Aspects of Investor Psychology: Beliefs, Preferences, and Biases Investment Advisors Should Know About," Journal of Portfolio Management, 24, pp.1–21.

Kahneman, D., and Tversky, A., (1971). Belief in the law of small numbers. Psychological Bulletin, 76(2), pp.105-110.

Kahneman, D., and Tversky, A. (1979). On the interpretation of intuitive probability: A reply to Jonathan Cohen. Cognition, 7(4), pp.409-411

Kotu, V., and Deshpande, B., (2019). Data Science (second Edition). Morgan Kaufmann, Chapter 4, pp 65-163.

Koehler, J. and Conley, C., (2003). The "Hot Hand" Myth in Professional Basketball. Journal of Sport and Exercise Psychology, 25(2), pp.253-259.

Korb, k. b., and Stillwell, M. (2003). The story of the hot hand: Powerful myth or powerless critique. In international conference on cognitive science.

Lantis, R. and Nesson, E., (2021). The Hot Hand in the NBA 3-Point Contest: The Importance of Location, Location, Location. SSRN Electronic Journal.

Laplace, P. S., (1951). A philosophical essay on probabilities (Trans. F.W. Truscott and F.L. Emory). New York: Dover (Original work published 1814).

Lee, M., AND Smith, G,. (2002). "Regression to the Mean and Football Wagers," Journal of Behavioral Decision Making, 15, pp.329–342.

Mack, M. G. and Stephens, D. E., (2000). An empirical test of Tylor and Demick's multidimensional model of momentum in sport. Journal of Sport Behavior, 23(4), pp.349-363.

Miller, J. and Sanjurjo, A., (2014). A Cold Shower for the Hot Hand Fallacy. SSRN Electronic Journal,. Working Paper.

Miller, J. and Sanjurjo, A., (2015). Surprised by the Gambler's and Hot Hand Fallacies? A Truth in the Law of Small Numbers. SSRN Electronic Journal,. pp.1-14.

Miller, J. and Sanjurjo, A., (2018). Surprised by the Hot Hand Fallacy? A Truth in the Law of Small Numbers. Econometrica, 86(6), pp. 2019-2047.

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

Morgulev, E., Azar, O, H., and Bar-Eli, M. (2020). Searching for momentum in NBA triplets of free throws. Journal of Sports Sciences, 38(4), pp.390-398.

Morgulev, E. and Avugos, S., (2020). Beyond heuristics, biases and misperceptions: the biological foundations of momentum (hot hand). International Revew of Sport and Exercise Psychology, pp.1-21.

Raab, M., Gula, B., and Gigerenzer, G., (2012). The hot hand exist in volleyball and is used for allocation decisions. Journal of Experimental Psychology, 18(1), pp.81-94.

Rabin, M. and Vayanos, D., (2010). The gambler's and Hot-Hand Fallacies: Theory and Applications. Review of Economic Studies, 77(2), pp.730-778.

Shefrin, H., (2009). How Psychological Pitfalls Generated the Global Financial Crisis. SSRN Electronic Journal, 10(4), pp.1-40.

Sinkey, M., and Logan, T., (2013). "Does the Hot Hand Drive the Market?" Eastern Economic Journal, Advance. online publication.

Stone, D. F, (2012). "Measurement Error and the Hot Hand," The American Statistician, 66, pp.61-66.

Wardrop, R., (1995). Simpson's Paradox and the Hot Hand in Basketball. The American Statistician, 49(1), pp.24-28.

Warpdrop, R., (1999). "Statistical Tests for the Hot-Hand in Basketball in a Controlled Setting," Unpublished manuscript, 1, pp. 1–20.

Wayne, A. Fuller and Michael, A,. (1978). Regression Estimation after Correcting for Attenuation, Journal of the American Statistical Association, 73(261), pp.99-104.

Wood, R. and Stanton, S., (2012). Testosterone and sport: Current perspectives. Hormones perspectives. Hormones and Behavior, 61(1), pp.147-155.

Xu, J., and Harvey, N., (2014). "Carry on Winning: The Gambler's Fallacy Creates Hot Hand Effects in Online Gambling," Cognition, 131, pp.173-180.

Yuan, J, Sun, G-Z, and Siu, r., (2014). "The Lure of Illusory Luck: how much are people willing to pay for random shocks," Journal of Economic Behavior & Organization, 106, pp.269-280.

Appendix

Appendix A

TABLE 4 – PLAY-BY-PLAY SUMMARY STATISTICS

Variable	Non-missing observations	Type	Description
Number of Play ID	6.596.414	Numeric	All events in a game have an id number
Shots	2.715.094	Numeric	Total number of shots that was taken within a game
Result	2.715.094	Categorical	Shot "made" or "missed"
Game ID	12.581	Categorical	Total number of games played
Player ID	1.333	Categorical	Total number of players in the data set
X Coordinates	2.364.412	Numeric	X axis value of the shot
Y Coordinates	2.364.412	Numeric	Y axis value of the shot
Shot Distance	2.364.412	Numeric	Shot distance (feet)
Location	6.596.414	Binary	If the team playing at home (=1) or away)=0)
Time Remaining	6.596.414	Numeric	Remaining time in the quarter
Score Margin	6.596.414	Numeric	Score difference of the teams at the moment of the shot
Shot type	2.715.094	Categorical	The type of shot. Layup, Dunk, 3 Point-Shot etc.
Shot Clock	6.596.414	Numeric	Length of the possession where the shot was recorded in seconds
Team	6.596.414	Categorical	Indicates which of the 30 NBA teams took the shot
Quarter	6.596.414	Categorical	Period in which the shot was taken (1-4) Overtime (5-6)
Closest Defender	2.364.412	Numeric	Distance between the shooter and the closest defender

.

TABLE 5 – TOTAL SHOOTING PERCENTAGE SUMMARY STATISTICS

Variable:	Mean	SD	p25	p50	p75	p99
Total Shooting percentage	.462	.046	.430	.450	.487	.615

Notes: This table provides summary statistics of NBA players' TS% and FG% over the duration of the entire sample period (2011-2021).

Appendix B

TABLE 6 – DO PLAYERS HAVE A HOT HAND K=1

	All Season	L=40	L=10	L=5
	(1)	(2)	(3)	(4)
Dependent variable:	(Hit|1Hit)	(Hit|1Hit)	(Hit|1Hit)	(Hit|1Hit)
State	-0.013***	-0.001***	0.006***	0.012***
	(0.001)	(0.001)	(0.001)	(0.003)
Shot Clock	-0.254***	-0.221***	-0.226***	-0.203***
	(0.003)	(0.012)	(0.013)	(0.012)
Shot Distance	-0.008***	-0.008***	-0.008***	-0.008***
	(0.000)	(0.000)	(0.000)	(0.000)
Time Remaining	0.001***	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Score Margin	-0.006***	0.004***	0.004***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
Location	0.004***	0.004***	0.004***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
Defender Distance	0.001*	-0.001*	-0.001*	-0.001*
	(0.000)	(0.000)	(0.001)	(0.000)
Constant	0.096***	0.098***	0.104***	0.114***
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	2.460.647	2.460.647	2.460.647	2.460.647
R-squared	0.044	0.044	0.044	0.043

Notes: This table summarizes the estimates for all player regressions using a single past hit as the dependent variable. Column (1) contains the estimates from the WLS regression over the whole season, likewise, for Column (2) - (4). All sequences disrupted by the previous shot and the first shot in the subsequent game produce a missing value that is removed. Standard errors in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

TABLE 7 - DO PLAYERS HAVE A HOT HAND K=2

	All Season	L=40	L=10	L=5
	(1)	(2)	(3)	(4)
Dependent Variable:	(Hit|2Hit)	(Hit|2Hit)	(Hit|2Hit)	(Hit|2Hit)
State	-0.001***	0.001***	0.004***	0.010***
	(0.001)	(0.001)	(0.001)	(0.003)
Shot Clock	-0.272***	-0.261***	-0.210***	-0.203***
	(0.003)	(0.012)	(0.013)	(0.012)
Shot Distance	-0.008***	-0.008***	-0.008***	-0.008***
	(0.000)	(0.000)	(0.000)	(0.000)
Time Remaining	0.002***	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Score Margin	-0.005***	0.004***	0.004***	0.004***
	(0.000)	(0.000)	(0.000)	(0.001)
Location	-0.006***	0.006***	0.005***	0.005***
	(0.001)	(0.001)	(0.001)	(0.001)
Defender Distance	0.001	-0.001	-0.001	-0.001
	(0.000)	(0.000)	(0.001)	(0.000)
Constant	0.092***	0.095***	0.099***	0.107***
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	2.218.555	2.218.555	2.218.555	2.218.555
R-squared	0.044	0.043	0.044	0.043

Notes: This table summarizes the estimates for all player regressions using a single past hit as the dependent variable. Column (1) contains the estimates from the WLS regression over the whole season, likewise, for Column (2) - (4). All sequences disrupted by the previous shot and the first two shots in the subsequent game produce a missing value that is removed. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

TABLE 8 - DO PLAYERS HAVE A HOT HAND K=3

	All Season	L=40	L=10	L=5
	(1)	(2)	(3)	(4)
Dependent Variable:	(Hit|3Hit)	(Hit|3Hit)	(Hit|3Hit)	(Hit|3Hit)
State	-0.006***	-0.001***	0.002***	0.007***
	(0.001)	(0.001)	(0.001)	(0.001)
Shot Clock	-0.277***	-0.277***	-0.276***	-0.276***
	(0.004)	(0.004)	(0.004)	(0.004)
Shot Distance	-0.008***	-0.008***	-0.008***	-0.008***
	(0.000)	(0.000)	(0.000)	(0.000)
Time Remaining	0.002***	0.002***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Score Margin	-0.004***	0.004***	0.004***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
Location	0.007***	0.007***	0.004***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
Defender Distance	0.001	-0.001	-0.001	-0.001
	(0.000)	(0.000)	(0.001)	(0.000)
Constant	0.094***	0.092***	0.090***	0.091***
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	1.989.891	1.989.891	1.989.891	1.989.891
R-squared	0.044	0.043	0.044	0.044

Notes: This table summarizes the estimates for all player regressions using a single past hit as the dependent variable. Column (1) contains the estimates from the WLS regression over the whole season, likewise, for Column (2) - (4). All sequences disrupted by the previous shot and the first three shots in the subsequent game produce a missing value that is removed. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Appendix C

The metrics is generated by evaluating the players performance by looking at (1) PPG = points per game (2) TS% (3) highest autocorrelation in the weighted least squares regression.

The following order applies in the choice of players:

(1) Most overlapping (i.e., name comes up in most criteria)

(2) Highest autocorrelation

(3) TS%

(4) PPG

TABLE 9 - HIGHEST AUTOCORRELATION IN THE WEIGHTED LEAST SQUARED

Player	(1) Coef	(2) T Statistics	(3) P Value	(4) Shot Count	(5) Shot per game
Stephen Curry	0.055	1.978	0.048**	1351	20.78
LeBron James	0.055	2.022	0.043**	1474	20.19
Kevin Durant	0.067	2.528	0.017**	750	15.01
James Harden	0.077	2.028	0.009***	787	15.74
Kyrie Irving	0.046	2.120	0.034**	416	11.74
Klay Thompson	0.089	7.442	0.017**	1404	18.10
Joel Embiid	0.066	2.110	0.035**	1034	16.67
Nikola Jokic	0.080	-2.853	0.004***	1439	17.54
Zach Lavine	0.063	-1.999	0.047**	1065	18.36
Gianni Antetokounmpo	0.065	-2.346	0.019**	1281	17,79

Notes: The above table shows players that were chosen in the metrics and their separate coefficients from the WLS. Column (1) shows their individual state during the regression. Column (2)-(3) are the statistics that follow. Column (4) represents the total number of shots they took during the season they were picked from in the regression and Column (5) represents the average amount of shots taken during the same year. The autocorrelation picked at k=1, L=5. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1.

FIGURE 10 - POINTS PER GAME

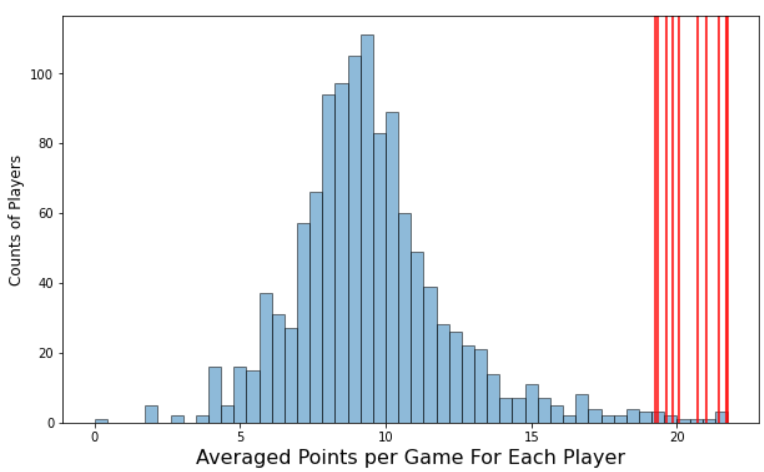
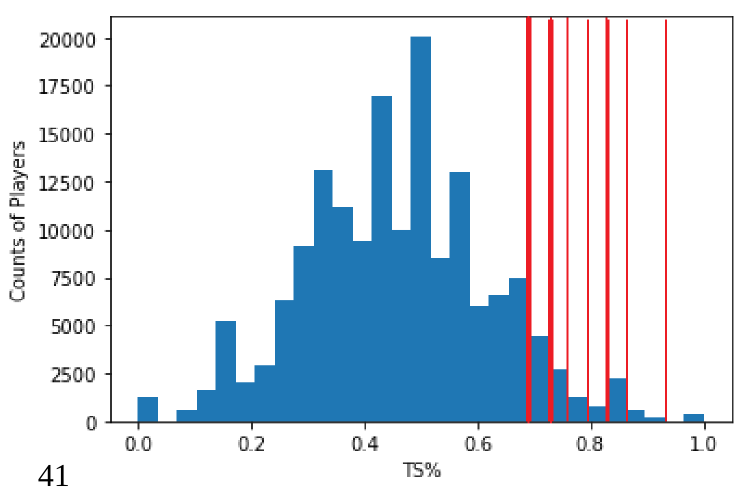


FIGURE 11 - TOTAL SHOOTING PERCENTAGE



Appendix D

TABLE 10 - NUMBER OF SHOTS (RDD)

	(1)	(2)	(3)
Dependent Variable:	Full Data	BW=2	BW=1
Streakiness	-1.830***	-2.816***	-0.900***
	(0.069)	(0.070)	(0.096)
Constant	10.959***	11.589***	11.396***
	(0.031)	(0.033)	(0.039)
Observations	73.723	69.823	47.515
R-squared	0.203	0.220	0.039

Notes: Number of observations is the total length of a sequence and not number of shots. Column (1) shows the results with all sequences. Column (2) shows the result with a bandwidth of 2 and Column (3) shows the result closest to the threshold. The threshold was set to 3.5. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE 11 - SHOT DISTANCE (RDD)

	(1)	(2)	(3)
Dependent Variable:	Full Data	BW=2	BW=1
Streakiness	-0.260*	-0.406***	-0.254
	(0.077)	(0.080)	(0.043)
Constant	15.582***	15.634***	15.139***
	(0.034)	(0.037)	(0.043)
Observations	73.723	69.823	47.515
R-squared	0.203	0.220	0.039

Notes: Number of observations is the total length of a sequence and not number of shots. Column (1) shows the results with all sequences. Column (2) shows the result with a bandwidth of 2 and Column (3) shows the result closest to the threshold. The threshold was set to 3.5. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE 12 - PLAYERS TOUCH (RDD)

	(1)	(2)	(3)
Dependent Variable:	Full Data	BW=2	BW=1
Streakiness	8.330***	4.738***	6.935***
	(0.175)	(0.184)	(0.235)
Constant	41.371***	42.366***	41.788***
	(0.078)	(0.085)	(0.094)
Observations	73.723	69.823	47.515
R-squared	0.546	0.502	0.310

Notes: Number of observations is the total length of a sequence and not number of shots. Column (1) shows the results with all sequences. Column (2) shows the result with a bandwidth of 2 and Column (3) shows the result closest to the threshold. The threshold was set to 3.5. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE 13 - CLOSEST DEFENDER DISTANCE (RDD)

	(1)	(2)	(3)
Dependent Variable:	Full Data	BW=2	BW=1
Streakiness	-0.047*	-0.058	-0.049
	(0.008)	(0.008)	(0.011)
Constant	2.261***	2.265***	2.251***
	(0.003)	(0.004)	(0.004)
Observations	73.723	69.823	47.515
R-squared	0.002	0.002	0.004

Notes: Number of observations is the total length of a sequence and not number of shots. Column (1) shows the results with all sequences. Column (2) shows the result with a bandwidth of 2 and Column (3) shows the result closest to the threshold. The threshold was set to 3.5. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE 14 - MINUTES PLAYED (RDD)

	(1)	(2)	(3)
Dependent Variable:	Full Data	BW=2	BW=1
Streakiness	1.830***	0.418	0.344
	(0.077)	(0.081)	(0.100)
Constant	17.338***	17.898***	17.947***
	(0.034)	(0.038)	(0.040)
Observations	73.723	69.823	47.515
R-squared	0.411	0.385	0.232

Notes: Number of observations is the total length of a sequence and not number of shots. Column (1) shows the results with all sequences. Column (2) shows the result with a bandwidth of 2 and Column (3) shows the result closest to the threshold. The threshold was set to 3.5. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.