Research article

Gregory M. Steeger, Johnathon L. Dulin and Gerardo O. Gonzalez*

Winning and losing streaks in the National Hockey League: are teams experiencing momentum or are games a sequence of random events?

<https://doi.org/10.1515/jqas-2020-0077>

Received April 8, 2020; accepted April 14, 2021;

published online May 7, 2021

Abstract: The Saint Louis Blues were hot at the end of the 2018–2019 National Hockey League season, winning eleven games in a row in January and February, and eight of their last ten. They parlayed this momentum to their first Stanley Cup Championship in franchise history. Or did they? Did the series of wins at the end of the season give the Blues the momentum needed to reach the pinnacle of the sport on June 12th, or was the Blues' path to victory the confluence of a series of random events that fell in their favor? In this paper we apply entropy as an unbiased measure to further refute the idea of momentum in sports. We show that game outcomes are not dependent on previous games' outcomes and conclude that the theory of momentum, across the season, is a fallacy that should not affect behavior.

Keywords: entropy; gambler's fallacy; hot hand fallacy; runs test.

1 Introduction

There is no shortage of academic literature on momentum in sports. Despite many studies that have concluded that sports momentum is a fallacy, most enthusiasts fail to accept that momentum does not exist. Amos Tversky may have said it best when he stated, "I've been in a thousand arguments over this topic. I've won them all and I've convinced no one" (Moskowitz and Wertheim 2012). Despite

decades of research, the debate is ongoing (Hales 1999). In fact, though the seminal research on the topic concluded that sports momentum is a fallacy and was published over thirty-five years ago (Gilovich et al. 1985), books continue to be published on the subject (Cohen 2020).

Researchers that study momentum in sports typically use one of five methods: conditional probabilities; autocorrelation tests; Wald-Wolfowitz Runs Tests; Chi-Squared Goodness of Fit Tests; or Markov analyses (Vergin 2000; Zhang et al. 2013). Many researchers do not believe these methods are powerful enough to adequately test for momentum in sports (Dorsey-Palmateer and Smith 2004; Huitema et al. 1996; Miyoshi 2000; Wardrop 1999; Zhang et al. 2013). Zhang et al. (2013) introduce new measures that detect unnatural groupings of events in data to measure momentum and overcome the limitations of existing methods. In this paper, we study momentum in hockey by using one of these new measures, namely entropy. We focus on the Saint (St.) Louis Blues' performance in the 2018–2019 season, because of how their season progressed.

On January 2nd, 2019, the St. Louis Blues were in last place in the National Hockey League (NHL) – 31st, out of 31 teams. The Blues only had 34 points with a record of 15 wins, 18 losses, and four overtime losses (Thomas 2019). By the end of January the Blues had moved up to sixth place in the Central Division, but were only given 100 : 1 odds of winning the Stanley Cup (Greenberg 2019). Despite these odds, on the 21st of May, the Blues beat the San Jose Sharks in a seven-game series to win the Western Conference finals and, on the 12th of June, they defeated the Boston Bruins in game seven of the finals to win the Stanley Cup and the first championship in the franchise's history. The question is, "How did the St. Louis Blues defy the odds and become the best team in the NHL?" Was it the change in the coaching staff, the change in the starting goaltender, the fact that they started to play Laura Branigan's song *Gloria* after every win (USA Today 2019), or was the team gaining

*Corresponding author: Gerardo O. Gonzalez, Department of Management, United States Air Force Academy, 2354 Fairchild Drive, USAF Academy, Colorado, USA, E-mail: gerry.gonzalez@usafa.edu

Gregory M. Steeger and Johnathon L. Dulin, Department of Management, United States Air Force Academy, 2354 Fairchild Drive, USAF Academy, Colorado, USA, E-mail: gsteeger@outlook.com (G. M. Steeger), Johnathon.Dulin@usafa.edu (J. L. Dulin)

momentum after every win? In this paper we ask, “Are NHL team’s wins and losses influenced by momentum?”

Many believe that the St. Louis Blues gained momentum in the second half of the season and carried that momentum into the playoffs (Greenberg 2019). The team recorded the longest winning streak in the franchise’s history by winning eleven consecutive games, from the 23rd of January to the 19th of February (Pinkert 2019). No other team in the NHL, in the 2018–2019 season, put together a winning streak that lasted longer. While it is clear that teams have winning and losing streaks, sequences of successive wins and losses, it is not clear what factors may contribute to those streaks. Hockey professionals, analysts, and enthusiasts point to a number of situational effects that may lead to wins and losses, successive or not: location of the game (home or away), quality of the opponent (i.e., winning percentage), scheduling factors (e.g., third game in four nights), injuries, and even psychological factors (when one team has a winning streak against another, regardless of other factors). But, in addition, those same people often point to momentum as a key driver of the outcome of any game.

We make an important distinction between streaks and momentum. Streaks are observed sequences of wins or losses that may or may not be dependent on each other, whereas momentum implies dependence between similar events. Sports psychologists use the term *psychological momentum* to describe changes in performance based on success or failure of recent events (Vergin 2000). More precisely, Adler (1981) states that psychological momentum is the tendency of an effect to be followed by a similar effect. If positive momentum exists then a success would increase the probability of another success, while negative momentum would make it so a failure would increase the probability of a subsequent failure (Adler 1981). The difficulty with determining if St. Louis and their road to the Stanley Cup was impacted by momentum is trying to do so objectively without letting intuition and our poor understanding of randomness cloud our judgement. Gilovich et al. (1985) further explain our poor understanding of randomness by stating, “*People’s intuitive conceptions of randomness depart systematically from the laws of chance. It appears that people expect the essential characteristics of a chance process to be represented not only globally in the entire sequence, but also locally, in each of its parts.*”

If sports momentum does not exist and instead the events are random, then perhaps behavior should be adjusted accordingly. It is important for owners, coaches, players, and fans to understand the differences between a sequence of random, independent events and a sequence

of events that are dependent on each other. Momentum is widely believed to influence the outcomes in sports and decisions are made based on these beliefs (Vergin 2000). Sports participants and observers adjust their behavior based on whether they believe their team is carrying momentum or not. Owners make hiring and firing decisions based on perceived positive and negative momentum, coaches adjust lineups and strategies to take advantage of players that are “streaky,” and players and fans often follow all sorts of strange rituals to prolong hot streaks or end cold streaks.

A common cliché in hockey (and other sports) is that a team needs to be hot going into the playoffs. But with playoff positioning secure, coaches are often tempted to rest their star players in the last few games for rest and/or recovery, or in an attempt to avoid injuries. But if this interrupts momentum, is it worth it? In the 2018–2019 NHL season, the Tampa Bay Lightning clinched the President’s Trophy with nine games remaining. The trophy is awarded to the team with the most points in the league and guarantees the team home-ice advantage throughout the playoffs. The Lightning’s coaching staff chose to keep most of their star players in the lineup throughout the playoffs, opting to pursue a number of records and maintain the momentum of a record-setting season. The Lightning finished the season winning seven of the last 10 games, consistent with what they had done during the entire season, yet were swept in the first round of the playoffs. Would giving the star players a few games off have made a difference? Resting key players and potentially losing more games towards the end of the season might have upset the fan base, but it might also have led to a fresher team going into that ill-fated series against Columbus.

Because their seasons consist of a sufficiently large number of games, when analyzing the existence of team momentum in winning and losing streaks, the team sports that are studied most often are basketball and baseball (Vergin 2000). There are few studies that analyze hockey, mainly because of the difficulties associated with modeling ties and overtime losses. Ties were eliminated in 2005, but overtime losses still create some difficulties. In the NHL, each game results in one of three outcomes: a win, a loss, or an overtime loss. If the game is tied at the end of regulation, each team will receive one point, and the teams will play one five-minute “sudden death” overtime period. The first team to score during the overtime period wins. If neither team scores during the overtime period then there will be a shootout in which each team gets three shots. If the score remains tied after the initial shootout then there will be a “sudden death” shootout. An overtime loss is not as good

as a win, but also not as bad as a regulation-time loss; teams are awarded two points for a win, no points for a loss, and one point for an overtime loss.

The remainder of this paper is organized as follows: Section 2 briefly reviews the literature on the psychology of momentum and on momentum in sports. Next, Section 3 presents the methodology we use to test for momentum. Section 4 then uses the methodology to analyze NHL teams from the 2018–2019 season. Following this, Section 5 provides a discussion on the results, and, last, Section 6 concludes the paper and provides areas for future work.

2 Literature review

To gain a better understanding of the topic, we review how the study of momentum has evolved over time. When we write about random sampling, we are generally referring to a number obtained from a probability distribution in a way so that it has no correlation with a previous number obtained from that same distribution. When we state that a process appears to be random, we do this based on measurement techniques that do not allow us to reject the observation of outcomes arising from patterns similar to those found from random sampling. The measurement techniques (explored later in this section) look for either random sequences or irregular clusters.

It does not seem possible to discuss momentum in sports without first discussing some of the foundational psychology literature on randomness. What is fascinating about this review is how the psychology literature naturally leads to the study of momentum in sports. Researchers in the field of psychology started studying people's understanding of random events in the late 1960s. However, Laplace questioned people's understanding of random events long before that, in the 1800s (Laplace 1951). Tversky and Kahneman (1971) find that people, even trained statisticians, have misperceptions about random sampling. Tversky and Kahneman (1974) go on to suggest that these misperceptions occur because people rely on heuristics to assess probabilities and that the commonly-used heuristics can lead to severe systematic errors. Oskarsson et al. (2009) expand on the causes of these misperceptions by stating that human beings use their senses to see patterns and use this information to predict future events, but this ability does not always lead to a perfect understanding of the outside world. Additionally, Tversky and Kahneman (1971) suggest that people view a random sample as representative of the population, even when the sample sizes are small. As an example, when subjects are instructed to construct random sequences of

coin flips, they construct sequences where the proportion of heads remains close to 50% and “long” sequences of heads or tails do not occur (Kahneman and Tversky 1972; Tunc 1964). In other words, every small sample of the sequence is representative of the population and the population's associated proportions.

Estes (1964) finds this representativeness error also to be true when subjects predict successive random events and/or outcomes of sequential games of chance; subjects create sequences of events so that every segment, or sample of the random sequence, reflects the population. When the sequence generates a proportion that deviates from the population proportion, they believe that a corrective action must occur (Estes 1964). This is known as the *gambler's fallacy*. Ayton and Fischer (2004) define the gambler's fallacy as “the belief that, for random events, runs of a particular outcome (e.g., heads on the toss of a coin) will be balanced by a tendency for the opposite outcome (e.g., tails).” More generally, this is referred to as *negative recency*.

Similarly, *positive recency* is the tendency to predict future outcomes the same as the most recent outcomes (Ayton and Fischer 2004). Positive recency is also referred to as the *hot-hand fallacy* (Ayton and Fischer 2004). The *hot-hand effect* is the belief that players will have success in streaks (whether success is defined as shooting a basketball through a hoop, shooting a puck into a net, kicking a football through field goal posts, or something else) and that those streaks will impact future successes. Bar-Eli et al. (2006) provide a thorough review on hot-hand literature up to 2006. Wald-Wolfowitz Runs and autocorrelation tests are common tools used to determine if a player has a “hot hand” (Raab et al. 2012; Stone 2012). Gilovich et al. (1985) are the first to study the hot-hand effect and team momentum from game to game in basketball. Gilovich et al. (1985) conclude that neither player streaks nor team streaks are predictive of future player or team performance. In sports it is easy to come up with reasonable explanations as to why successes or failures, for individuals or teams, would be dependent on each other. This dependency could be attributed to confidence or fleeting confidence, muscle memory, and/or fatigue (Ayton and Fischer 2004). However, all of these cases should show evidence of positive recency (Ayton and Fischer 2004). Other literature on human behavior refutes the work above that attributes human error in assessing probability to incorrectly using a representativeness heuristic. Gigerenzer et al. (1988) suggest that humans may not rely on this heuristic if they have the proper information and proper presentation of the required information. We present this counter argument, not to refute

the findings on negative recency and positive recency, but to expose a different and more complex perspective on human behavior.

Undoubtedly, streaks exist in sports. Most streaks seen in sports are not surprising and are even expected, but we sometimes experience unexpected streaks (Reifman 2011). At the individual level, Reifman (2011) calls these streaks extreme events, stating the chance of the streak occurring is extremely improbable and even close to miraculous. The most well-known extreme event is Joe DiMaggio's 56-game hitting streak in 1941. Computer simulation shows that, when considering all relevant base-rate information, DiMaggio's streak had a one in 5000 seasons chance of occurring. But when looking at all players and all seasons, a streak like DiMaggio's is not unlikely (Reifman 2011). Reifman (2011) issues one final note about only using mathematics and simulation to gauge these probabilities; these techniques cannot easily capture the effects of human awareness. Once coaches, opponents, and even the players themselves know about the streak, behaviors may change to either help or hurt the possibility of the streak continuing (Reifman 2011). Reifman (2011) goes on to talk about the much harder task of gauging probabilities of streaks related to team performance.

Utilizing big data techniques, Yaari and Eisenmann (2011) use five seasons worth of free-throw data to argue that seemingly non-random patterns in the data are more likely evidence of "better" or "worse" periods than they are of momentum. Similarly, Attali (2013) use an entire season of National Basketball Association (NBA) data to analyze pairs of consecutive shots to conclude that the hot-hand belief impacts players' shot allocations, but decreases the likelihood that the shots are successful. With few exceptions, sports researchers do not find statistical evidence to support positive recency (Moskowitz and Wertheim 2012; Tversky and Gilovich 1989; Vergin 2000). In fact, Avugos et al. (2013) use a meta-analytic approach, studying over 250 papers on the subject to support the claim that in most cases the hot-hand effect does not exist in sports.

Some believe that skeptics are looking for too rare of events when attempting to detect hot hands in sports. Hales (1999) suggests that while skeptics have proven that recent success does not impact the likelihood of future success and that long streaks are not statistically unexpected, we may need to reconsider how we define what a hot hand is. Hales (1999) goes on to suggest that a player has a hot hand when he or she is simply playing better than average and that both players and fans can detect when a player is playing better than average. For these reasons, Hales (1999)

concludes that the belief that the hot-hands effect exists is justified.

Though it is rare in sports, positive recency has been found to exist in golf putting and dart throwing (Gilden and Wilson 1995), billiards (Adams 1995), horse-shoes (Smith 2003), and bowling (Dorsey-Palmateer and Smith 2004). Note that these exceptions occur in individual sports, as opposed to team sports, in which defensive play does not dramatically interfere with one's play (Csapo et al. 2015). Iso-Ahola and Dotson (2014) discuss why psychological momentum occurs more in individual sports than it does in team sports.

There are a few exceptions where researchers do find evidence of psychological momentum in team sports. Arkes (2010) furthers hot-hand research by testing for conditional probabilities in a multivariate framework. Arkes' (2010) research focuses on basketball free throws using NBA data from 2005 to 2006. The research is one of the first to find positive recency in an example hampered with interference by defensive play. By pooling all shooting data for all players throughout the entire season, Arkes is able to overcome deficits from previous studies with insufficient power to detect evidence of a hot hand. Arkes and Martinez (2011) further NBA momentum research by developing an economic model using correlation tests. They propose that, with a sufficiently large sample (2007–2009 seasons), positive momentum between wins can be displayed. Arkes and Martinez (2011) find positive momentum effects for winning by accounting for strength of opponent and by pooling teams in home versus away matchups. However, the authors also state that their results may not hold for other sports and that their results do not account for changes in team composition. Another exception is volleyball (Raab et al. 2012). Raab et al. (2012) conclude that the hot-hand exists in volleyball and that both players and coaches can detect it and take advantage of it. Stone (2012) provides yet another exception and concludes that, due to measurement error, previous studies prevent the hot-hand effect from being observed. Similar to Hales (1999), Stone (2012) proposes an alternative definition for the hot hand based on autocorrelation; he suggests that the hot hand exists if the probability of a shot being successful is positively correlated with the probability of the next shot being successful. Stone (2012) uses an autocorrelation model to show that measurement error exists in basketball and finds strong evidence for the hot-hand effect when that measurement error is removed. Last, Bocskocsky et al. challenge extant literature on the hot-hand fallacy in basketball by challenging the assumption that player shot selection is made independent of the

player's belief of being "hot" or "cold." Bocskocsky et al. suggest that when shot difficulty is controlled for, "hot" players are more likely to make subsequent shots. In other words, Bocskocsky et al. argue that the hot-hand effect is real in basketball.

In our research we make comparisons between our methodology and the Wald-Wolfowitz Runs Test since it is often used for finding unlikely streaks in sequences of games (Peel and Clauset 2015). Runs tests are used for finding unlikely streaks since, in general, they look for uninterrupted sequences of alike elements bordered at each end by elements of different types (Koutras et al. 2007). Using this technique in sports applications allows analysts to identify either positive momentum or anti-restorative effects in within-game scoring (Peel and Clauset 2015). The Runs Test is also well known in the field of quality control. Specifically, its use dates as early as the 1940s in testing of manufactured articles when trying to ascertain if measurement differences in the same inspection point are random or whether they contain trends (Wolfowitz 1943). More recently, the field of quality control found benefit in using Hotelling's chi-square control charts and Markov chain approaches (Koutras et al. 2007). These techniques have also made it into sports analysis.

Along with Runs Tests, Chi-Squared Goodness of Fit Tests, and Markov Analyses, sports research on momentum typically utilizes conditional probabilities and autocorrelation tests (Stone 2012; Vergin 2000; Zhang et al. 2013). However, because all of these methods are limited by stationarity (*i.e.*, the assumption that success probabilities do not change over time), many researchers do not believe they are powerful enough to adequately test for momentum in sports (Dorsey-Palmateer and Smith 2004; Huitema et al. 1996; Miyoshi 2000; Wardrop 1999; Zhang et al. 2013). Zhang et al. (2013) define data *clumpiness* as irregular clusters of activity that are gathered together. More specifically, they state "clumpiness indicates non-constant propensity, specifically temporary elevations of propensity – *i.e.*, periods during which one event is more likely to occur than the average level." Zhang et al. (2013) further suggest that entropy is a better measure for momentum in sports. We agree with this conclusion and base our methodology largely on the premise that we do not want stationarity to be a limiting factor of our work.

In general, the term entropy is used to describe a system that lacks order or predictability. In thermodynamics, entropy measures the amount of disorder or randomness in a system (Key and Ball 2014). Entropy is also used in

information theory, the mathematical theory of communication, to measure the amount of information received from a random process (Gray 2011). Players that experience the hot-hand effect or teams that are influenced by momentum should generate data sets with activities that are grouped together, or clumpy data.

Zhang et al. (2013) expand on the limitations of existing methods that seek to detect momentum (or data clumpiness) by stating that existing measures do not possess two desired properties: continuity and convergence. Both of these properties can exist in a measure so long as the measure is based on inter-event times or the amount of time that elapses between event occurrences. Continuity ensures that when event occurrence times shift by a small amount, the measure only changes by a small amount. Convergence ensures that as events move closer together (further apart) the measure increases (decreases). These properties, Zhang et al. (2013) argue, make entropy a better measure to use when testing for momentum. Zhang et al. (2014) use computer-generated clumpy and non-clumpy data sets to test different measures and conclude that entropy is better at detecting clumpy data than other, widely-used measures.

3 Methodology

To measure whether or not a team's wins and losses were impacted by momentum we use the entropy measure described in Zhang et al. (2013). Entropy is defined as

$$H_p = \sum_{i=1}^{n+1} x_i \log(x_i),$$

where x_i is the inter-event time of the i th event. More specifically,

$$x_i = \begin{cases} t_1, & \text{if } i = 1 \\ t_i - t_{i-1}, & \text{if } i = 2, \dots, n, \\ N + 1 - t_n, & \text{if } i = n + 1, \end{cases}$$

where t_i is the occurrence time of the i th event, and n and N are the number of events and the total number of trials (Zhang et al. 2014). For the purposes of this paper, an event is a win, t_i is the match number of the i th win, n is the number of wins, and N is the total number of games played. Because the entropy measure applies when dealing with binary situations, *e.g.*, wins and losses, we redefine the outcomes that hockey teams can achieve in the regular season as "wins" and "not wins." Not only does this approach allow us to measure the impact of momentum as described, it also captures the prevailing opinion of hockey

players, teams, coaches and fans, reinforced by the fact that it earns only a single standings point; an overtime loss is simply not the same as a win.

As an example, the St. Louis Blues lost seven of their first 10 games. The sequence was L, L, W, L, L, L, W, L, L, W. At the end of this sequence, $n = 3$ and $N = 10$. x_1 , the inter-event time for the first win, equals t_1 , which is equal to 3. Similarly, $x_2 = t_2 - t_1 = 4$, and $x_3 = t_3 - t_2 = 3$. Last, $x_{n+1} = x_4 = N + 1 - t_3 = 10 + 1 - 10 = 1$.

To scale the entropy measure, so that each inter-event time is computed relative to the number of trials (see Zhang et al. 2014), we divide each x_i by $N + 1$. Mathematically, for $i = 1, \dots, n + 1$, we compute scaled inter-event times as

$$\hat{x}_i = \frac{x_i}{N + 1}.$$

Scaling the measure in this way also ensures all of the scaled $\hat{x}_i$ sum to one. In other words,

$$\sum_{i=1}^{n+1} \hat{x}_i = 1.$$

Finally, the normalized entropy is computed as

$$\hat{H}_p = 1 + \frac{\sum_{i=1}^{n+1} \hat{x}_i \log(\hat{x}_i)}{\log(n + 1)}.$$

Generally, teams that experience more momentum throughout their seasons will have wins and losses clumped more tightly together, and consequently larger $\hat{H}_p$ values, than teams that experience less momentum.

$\hat{H}_p$ displays a few interesting characteristics. First, it does not rely on stationarity. In the context of our application, this means that the winning percentage does not have to be fixed to measure a team's entropy. Second, $\hat{H}_p$ is minimized when events are equally spaced and maximized when all events are grouped together (Zhang et al. 2013). Third, $\hat{H}_p$ decreases with n , the number of events that occur. This means that for two teams that exhibit an equal amount of momentum, the team with the lower winning percentage will have a larger value for $\hat{H}_p$. Because we define an event as a win, teams with higher winning percentages are expected to have more wins (events) clumped together than teams with lower winning percentages. Fourth, as discussed at the end of Section 2, the measure demonstrates the properties of continuity and convergence (Zhang et al. 2013). Last, streaks of events (or non-events) will have a larger impact on the measure when the number of trials (N) is small. In other words, streaks of events (or non-events) will not have a significant impact on the measure until the probability of them occurring is "sufficiently" small.

$\hat{H}_p$, is calculated for each of the teams at a set point in the season and then used as a test statistic, in the following hypothesis test:

H_0 : wins and losses are random

H_a : wins and losses are not random.

Like in many hypotheses tests the test statistic, $\hat{H}_p$, alone does not provide a great deal of insight into the process being studied. This is especially true in our case, because of the inverse relationship between the measure and the team's number of wins (or the winning percentage). It is only fair to compare entropy measures for two teams if the two teams have the same winning percentage. In all other cases we must either compare the test statistic to a critical value or compute an associated p -value. We choose to compute p -values.

To determine how $\hat{H}_p$ is distributed under the null hypothesis, we simulate each team's 82-game season, based on their overall win percentage. The outcome of each of the 82 games in the simulation is randomly selected where the probability that "win" is selected equals the team's winning percentage. Consequently, neither positive nor negative momentum exists in the simulated seasons. For each simulated season we compute $\hat{H}_p$. To ensure consistency, we replicate each simulation 10,000 times and use these results to build distributions for $\hat{H}_p$, under the null hypothesis. We then use these distributions for $\hat{H}_p$ to obtain p -values so that we can test whether each team's sequence of wins and losses are random, and thus independent of each other, or not. By fixing the winning percentage in these simulations we are imposing stationarity under the null hypothesis and this is a limitation of our method. That said, the limitation can be lessened by testing for momentum at multiple points throughout the season; this allows the winning percentage to vary incrementally. To alleviate this limitation completely, one could test for momentum after every game in the season.

To gain a better understanding of the entropy measure, and how it can be used in this context, we now examine several hypothetical examples. We first examine a season that is 10 games long and then a season that is 20 games long. In both sets of examples we assume the team we are analyzing wins approximately half of their games. We analyze different 10 and 20-game win/loss sequences by computing entropy after the 11th or 21st game. We ensure that the last game (either game 11 or game 21) is a win so that we capture the last inter-event time and the entropy measure accounts for everything in the season's sequence of games. Note that adding an additional game, that is a win, at the end of the season makes it so the team's

winning percentage is slightly higher than 50%, and this will also have a small impact on the p -values. Also note that, based on how the p -values are computed, the results and takeaways will hold true regardless of the value we select for the team's win percentage.

- (1) For the first set of examples, assume that the season is only ten games long. A team that experiences the least amount of momentum will alternate between wins and losses, whereas a team that exhibits the most momentum possible (both positive and negative) will experience wins and losses in exactly two clumps of five games. In these two cases $\hat{H}_p$ goes from 0.0198 to 0.1834 and the p -value goes from 0.9891 to 0.0570. Based on the property of convergence, as the number of wins and losses that are grouped together increases, the entropy measure increases and the p -value decreases, within the aforementioned ranges. Sequences with groups of two wins and losses have an entropy measure slightly larger than 0.0198 and a p -value that is slightly smaller than 0.9891. Similarly, sequences with groups of four wins and losses will have an entropy measure slightly smaller than 0.1834 and a p -value that is bigger than 0.0570, approximately equal to 0.1812. Thus, in a 10-game season with $\alpha = 0.10$ and assuming a 50% winning percentage, we only reject the null hypothesis that wins and losses are independent of each other in the extreme case in which the team experiences wins and losses in exactly two clumps. We use $\alpha = 0.10$ because it offers the most conservative, yet reasonable, value at which to assess the statistical significance of the entropy measure and therefore the existence of momentum. To examine more instances in which we reject the null we double the length of the season.

- (2) Figure 1 plots $\hat{H}_p$ and associated p -values for different groupings of wins and losses in a 20-game season. For example “6/6” denotes a grouping of six wins and a grouping of six losses, within a 20-game season of otherwise alternating win/loss sequences. From the plot we see that, as expected, $\hat{H}_p$ increases and the p -values decrease as the number of wins and losses that are grouped together increases. With $\alpha = 0.10$, we cannot reject the null hypothesis until there is approximately a grouping of seven wins and seven losses. Running through these hypothetical examples we also note that, in each of the groups of wins and losses, moving the groupings around in the sequence and adding small deviations to the alternating win/loss patterns (i.e., adding a couple of wins in a row or a couple of losses in a row) does not

Figure 1: Entropy measures and p -values for different win/loss groupings in a twenty-game season.

have a significant impact on $\hat{H}_p$ or the associated p -value. This realization demonstrates the property of continuity.

4 Analysis

To determine if NHL teams exhibit momentum we study each team's performance over the 2018–2019 season and utilize the methodology described above to test for momentum. Data for each team is collected from www.hockey-reference.com. Figures 2 and 3 provide incidence graphs for all the teams in the NHL, separated by conference and division. The plots show wins over each team's 82-game season and can be used to visually check for winning and losing streaks. However, incidence plots can be deceptive when looking for momentum in a team's season because teams with higher winning percentages will have plots that are more densely populated than those with lower winning percentages. Instead, to test for momentum we must turn to statistical analysis.

One of the primary benefits of using entropy to check for momentum in a team's season is that, unlike other measures, it does not require stationarity. In our context non-stationarity means that a team's winning percentage can change over time, without negatively influencing the measure. This is important because, as shown in Figure 4, teams' winning percentages typically do not stabilize until about midway through the season. We plot Tampa Bay's and Chicago's win percentages over time because Tampa Bay had the highest win percentage in the league and Chicago had one of the lowest, and because the fluctuations in their win percentages are representative of most any team's win percentage in the league. In the case of the St. Louis Blues, the winning percentage did not stabilize until well after half of the games had been played.

(a) Metropolitan Division

(b) Atlantic Division

Figure 2: Incidence graphs for the teams in the Eastern Conference.

To gain more insight into the entropy measure, $\hat{H}_p$, we analyze how the measure changed over the season for both Anaheim and St. Louis. We analyze the measure for Anaheim, because they end up having the largest entropy value (smallest p -value), and for St. Louis, because they end up having the least significant entropy value and because that is where our story started. In this analysis we use a rolling season length, and adjust the number of games, N . Figure 5 plots the entropy measure and each team's incidence graph over the length of the season.

The plots show how entropy changed over the course of the season and what the entropy would be if the season was to end at each specific moment along the x -axis. They also show how the entropy measure does not require

stationarity; the win percentage is constantly changing with every game as the season progresses, and the entropy measure accounts for these changes. However, to determine if the team is experiencing momentum we must stop the process and test for it via the method described above at fixed points in the season. Whenever we test for momentum, at that point we fix the win percentage so p -values can be computed. The plots also show how winning and losing streaks and groups of wins and losses impact the measure. More and longer streaks of wins or losses will cause the measure to increase, while frequent fluctuations between wins and losses will cause the measure to decrease.

Anaheim experienced several losing and winning streaks throughout the season. In the plot we can see that

(a) Central Division

(b) Pacific Division

Figure 3: Incidence graphs for the teams in the Western Conference.

(a) Tampa Bay and Chicago

(b) St. Louis

Figure 4: Win percentage over time.

they have an entropy of zero for games one through three. This is because their win percentage is 100% for the first three games and we expect them to win every game. Their

first significant losing streak occurred from games eight to 14. This was followed by a five-game winning streak, from games 26 to 30; a 12-game losing streak, from games 36 to

Figure 5: Entropy over time.

47; and a seven-game losing streak, from games 50 to 56. Note that the last two losing streaks were only separated by two wins; they lost 19 of 21 games from game 36 to 56. In the plot, Anaheim's measure spikes after the first seven-game losing streak. This losing streak is significant because, at this point, the season is only 14 games long and a losing streak this long is very unlikely, especially following the team winning five of their first seven games. In other words, at the 14-game mark, their winning percentage is 36% and Anaheim's wins and losses are occurring in clumps. The next 12-game losing streak, starting with game 36, was preceded by them winning 11 of their previous 13 games. These groups of wins and losses cause the measure to increase as well, but not as drastically, because the sample size N is larger. At the end of the season Anaheim does not experience any significant streaks of wins or losses and the measure decreases as a result.

St. Louis's wins and losses, on the other hand, are fairly evenly distributed across the season. Particularly during the first half of the season, St. Louis would hardly have been considered streaky; their longest string of consecutive wins or losses was only three games long until they began their lengthy winning streak at game 49. After this point, they fell back into their somewhat alternating pattern with just one three-game losing streak and one four-game winning streak over the remainder of the season. As shown in Figure 5, $\hat{H}_p$ spikes early when the team earns their first and second win, at games three and seven, and then stabilizes throughout the remainder of the season. Note that entropy cannot be computed until the first win occurs and this is why there is such a dramatic spike at game three. If the season concluded with game seven, St. Louis would have had five losses and two wins and the losses would have been grouped in a pair

and a triplet. The team, at this point, may have been considered to have been experiencing negative momentum. The remainder of the season, however, does not show a lot of evidence to support momentum. The 11-game winning streak does not have a huge impact on the measure, because it is only one grouping of wins within 60 trials and not preceded or followed by groups of wins or losses. Note that the eleven-game winning streak does cause the measure to rise slightly, and because $\hat{H}_p$ typically decreases with n , a small increase in the measure is not expected and more impactful than it may appear. To be certain which teams are experiencing momentum and which are not, we look at how $\hat{H}_p$ is distributed for each of the teams, based on each team's winning percentage over the season.

Distributions for $\hat{H}_p$ are generated by simulating a team's NHL season based on their overall, end-of-season winning percentage. Figure 6(a) shows that entropy ($\hat{H}_p$)

is strongly and negatively correlated with a team's win percentage. The measure decreases as win percentage increases. We also see this relationship when analyzing the distributions for the different teams. Figure 6(b) plots the distribution for $\hat{H}_p$ for three teams, from left to right: Tampa Bay, which had the highest winning percentage of all the teams (75%); Las Vegas, which had a winning percentage of 52%; and Ottawa, which had the lowest winning percentage of all the teams (35%). As the winning percentage decreases the distribution for $\hat{H}_p$ shifts to the right and becomes more spread out. In terms of the hypothesis test described above, this means teams with lower winning percentages will require larger critical values in order for the null hypothesis to be rejected.

To make equitable comparisons between the teams, we use the distributions for $\hat{H}_p$ to generate p -values for each of the team's seasons. Figure 7 shows the p -values associated with each team's measure of entropy. Figure 8 plots each

(a) Correlation between entropy and win percentage.

(b) Entropy distributions

Figure 6: Entropy vs. win percentage.

Figure 7: Entropy p -values.

team's entropy measure against its associated p -value and shows that p -values decrease as entropy increases. Deviation in the expected negative correlation between entropy and its p -value, shown in Figure 8, can be explained by the teams' different win percentages. Anaheim has the smallest p -value by far, and it is the only p -value that is significant at the 0.90 or 0.95 level of confidence. We have to be careful not to overstate this finding. If we treated each team's season as an independent trial, we would expect to reject about 1.55 (31×0.05) of the teams' seasons, at the 0.05 level of significance. We also see that St. Louis's season had the largest p -value of 0.8751. Most people believe St. Louis was experiencing momentum because, as described in the quote from Gilovich et al. (1985) at the beginning of this paper, and in the foundational literature on randomness, the 11-game winning streak was seemingly not representative of a team that was not experiencing momentum.

Many analysts, players, and sports fans believe that the St. Louis Blues gained momentum in the second half of the season and carried that momentum into the playoffs. If we break the St. Louis Blues season in half and analyze the first and second half in separate 41-game chunks we still cannot reject the null hypothesis and therefore do not find any evidence of momentum. In the first half of the season, the Blues had won 17 of 41 games, giving

Figure 8: Entropy versus p -values.

them a win percentage of 41.5%. At this point $\hat{H}_p = 0.0296$ which has an associated p -value of 0.9807. In the second half of the season, starting with game 42, the Blues won 28 of 41 games, giving them a win percentage of 68.3%. Analyzing the second half of the season separately results in $\hat{H}_p = 0.0346$ and an associated p -value of 0.5245. Thus, though there is more evidence of momentum in the second half of the Blues season than in the first half, there is not sufficient statistical evidence to reject the null and claim that their wins and losses were not randomly distributed.

Our findings are consistent with what has been found in the past; team momentum is rare. Owners, coaches,

Figure 9: Runs test p -values. Figure 10: Runs test p -values vs. entropy p -values.

players, and fans are convinced their team is experiencing momentum because they rarely view the season as a whole. Instead, they focus on just a small sample of the most recent games. The perception of momentum can be explained by recency bias, or the tendency to place a disproportionate amount of weight on recent events and less on events from the more distant past. Coaches, players, and analysts may study the past five games to improve their performance in the next five games. Or, they may analyze a series of games on the road, or the next series of home games, to improve the next series of road or home games. They rarely look at the season as a whole. And, because small random samples are not necessarily representative of their

populations, analyzing these small samples does not shed light on the team's performance over the entire season.

5 Discussion

To more definitively determine how our findings compare with what has been done in the past we compare our results with the results achieved using the Wald-Wolfowitz Runs Test (Bradley 1968). The Runs Test counts the number of runs that occur during the season and compares it to the number of runs that are expected to occur during the season, based on the success probability (*i.e.*, win percentage). A run is defined as an uninterrupted sequence of wins

or losses. When teams have significantly more or fewer than the expected number of runs, then we reject the null hypothesis that wins and losses are independent of each other. Figure 9 shows the p -values for each team's season, based on the Runs Test and Figure 10 shows that, overall, there is little correlation between p -values obtained using entropy and the p -values obtained using the Runs Test. In other words, the two measures produce much different results, though both conclude that very few teams exhibited momentum during the season.

Despite the four to five significant winning or losing streaks that Anaheim exhibited in their season, the Runs Test does not conclude that these wins and losses are dependent on each other. The Runs Test does not detect the sequences because there are enough single-game-win and single-game-loss runs to mask the significant long winning and losing streaks. This fact limits the effectiveness of the Runs Test in detecting momentum. Perhaps one of the biggest differences between the two approaches is evident when looking at Montreal. The p -value for Montreal for the Runs Test equals 0.0663, which is significant at the 0.90 level of confidence, whereas the p -value for Montreal using the entropy test is 0.8385, one of the largest in the group. Montreal never experienced a winning streak longer than four games or a losing streak longer than five games. Because the team routinely experienced single-game wins followed by single-game losses, their season displayed more runs than expected and this generated a small p -value for the Runs Test. Note that a hypothetical team that has a 50% winning percentage and alternates between winning and losing would generate a small Runs Test p -value, and we would reject the null hypothesis for this team as well because the number of runs is much more than expected. Stated another way, teams with either very clumpy or very non-clumpy data (lots of streaks or very few streaks) will have low p -values using the Runs Test, so in either case results for those teams suggest momentum. But because a high number of runs does not fit our definitions for positive or negative momentum, results from the Runs Test can be misleading. In the case of Montreal, the entropy measure indicates the team did not exhibit momentum over the course of the season because the p -value for the entropy measure is large since the wins and losses are evenly distributed across the season; the Runs Test suggests otherwise although simple observation of wins and losses depicts less "streakiness" than many other teams exhibited.

The Runs Test and the test using entropy described herein generate two different sets of results. Determining which test to use depends on what one is testing. Based

on our results, the entropy measure does a better job of detecting data that is grouped or clumped together and is not limited by a success rate (i.e., winning percentage) that is not constant or stationary. Additionally, the Runs Test counts winning or losing streaks that are broken by a single loss or win as separate runs while the entropy measure considers the overall clumpiness of the data. With momentum defined as the tendency of an effect to be followed by a similar effect, the most accurate way to detect momentum is by testing for events that are grouped or clumped together. Consequently, we believe that entropy is the better way of detecting momentum, either positive or negative, in NHL teams' seasons.

6 Conclusion

In this paper we have demonstrated how entropy can be used to test for momentum in a hockey team's season. Testing for momentum using entropy results in more reliable results than existing measures. That said, our overarching research question asked if NHL teams' wins and losses are influenced by momentum. Based on statistical tests using entropy and with $\alpha = 0.10$, we fail to reject that wins and losses are random for 30 of the 31 teams. Anaheim is the only team, in the 2018–2019 season, with a win/loss record that generated a p -value smaller than 0.10. Consequently, Anaheim is the only team for which we can reject the null hypothesis in favor of the alternate that wins and losses are not random. So momentum may exist, but it is much more rare than the typical fan may believe, especially when examining the entire 82-game season. Though, statistically speaking, the St. Louis Blues did not show any signs of momentum, they did look like a different team in the second half of the season. In the first forty-one games the Blues had a winning percentage of 41%. From game 42 to game 82, the Blues had a winning percentage of 68%. Something caused the team to start consistently winning games and though we cannot be sure what caused this change, we cannot conclude that it was momentum.

Areas of future work include finding a methodology that accounts for opponent dynamics and playing environments. We hypothesize that we would find even less evidence of momentum if we used a team's winning probability with respect to the specific opponent of each game played. If a winning streak coincides with a team playing against multiple inferior teams, this would detract from the belief that momentum is at play. We also suspect that changes in win probability due to playing away from or at home would further strengthen the findings that winning and losing streaks are more likely due to randomness or

other measurable factors than momentum. Again, if a winning streak coincides with a team playing multiple home games in a row, this would also detract from the belief that momentum is at play. Finally, if a team were interested in patterns other than multiple wins/losses in a row, some of the aforementioned statistical tools like runs tests may help identify if the perceived patterns are significant. Also, if teams were interested in investigating point differential patterns, other tools like Shewhart Control Charts (from the quality control body of knowledge) may help identify if perceived patterns in mean and variability are significant.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: None declared.

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

- Adams, R. M. 1995. "Momentum in the Performance of Professional Tournament Pocket Billiards Players." *International Journal of Sport Psychology* 26 (4): 580–7.
- Adler, P. 1981. *Momentum: A Theory of Social Action*. Beverly Hills: Sage Publications.
- Arkes, J. 2010. "Revisiting the Hot Hand Theory with Free Throw Data in a Multivariate Framework." *Journal of Quantitative Analysis in Sports* 6 (1). <https://doi.org/10.2202/1559-0410.1198>.
- Arkes, J., and J. Martinez. 2011. "Finally, Evidence for a Momentum Effect in the NBA." *Journal of Quantitative Analysis in Sports* 7 (3). <https://doi.org/10.2202/1559-0410.1304>.
- Attali, Y. 2013. "Perceived Hotness Affects Behavior of Basketball Players and Coaches." *Psychological Science* 24 (7): 1151–6.
- Avugos, S., J. Köppen, U. Czienskowski, M. Raab, and M. Bar-Eli. 2013. "The 'hot Hand' Reconsidered: A Meta-Analytic Approach." *Psychology of Sport and Exercise* 14 (1): 21–7.
- Ayton, P., and I. Fischer. 2004. "The Hot Hand Fallacy and the Gambler's Fallacy: Two Faces of Subjective Randomness?" *Memory & Cognition* 32 (8): 1369–78.
- Bar-Eli, M., S. Avugos, and M. Raab. 2006. "Twenty Years of 'Hot Hand' Research: Review and Critique." *Psychology of Sport and Exercise* 7 (6): 525–53.
- Bocskocsky, A., J. Ezekowitz, and C. Stein. *Heat Check: New Evidence on the Hot Hand in Basketball*. New York. Available at SSRN 2481494.
- Bradley, J. V. 1968. *Distribution-Free Statistical Tests*. Ch 12. Englewood Cliffs, NJ: Prentice-Hall.
- Cohen, B. 2020. *The Hot Hand: The Mystery and Science of Streaks*. New York: Custom House.
- Csapo, P., S. Avugos, M. Raab, and M. Bar-Eli. 2015. "How Should 'Hot' Players in Basketball Be Defended? The Use of Fast-And-Frugal Heuristics by Basketball Coaches and Players in Response to Streakiness." *Journal of Sports Sciences* 33 (15): 1580–8.
- Dorsey-Palmateer, R., and G. Smith. 2004. "Bowlers' Hot Hands." *The American Statistician* 58 (1): 38–45.
- Estes, W. 1964. "Probability Learning." In *Categories of Human Learning*, edited by A. W. Melton. New York: Academic Press.
- Gigerenzer, G., W. Hell, and H. Blank. 1988. "Presentation and Content: The Use of Base Rates as a Continuous Variable." *Journal of Experimental Psychology: Human Perception and Performance* 14 (3): 513.
- Gilden, D. L., and S. G. Wilson. 1995. "Streaks in Skilled Performance." *Psychonomic Bulletin & Review* 2 (2): 260–5.
- Gilovich, T., R. Vallone, and A. Tversky. 1985. "The Hot Hand in Basketball: On the Misperception of Random Sequences." *Cognitive Psychology* 17 (3): 295–314.
- Gray, R. M. 2011. *Entropy and Information Theory*. New York: Springer Science & Business Media.
- Greenberg, N. 2019. *Blues Streak: St. Louis is Red-Hot as NHL Playoffs Approach*. Washington, D.C.: The Washington Post.
- Hales, S. D. 1999. "An Epistemologist Looks at the Hot Hand in Sports." *Journal of the Philosophy of Sport* 26 (1): 79–87.
- Huitema, B. E., J. W. McKean, and J. Zhao. 1996. "The Runs Test for Autocorrelated Errors: Unacceptable Properties." *Journal of Educational and Behavioral Statistics* 21 (4): 390–404.
- Iso-Ahola, S. E., and C. O. Dotson. 2014. "Psychological Momentum: Why Success Breeds Success." *Review of General Psychology* 18 (1): 19–33.
- Kahneman, D., and A. Tversky. 1972. "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3 (3): 430–54.
- Key, J. A., and D. W. Ball. 2014. *Introductory Chemistry*, 1st Canadian ed. BCcampus Open Education.
- Koutras, M., S. Bersimis, and P. Maravelakis. 2007. "Statistical Process Control Using Shewhart Control Charts with Supplementary Runs Rules." *Methodology and Computing in Applied Probability* 9 (2): 207–24.
- Laplace, P. S. 1951. *A Philosophical Essay on Probabilities*. New York: Dover (Original work published in 1796).
- Miyoshi, H. 2000. "Is the 'Hot-hands' Phenomenon a Misperception of Random Events?" *Japanese Psychological Research* 42 (2): 128–33.
- Moskowitz, T., and L. J. Wertheim. 2012. *Scorecasting: The Hidden Influences behind How Sports Are Played and Games Are Won*. New York: Three Rivers Press (CA).
- Oskarsson, A. T., L. Van Boven, G. H. McClelland, and R. Hastie. 2009. "What's Next? Judging Sequences of Binary Events." *Psychological Bulletin* 135 (2): 262.
- Peel, L., and A. Clauset. 2015. "Predicting Sports Scoring Dynamics with Restoration and Anti-persistence." In *2015 IEEE International Conference on Data Mining*, 339–348. IEEE.
- Pinkert, C. 2019. *Blues Set New Franchise Record for Consecutive Wins*. St. Louis. NHL.com.
- Raab, M., B. Gula, and G. Gigerenzer. 2012. "The Hot Hand Exists in Volleyball and Is Used for Allocation Decisions." *Journal of Experimental Psychology: Applied* 18 (1): 81.
- Reifman, A. 2011. *Hot Hand: The Statistics behind Sports' Greatest Streaks*. Washington, D.C.: Potomac Books, Inc.
- Smith, G. 2003. "Horseshoe Pitchers' Hot Hands." *Psychonomic Bulletin & Review* 10 (3): 753–8.

- Stone, D. F. 2012. "Measurement Error and the Hot Hand." *The American Statistician* 66 (1): 61–6.
- Thomas, J. 2019. *A Great Eight: Blues Enjoy Longest Winning Streak in 16 Years*. St. Louis: St. Louis Post Dispatch.
- Tune, G. 1964. "Response Preferences: A Review of Some Relevant Literature." *Psychological Bulletin* 61 (4): 286–302.
- Tversky, A., and D. Kahneman. 1971. "Belief in the Law of Small Numbers." *Psychological Bulletin* 76 (2): 105.
- Tversky, A., and D. Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–31.
- Tversky, A., and T. Gilovich. 1989. "The Cold Facts about the "Hot Hand" in Basketball." *Chance* 2 (1): 16–21.
- USA Today. 2019. *Blues Beat Coyotes 4-0 to Stretch Winning Streak to 8 Games*, USA Today, <https://www.usatoday.com/story/sports/nhl/2019/02/14/blues-beat-coyotes-4-0-to-stretch-winning-streak-to-8-games/39062065/>.
- Vergin, R. C. 2000. "Winning Streaks in Sports and the Misperception of Momentum." *Journal of Sport Behavior* 23 (2): 181–98.
- Wardrop, R. L. 1999. "Statistical Tests for the Hot-Hand in Basketball in a Controlled Setting." *The American Statistician* 1: 1–20.
- Wolfowitz, J. 1943. "On the Theory of Runs with Some Applications to Quality Control." *The Annals of Mathematical Statistics* 14 (3): 280–8.
- Yaari, G., and S. Eisenmann. 2011. "The Hot (Invisible?) Hand: Can Time Sequence Patterns of Success/failure in Sports Be Modeled as Repeated Random Independent Trials?" *PloS One* 6 (10): e24532.
- Zhang, Y., E. T. Bradlow, and D. S. Small. 2013. "New Measures of Clumpiness for Incidence Data." *Journal of Applied Statistics* 40 (11): 2533–48.
- Zhang, Y., E. T. Bradlow, and D. S. Small. 2014. "Predicting Customer Value Using Clumpiness: From RFM to RFMC." *Marketing Science* 34 (2): 195–208.