SMU Data Science Review

Volume 1 | Number 2 Article 12

2018

Goalie Analytics: Statistical Evaluation of Context-Specific Goalie Performance Measures in the National Hockey League

Marc Naples
Southern Methodist University, mnaples@smu.edu

Logan Gage
Southern Methodist University, jlgage@smu.edu

Amy Nussbaum
amyenussbaum@gmail.com

Goalie Analytics: Statistical Evaluation of Context-Specific Goalie Performance Measures in the National Hockey League

Marc Naples, Logan Gage, Amy Nussbaum

Master of Science in Data Science
Southern Methodist University
6425 Boaz Lane, Dallas, TX 75205

Abstract. In this paper, we attempt to improve upon the classic formulation of save percentage in the NHL by controlling the context of the shots and use alternative measures other than save percentage. In particular, we find save percentage to be both a weakly repeatable skill and predictor of future performance, and we seek other goalie performance calculations that are more robust. To do so, we use three primary tests to test intra-season consistency, intra-season predictability, and inter-season consistency, and extend the analysis to disentangle team effects on goalie statistics. We find that there are multiple ways to improve upon classic save percentage, including controlling for shot type, measuring performance against an "expected goals" metric, and perhaps most importantly, calculating a save percentage that includes shot attempts that go wide. Despite these avenues for improvement, many questions remain due to the questionable robustness of all measures, and the clear presence of team effects.

1 Introduction

The application of high volume data and statistical theory has spread to every major sport in the past decades. Hockey has always been something of a laggard in this area, partially because of the difficulty in meaningfully reducing the plays of a free-flowing game such as hockey into discrete, objective classifications. Nonetheless, a new wisdom eventually emerged in scrutinizing every single shot attempt. Thus far, such approaches are proving useful for overall team stats as well as individual skaters, but have not advanced the ball in analyzing goalies with measures that are reliably repeatable or predictive. In this paper, we explore methods of controlling existing shot data to identify ways that measure goalie performance with more consistency and predictive power.

Hockey has never been a sport for stat-heads. The statistical tracking anyone thought to record on games in the National Hockey League ("NHL") grew at a glacial pace, beginning with goals. This led to the counting of assists (and goals plus assists, points). Coaches and writers tossed in a few more numbers such as penalty minutes, and got a little more clever in tracking the goal differentials of skaters while each individual was on the ice ("plus-minus"). Another piece of the puzzle was counting shots to track how

many shots each skater and team was attempting, while simultaneously calculating save percentages of the shots a goalie turned away.

While other American professional sports discovered the revolutionary usefulness of "big data" and powerful mathematics, only recently has hockey developed better metrics based on on-ice shot counting. By more carefully focusing on shot counts, both shots-on-goal and shot attempts that do not reach goal, analysts realized that teams that controlled the shot counts also controlled possession of the puck, and thus the game itself. While conventional wisdom insisted that all shots are not created equal, it turned out that it is actually very likely that scoring chances will predictably follow shot totals, and goals predictably follow scoring chances. In this way, hockey found a basic indicator like hits and batting average in baseball or yards in football. Additionally, with large piles of data, this shot counting can differentiate the performance of players on the ice even if they are not directly involved in scoring plays.

Ultimately, (most of) the doubts of the old-school thinking regarding shot data was overcome by the demonstration that shot-based performance was both more repeatable and more predictive of future goals and wins than past goals and wins themselves. This result has served as a powerful lodestar to guide many different metrics by which to judge individual skater and team performance. Unfortunately, to date, a similar lodestar to judge goalie performance has not been discovered.

In this paper, we explore several goalie performance metrics derived from detailed shot data. The traditional goalie measures are wins, goals against average, and save percentage. These measures are generally acknowledged to be insufficient to capture actual goalie performance, necessitating more nuanced analysis and statistics.

We perform three primary statistical analyses to explore if there are measures better than classic save percentage to accurately capture goalie performance. We control for the context of the shot, attempting to reduce noise and other factors that affect the probability of scoring on a shot that are beyond the control of the goalie himself. Testing for repeatability and predictability, we see there are a few promising approaches to isolate goalie skill.

We find that controlling for shot type, using an expected goals metric, and counting all shot attempts instead of the classic formulation of only shots on goal, generally improve the repeatability and predictive power of goalie statistics. These findings, however, are marginal improvements, and are not as robust as we would prefer. Furthermore, we observe clear "team effects" that need to be sorted out here and further in the future.

In this paper we begin with a background that generally describes the development of sports and hockey analytics. We continue with a related work section about public work that has been completed relative to the specific topic of goalie performance. We then describe the data and tests to be completed, before laying out our battery of test results. Finally, we discuss ethical issues around slow adoption of analytics in hockey, before tying all the findings together into a set of discrete conclusions.

2 Background

"Moneyball" is something of a buzz-word that has penetrated the popular lexicon in recent years. Driven by the aforementioned book by Michael Lewis (and a subsequent movie version starring Brad Pitt), "Moneyball" is often used as a general term or verb to describes the use of sophisticated data and statistics to comprehensively re-think the way teams evaluate players. Armed with new and superior methods for evaluating players, teams can exploit an informational advantage.

Baseball was the sport specifically covered in "Moneyball," and remains the most statistically advanced sport. Baseball has long tracked complicated boxscores to summarize every play, but modern analytics can be more accurately traced to the "Sabermetrics" movement in the early 80s. Now, professional baseball teams have staffs specifically dedicated to statistical analysis.

Every sport has eventually followed baseball into its own version of Moneyball-ing their game. Professional basketball teams seemed eager to create analytics departments, but perhaps hockey was the last sport to find the place for analytics. In recent years, however, "hockey analytics" has gained traction and perhaps a reluctant mainstream acceptance.

The watershed moment for hockey analytics is the use the of shot-based play-by-play data. Previously, few plays that occurred during a game were noted; basically, just goals and penalties. Otherwise, it was assumed that little could be tracked in a game where the puck never stops moving, and the players fluidly substitute for one another.

The problem with tracking only goals is that goals are relatively rare events. Furthermore, there are 12 players on the ice at all times (except special teams situations). As a result, most players would have a blank stat-sheet at the end of a game, and even those players that did have statistical entries might essentially be bystanders to the rare event of interest.

Eventually, someone realized how much of the game could be captured with detailed shot-data. Of course, the point of the game is to create scoring chances and ultimately goals, but measuring the quality of a scoring chance is a subjective exercise and the "old-school" stressed that not all shots are created equal. Shots on goal, on the other hand, are simple to track and occur frequently. Even though not all shots are equal, so long as a player or team is not systematically biasing their shot attempts (which can be shots-on-goal, all shot attempts [including those blocked or shot wide] commonly referred to as "Corsi", or unblocked shot attempts commonly referred to as "Fenwick"), shots usually project to scoring chances, and scoring chances usually project to actual goals.

Furthermore, shot data can be applied to individual players even if they aren't the player shooting the puck. With the fluid substitutions of hockey, differentials of shot attempts when any particular player is on the ice can be tracked, providing numerous data points and on-ice/off-ice shot differentials that indicate the impact an individual player has on the flow of play even if that player never shoots or scores.

The effectiveness of these shot measures is proven by two key, demonstrable results. The first is repeatability. Teams and players that perform well in these shots measures tend to consistently do so in the future. While actual goal-scoring tends to be streaky,

inconsistent, and prone to anomalies, shot generation and suppression is much more consistent.[1]

Second, shot measures are predictive. A team or player that is generating many shots will likely score goals in the future, regardless of whether or not they are scoring many goals at the present moment. On the other hand, simply seeing a team or player scoring goals at the present does not reliably project to continued goal scoring in the future.

These breakthroughs, however, pass completely over goalies. Little has been discovered that better captures goalie performance beyond the classic, flawed save percentage calculation.

3 Related Work

The present challenge facing hockey analysts is to develop methods that build upon the basic insights discussed above. Plain old shot counts are powerful, and can furthermore be evaluated by computers at the individual player level to create all kinds of ratios and differentials. At some point, however, you end up with an ever-growing pile of differentials all built upon raw, "dumb" shot events. The next natural step is to perform second-level analysis upon the raw shot data. To this end, we cannot know what is happening behind closed doors in the analytics departments of individual teams, or at companies that provide propriety data and tracking to customers, but there is an active community doing public work to this end.

One prominent route of second level analysis of hockey data is "expected goals" measures. Here, there are at least four prominent public "expected goals" measures.[2] [3] [4] [5] The primary function of these measures is to create a logistic regression model that considers several facts about each shot. Upon consideration of shot location, type of shot, and any information that can be gathered or calculated about events preceding the shot, a probability of a goal being scored can be calculated. This information provides a different, and perhaps superior perspective than simple shot counts, and can also be plugged into any detailed tracking of shots. These models help inform our analysis of goalie performance, and simply judging goalies against such "expected goals" has been found to be an improvement.[6]

---

[1] Travis Yost, "How Analytics forecast future success and failure," https://www.tsn.ca/how-analytics-forecast-future-success-and-failure-1.355108, (September 3, 2015).

[2] Dtmaboutheart, "Expected Goals are a better predictor of future scoring than Corsi, Goals," https://hockey-graphs.com/2015/10/01/expected-goals-are-a-better-predictor-of-future-scoring-than-corsi-goals/, (October 1, 2015).

[3] Peter Tanner, "Shot Prediction Expected Goals Model," http://moneypuck.com/about.htm.

[4] Cole Anderson, "xG by last event type, zone & time since event," https://twitter.com/CrowdScoutSprts/status/866007034038280193, (May 20, 2017).

[5] Emmanuel Perry, "Shot Quality and Expected Goals: Part 1," http://www.corsica.hockey/blog/2016/03/03/shot-quality-and-expected-goals-part-i/, (March 3, 2016).

[6] Cole Anderson, "Expected Goals (xG), Uncertainty, and Bayesian Goalies," http://www.crowdscoutsports.com/game-theory/expected-goal-xg-model/, (June 17, 2017).

Another avenue that is actively being pursued by the analytics community are unified stats. Inspired by the now well-known baseball "WAR" statistic (Wins-Above-Replacement), several analysts have worked on hockey equivalents or near equivalents. A number of such models have been published,[7] [8] [9] as well as a "Game Score" model which measures contributions made by individuals to a single game.[10] This paper and research is too premature and thus is not specifically targeted towards creating a goaltender "WAR" figure, but any insights herein would be applicable towards a future unified goalie statistic.

As mentioned above, progress in finding the core of repeatable goaltender performance has not been as easy. Several analysts have poked and groped around the topic, taking approaches such as regressing save percentage by danger zone,[11] focusing on "quality starts",[12] or building complex, entirely new statistics.[13] We initially approach the problem guided by the value of simplicity and being agnostic as to what relatively simple and controlled save statistics may prove to be both repeatable and predictive of future success. This comparative analysis reveals statistics that are worthy of more focus to analysts in isolating goalie performance amidst a fog of factors that cause goals that are actually beyond the goalie's control.

4 Data and Methods

To perform our analysis, our primary source is data collected and made publicly available from moneypuck.com.[14] This database is bulk shot/event data. Specifically, our analysis uses the data for seven NHL seasons (from 2010 to 2016), containing 726,969 shot events, with 134 variables for each event. Most of these variables are raw data from the NHL, including many Boolean values such as rush shot, rebound shot, etc... Other variables in the data set have previously been cleaned and adjusted by the data provider to correct for known specific team and arena biases in recording events. Furthermore, other variables in the data set are calculated by the data provider and

[7] Acthomasca, "The Road to WAR Series," http://blog.war-on-ice.com/index.html%3Fp=429.html, (April 8, 2015).

[8] Emmanuel Perry, "The Art of WAR," http://www.corsica.hockey/blog/2017/05/20/the-art-of-war/, (May 20, 2017).

[9] Cole Anderson, "The Path to WAR*," http://www.crowdscoutsports.com/game-theory/the-path-to-war/, (June 28, 2016).

[10] Dom Luszczyszyn, "Measuring Single Game Productivity: An Introduction to Game Score," https://hockey-graphs.com/2016/07/13/measuring-single-game-productivity-an-introduction-to-game-score/, (July 13, 2016).

[11] FooledByGrittiness, "Regressing Sv% by Danger Zone," http://fooledbygrittiness.blogspot.com/2016/05/regressing-sv-by-danger-zone_31.html, (May 31, 2016).

[12] DragLikePull, "Are Quality Starts A Repeatable Skill?" http://nhlnumbers.com/2016/10/27/are-quality-starts-a-repeatable-skill, (January 11, 2018).

[13] Micah Black McCurdy, "Standardized Goals Against," http://hockeyviz.com/txt/sGA, (May 31, 2017).

[14] http://moneypuck.com/about.htm

bundled into this public data set based on chains of events, such as time since last event, change in angle since last shot, or if the shot is or produces a rebound.

Furthermore, we supplement this primary data with more NHL data that is publicly available on the internet for free from sources such as https://www.hockey-reference.com/ and http://www.corsica.hockey/. These sites function by aggregating official data from the NHL, as well as scraping official NHL play-by-play shot data, then compiling and calculating the events to create a repository of data for teams and players, respectively.

Lastly, although our data set covers seven seasons from 2010-11 from 2016-17, we elected not to include the 2012-2013 in our analysis. This is because that season was shortened by a lockout, which shortened the season by four months and reduced the number of per-team games played from the usual 82 to 48 games. This dramatic reduction in season length introduced additional unpredictability and uniqueness to the numbers from that season, and furthermore would have required that we lower our threshold of minimum shot attempts faced by any goalie to be included in our sample for that season. All things considered, we decided to exclude this season entirely.

We also chose to test regular season games only, although the data set included playoff games. The playoffs in the NHL have a reputation for being a different type of hockey than the regular season, and it is not unusual for both goalies and skaters to record performance statistics in the playoffs that are markedly different from the regular season. Including playoff games would skew the statistics of the (minority) of goalies that competed in the playoffs, which skewness would be exacerbated even more severely upon those goalies who played in a large number playoff games.

4.1 Tests Performed

Armed with this data, our analysis is built upon three types of tests to apply to the data. The three tests are designed to evaluate the quality and robustness of the given individual goalie performance measure or sub-measure. The tests are; 1) intra-season consistency/repeatability, 2) inter-season consistency/repeatability, and 3) intra-season predictive power.

Beginning with intra-season repeatability, our method is to break down every shot by season. Once grouped by season, each season group is randomly divided into two groups; group A or group B. Then we calculate the save percentage (or other measure, as applicable) for each individual goalie independently for events in both group A and group B in each season. After eliminating individual save percentages based upon fewer than 500 shot events (in either group A or group B), the save percentages are paired such that we can compare, for instance, the save percentage of "Goalie X" in group A and group B of in any season. Our final takeaway from this data preparation then is to calculate the correlation and its p-value of those paired save percentage values.

This test is meant to capture how consistently goalies maintain their save percentage over the course of a single season. If a goalie's save percentage was perfectly consistent, the correlation of these randomly-split shot event values would be a 1.0.

Regarding inter-season repeatability, similar analysis is completed. The difference here is that rather than creating split-pairs of save percentage by random sampling within a season, data from entire seasons is utilized. Again, individual goalies' save percentage is paired, such that, for instance, the save percentage of "Goalie X" in 2010-

11 (“2010”) is compared his save percentage in 2011-12 (“2011”). Lastly, a correlation test is performed on a combined data set of four paired seasons (2010/2011, 2013/14, 2014/15, and 2015/16). A variation of the preceding intra-season repeatability test, this correlation will measure how consistently a goalie maintains his save percentage across seasons.

Our third and final statistical test is testing intra-season predictability. At first glance this appears to be a similar measure to intra-season repeatability, but this measure splits data based on time, and utilizes cross-subset save percentage analysis.

To perform this test, we again divide the data by season. Once grouped by season, each season’s data is split by game ID number. Our data set does not contain date information specifically, but it does contain the unique game ID that the NHL assigns to every game that proceeds sequentially through all 1,230 NHL games played each year. Once separated, this test finished like the others—pairing an individual goalie’s performance in each split in every year.

We add one additional wrinkle to this test, however. For intra-season predictability we further test whether an alternative save percentage measurement in the first half of the season can predict classic save percentage in the second half of the season. The reason for this is that, despite flaws in classic save percentage that is the very impetus for this study, classic save percentage is a bottom-line result. If a measure can be identified that predicts such a final, broadly applicable outcome, that might be the best indicator of value of all. If found, this would be just like how “Corsi” for skaters predicts future goals better than current goals predicts future goals.

4.2 Test Application on Shot Context

With the test protocol defined, we first apply each test to the classic measurement of save percentage—the percentage of all shots on goal that are faced and saved by an individual goalie. This classic, raw, calculation then establishes a baseline for each of our three tests.

We then compare this baseline of classic save percentage to several variations and modifications to save percentage. Each test is thus repeated several times, to be performed on each save percentage variation we consider and explain below.

The first type of data subset we work with is data based on general game situation. To accomplish this, we look at shots that occurred in 5 skaters versus 5 skaters situations, as well as 4 skaters versus 5 skaters situations (commonly referred to as “shorthanded”, when the goalie’s team is forced to play with one fewer skater for a brief period because of a penalty).

Five-on-five (“5v5”) play would presumably be particularly important, as it is the normal way NHL hockey is played. It is the most common situation for a team and goalie to be in, and places both teams with equal opportunity to score. For these reasons, much of advanced hockey analytics is performed on 5v5 data only.

Shorthanded situations present a very different challenge for the goalie. In these situations, the goalie is under intense pressure as the opposing team generally takes their time holding the puck in the offensive zone. They control the puck and may selectively pick a shot rather than taking whatever they can muster at 5v5. On the one hand, this creates a situation where the goalie appears particularly helpless. On the other

hand, it is old adage that "a goalie is a team's best penalty killer", implying this is where a truly skilled goalie shines and separates himself.

Aside from general game situation, we also consider a data set of all shot attempts, even shots that do not hit the net. This is to account for the possibility that some goalies' save percentage may be padded by habitually "saving" shots that were going wide, or that other goalies may in fact be causing shooters to miss the net entirely. It is often said (anecdotally) that a hot goalie "gets in the shooters' heads," and makes them think they have to shoot a perfect shot into the top corner of the net. If true, a good goalie might leave a fingerprint of forcing more shots wide without ever actually touching the shot himself.

Our third philosophical approach to sub-setting the data is to adjust for shot quality. Furthermore, we adjust for shot quality in two separate ways; by controlling for a few shot conditions in the data set, and by utilizing an "expected goal" measure.

Of the two methods, controlling for shot conditions is simpler. To do so, we propose a method of "Clean" shot attempts. This set of shot attempts exclude shots that are classified as tip-ins or deflections, and also excludes shots that are rebound opportunities. The goal is to limit shots to those in which the goalie has a clean opportunity to stop a shot, rather than scrambling on a shot that makes a last-second change of direction or is a second-effort after the goalie has stopped the first attempt.

This control, however, we still fail to capture passing plays that precede a shot that can dramatically increase the difficulty of stopping a shot, nor does it capture shots that deflect off a defensive player. This is simply a short-coming of the data set used, which is not tracked by the NHL.

The other method, "expected goals", is briefly explained in the Background section of this paper. In short, this data set includes an expected goal value for each shot event per the moneypuck.com expected goals model, also cited in the Background. We can apply this expected goal value to every shot subset and split we create by adding up the total expected goals and subtracting them from the total actual goals. We further divide that difference by shot attempts to compensate for the different number of shots each goalie faces. While the specific calculations of expected goal for each is a "black-box", it represents a comprehensive measure of shot quality that may prove to reveal a more repeatable or predictive goalie skill measure.

In sum, we have one baseline set of shot data and four subsets of data to evaluate repeatability and predictability thereof against the baseline, all situation classic save percentage. Thus we have these five groups:

Table 1. Shot-control groups for analysis.

Group	**Description**
Classic Save Percentage – Baseline	Percentage of shots on net saved – all situations
5v5	Percentage of shots on net saved – 5v5 situations only
4v5 or shorthanded	Percentage of shots on net saved – 4v5 situations only
Clean	Percentage of shots on net saved in all situations, but excluding tip-ins, deflections, and rebound attempts

All Attempts	Percentage of shots that do not result in goal, including shots that miss the net

Furthermore, each of these five subsets can be re-evaluated to calculate performance against expected goals rather than save percentage. This creates up to ten measures upon which to apply each of the three tests we designed, including the baseline comparison value.

5 Results and Analysis

5.1 Intra-season Repeatability

The first and most basic result of our analysis is the intra-season repeatability of classic save percentage. Indeed, this entire paper is premised on the idea that this value will be relatively low.

We find class or raw save percentage to have an intra-season correlation of 0.1458, with a p-value of 0.0425 over six seasons. This therefore represents a statistically significant result, albeit this is not a very strong trend. Figure 1 below visualizes this finding.

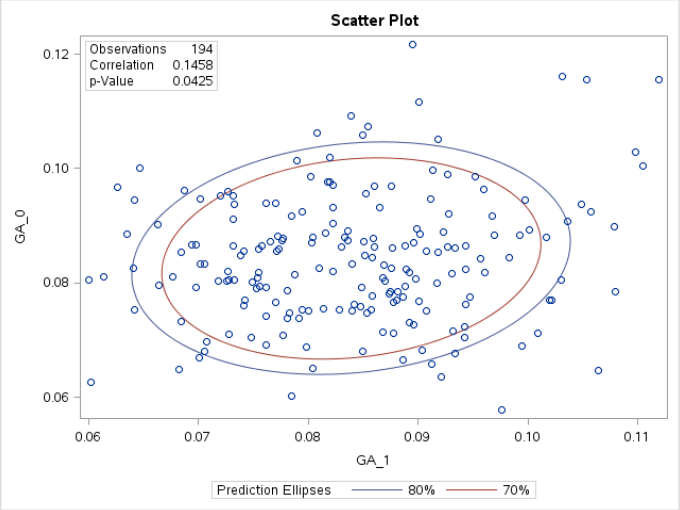

Fig. 1. Scatter plot of intra-season repeatability save percentage pairs. Each data point represents an individual goalie's performance in a given season, randomly split, presuming a minimum of 500 shots faced in random split group.

We would also note that this correlation can vary wildly from season to season. Some seasons this value is as high as 0.37, while other seasons it's 0.

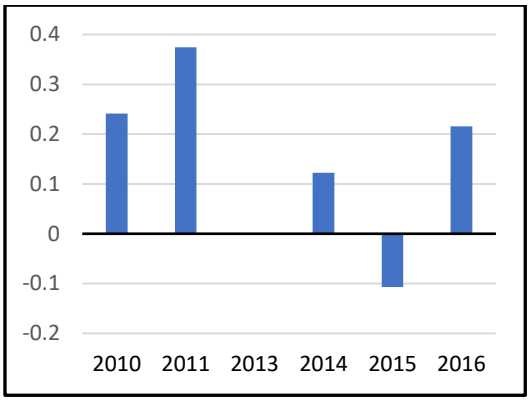

Fig. 2. Intra-season correlation (repeatability) of classic save percentage, by season

This large variance from season to season reveals a lack of robustness in raw save percentage as a statistic measuring a repeatable skill. Overall, we observe that this statistic gives a baseline measure that is weakly significant, and not robust.

From this baseline, we re-run our testing process on our four alternative types of save percentage: 5v5, 4v5, Clean shots, and All Attempts.

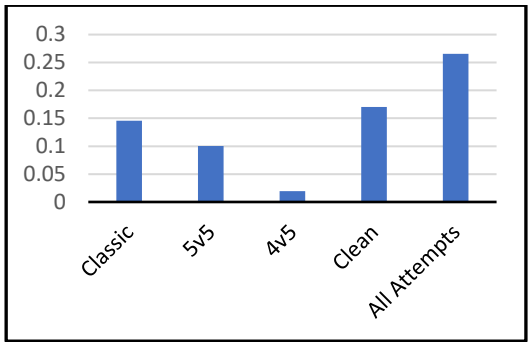

Fig. 3. Intra-season correlation of save percentage, by group, aggregate of 6 seasons.

Compared to the baseline classic save percentage, we see a modest decrease to correlation using only 5v5 shots on goal, and a dramatic decrease using only 4v5 shots on goal. Conversely, we observe a modest increase in correlation restricting the sample to only Clean shot attempts, and a large increase simply using All Attempts.

The first conclusion we draw from these results is that 4v5, or shorthanded, save percentage is not at all reliable. Part of this is simply due to a smaller sample size. Goalies simply do not see the same total of shots in this situation. They play relatively few minutes shorthanded, and shooters are picky about waiting for the right shot. Second, this implies that the adage about a goalie being your best penalty killer is misguided. Goalies simply do not perform consistently in this situation over the course of a season. No doubt a hot goalie will go a long way towards a successful penalty kill performance in the short term, but not even a good goalie cannot do this time after time.

The minor decrease in correlation when restricting the sample to 5v5 shots is another data point that simply controlling for broad game situation does not improve the power of save percentage. If one were evaluating goalie performance in a season, it appears to be a better idea to look at 5v5 performance rather than 4v5 performance, but perhaps they shouldn't restrict for either unless there is a compelling independent reason to do so.

More encouraging is the slight improvement to correlation of Clean shots on goal. We would certainly expect this to be the case based on our subjective understanding of how a goalie performs and limiting the factors out of his control, so this result is a bit of a sanity check. That said, the improvement is only modest.

Indeed, the real improvement comes in using All Attempts, not limiting it to shots on goal as classic save percentage does. This simple change nearly doubles intra-season correlation from the base of 0.1458 all the way up to 0.2656. Clearly there is something going on here, although there are a few possible explanations for this.

One possible explanation for this increased correlation is that this increases the sample size. Opposite of the problem with 4v5 shots on goal, many shots go wide, so the number of shot attempts faced will be significantly higher than shots on goal. This increased shot count can eliminate some of the random noise in the results and drive up correlation.

Another explanation for the increased correlation found with all shot attempts is that forcing shooters to shoot wide appears to be a repeatable skill. Perhaps it is even a more repeatable skill than actually saving the shots that are on goal. The interesting question presented here is, is that a repeatable skill of the goalie, or of the entire team? We return to this question later.

We can also take another look at the three top groups, raw save percentage, Clean save percentage, and all attempts save percentage, but instead of calculating save percentage we measure performance against expected goals.

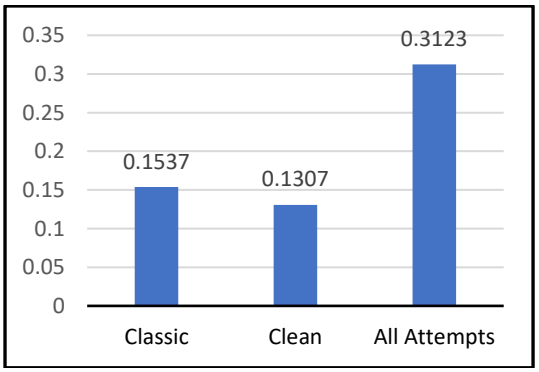

Fig. 4. Intra-season correlation of expected goals performance, by group, aggregate of 6 seasons

Recall that raw save percentage had a correlation of 0.1458, so calculating expected goals increases this value by a very small amount. Conversely, using performance against expected goals reduces the intra-season correlation of Clean shots from 0.1705 to 0.1307. This result is mildly surprising, although it may be due to the fact that both

Clean shots and expected goals attempt to control for shot quality, and thus using the measures together may be redundant and/or counterproductive.

Finally, using expected goals on all shot attempts further raises the correlation from 0.2656 to 0.3123. In sum, we see preventing gross shot attempts from becoming goals appears to be a repeatable skill as measured by either save percentage or expected goals. The question of whether this a phenomenon of goalie talent or of team performance, however, remains.

5.2 Intra-Season Predictability

Switching over to intra-season predictability correlation, we see generally similar results, but with less variance between data groups.

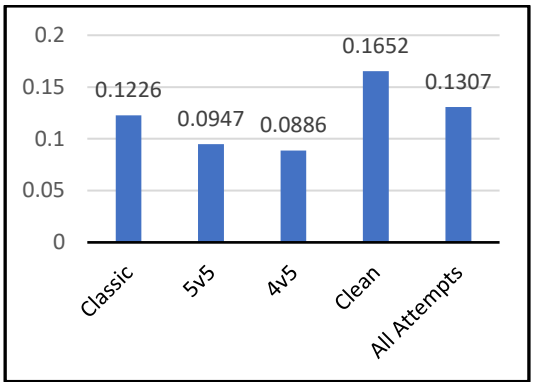

Fig. 5. Intra-season predictability (correlation of first half group percentage to second half classic save percentage), by group, 6 year aggregate

Recall that this test differs from intra-season repeatability correlation, which is a random division of shot events. This measures different types of save percentage in the first half of a given season, and measures how well it matches classic save percentage in the second half of the season.

Here we find a baseline correlation value of 0.1226 between raw first half save percentage, and second half raw save percentage, with a p-value of 0.0568. This is nearly the same as the intra-season repeatability correlation of 0.1458, although the p-value moves just enough to cross onto the wrong side of the 0.05 significance barrier.

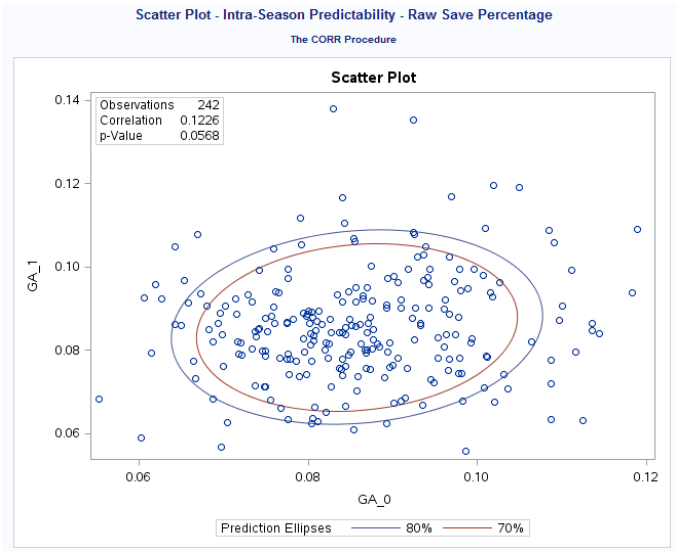

Fig. 6. Scatter plot of intra-season predictability save percentage pairs. Each data point represents an individual goalie's performance in a given season, randomly split, presuming a minimum of 500 shots faced in random split group.

Building off a largely unchanged baseline intra-season correlation from repeatability to predictability, we do see a noteworthy change in contrasting 5v5 and 4v5 save percentage. In testing for predictability, these two subgroups now have very similar correlations of 0.0947 and 0.0886, respectively. This is a dramatic change from repeatability, where the values were 0.1004 and 0.0193 respectively. While 5v5 performance is about the same across the tests, 4v5 is much better for predictability than for repeatability.

Another change we see from repeatability is that All Attempts save percentage is dragged almost entirely back to the correlation of only shots on goal. This is an interesting test result that, while save percentage on All Attempts stays relatively consistent over the course of the season, having a high All Attempts save percentage in the first half does not necessarily cross-over to predict classic save percentage of shots on goal in the second half much better than first half classic save percentage.

The value that remains largely unchained across repeatability and predictability is Clean shots, with a correlation here of 0.1652. We interpret this a piece of evidence that restricting for Clean shots as we have designed does improve upon general save percentage by clarifying actual goalie performance by looking at shots upon which the goalie has a more fair chance to make a save.

Furthermore, we again performed this intra-season predictability test using expected goals instead of save percentages. In this case, measuring performance against expected goals does not prove to be a predictor of raw save percentage later in the season. No group of shot types proves to be statistically significant, although expected goals on Clean shots comes the closest with a correlation of 0.123 and a p-value of 0.0623. This

adds more evidence to the conclusion that that the best measure for predicting performance later in the season is current performance on Clean shots on goal.

5.3 Inter-Season Repeatability

Lastly, we performed our gamut of tests to measure inter-season consistency of all our shot groups. These results diverge significantly from the intra-season tests and are shown below.

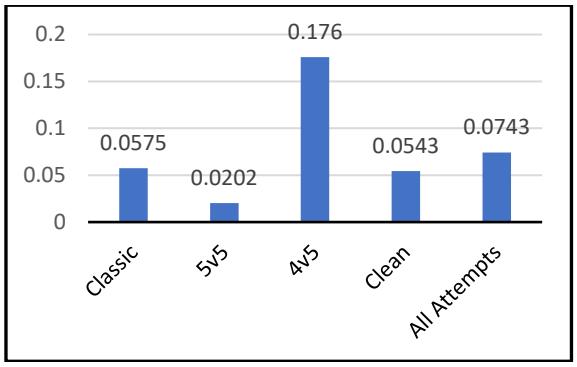

Fig. 7. Inter-season correlation of save percentage, by group, 6 season aggregate

Immediately we see a dramatic reduction in the correlations for nearly every shot group. The baseline value, classic save percentage, sees its intra-season correlation of 0.1458 plummet to 0.0575. Additionally, none of these correlations prove to be statistically significant at the standard 0.05 level.

Throughout the intra-season testing, Clean shots and all shot attempts performed the best. In this test, Clean intra-season correlation is reduced by approximately 75%, while all attempts' intra-season correlation is reduced by about 72%. The only solace for the power of these measures is that All Attempts remains a more reliable indicator than raw save percentage.

Sticking out like a sore thumb in this set of results is 4v5 inter-season repeatability. While 4v5 had a nearly zero correlation of intra-season repeatability, it has a large inter-season repeatability correlation of 0.176. This has all the indicators of an outlier and anomalous result, so we regard this result with caution. It is also difficult to reconcile how this statistical grouping can simultaneously be worthless on an intra-season basis, but standout on an inter-season basis. Perhaps this is another example of the small sample size of 4v5 shot totals wreaking havoc and yielding untrustworthy indicators.

Alternatively, we repeat the test measuring performance against expected goals rather than save percentage for inter-season consistency for key groupings. In so doing, we see a bit of a rebound in correlations.

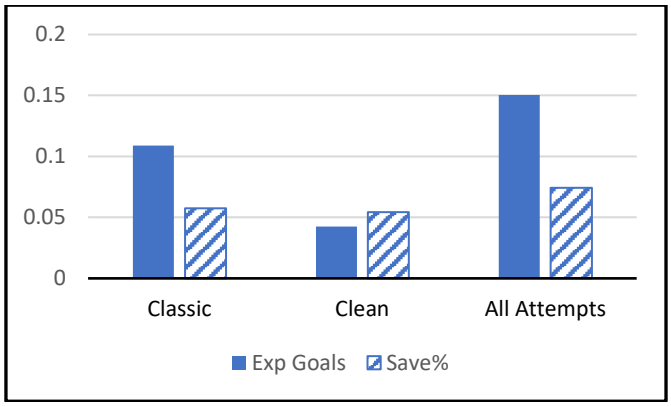

Fig. 8. Inter-season repeatability, by group, save percentage versus expected goals

As shown in the chart, using expected goals instead of save percentage roughly doubles the correlation of classic shots on goal and All Attempts save percentage. This is encouraging, but only All Attempts produces a statistically significant p-value, and the correlation of 0.1503 is hardly powerful.

Overall we are left with surprisingly low measurements of inter-season consistency across the board. If we were beginning to identify some grouping of shots across which goalies perform more predictably and consistently within a single season, the measures do not prove to be robust from one season to the next.

5.4 Extended Analysis Regarding Team Effects

Having completed our primary testing methods on the main data set, a few takeaways emerge. The most prominent takeaway is that looking at all shot attempts proves to be a more repeatable and more predictive statistical measure than classic save percentage, including when we are to predict future classic save percentage. Despite this finding, not even the All Attempts save percentage is reliable from season to season. This is a curious and frustrating result, and in response, we extend the analysis to look at team factors hinted at earlier in the paper.

The first step in this extended analysis is to re-run our test for inter-season repeatability, but instead of testing the repeatability of individual goalies, we look at repeatability of team save percentage. The results are dramatic.

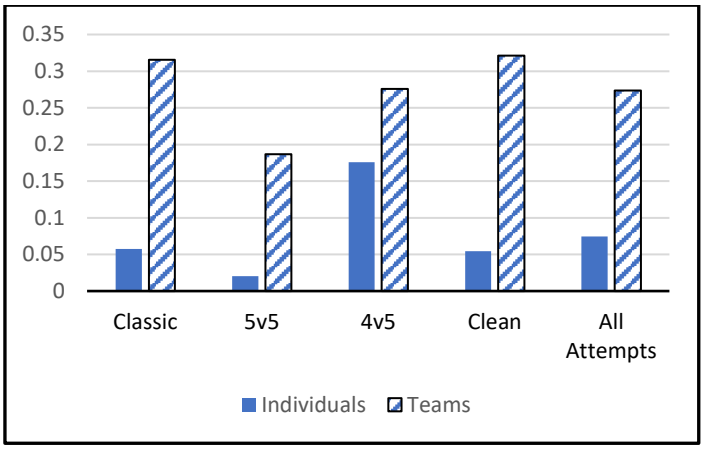

Fig. 9. Inter-season classic save percentage repeatability, 6 season aggregate, by group, individual versus team

The chart above shows that all flavors of team save percentage are much more consistent across season than when you grade out individual goalies. This strongly implies that save percentage is a team effort.

Additionally, all groups are comparable, apart from 5v5. This is significant, because while using All Attempts often differentiates individual goalies, it does not stand out when measuring team save percentage. While the results imply save percentage is a team measure, it also implies that individual goalies are affecting missed shot attempts they never touch. Even if we cannot sort out how much of save percentage is on the team and how much is on the individual, All Attempts appears to be an improved metric for judging individuals.

To further inquire into the team versus individual inquiry, we also created another subset of only goalies that changed teams within the seasons of our primary data set. This creates a relatively small sample set of 30 goalies, but in it we can directly compare which is a better predictor of save percentage in the year after switching teams—the individual's save percentage the prior year while playing for another team, or the team's save percentage the prior year with an entirely different goalie.

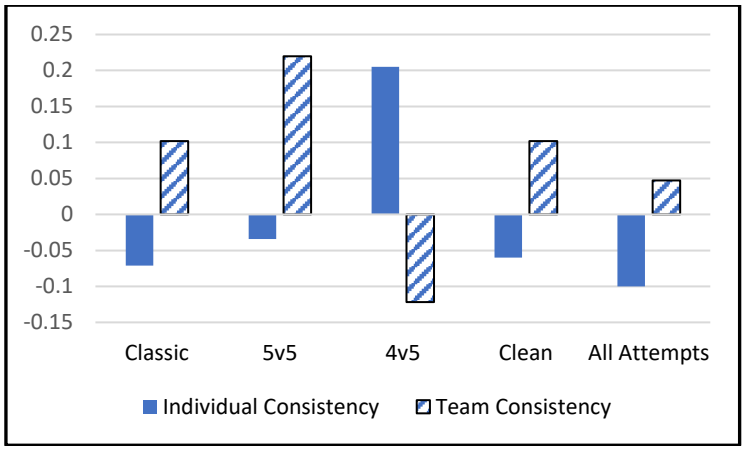

Fig. 10. Inter-season consistency of goalies that change teams, 6 year aggregate, by shot group. Showing individual consistency across teams, versus team consistency whereby the goalie inherits his new team's old save percentage

The chart above shows the results. For individual goalies that changed teams, their save percentages for their old team was a very poor predictor of their save percentage with their new team. The absolute values of the correlations is low, often going negative. In fact, the relatively higher striped bars show that the better way to predict a newly-arrived goalie's save percentage is to look at the team's save percentage the previous season with different personnel. This reinforces the notion that save percentage is a team statistic.

Another way to illustrate this point is to look at examples of teams who have an entrenched starting goalie while the number two goalie changes from year to year.

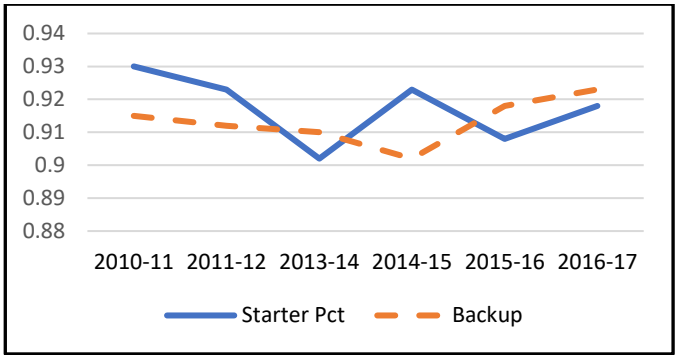

Fig. 11. Season to season save percentage of Nashville Predators starting goalie, Pekka Rinne, versus a changing cast of 3 backups.

One such case is shown above. Here, the Nashville Predators kept a high-paid starter, Pekka Rinne, for the six years pictured (the 2012-13 lockout-shortened season is excluded), while they went through three different backup goalies. In this instance,

Rinne was not all that consistent, while the revolving door of backups were slightly more consistent.

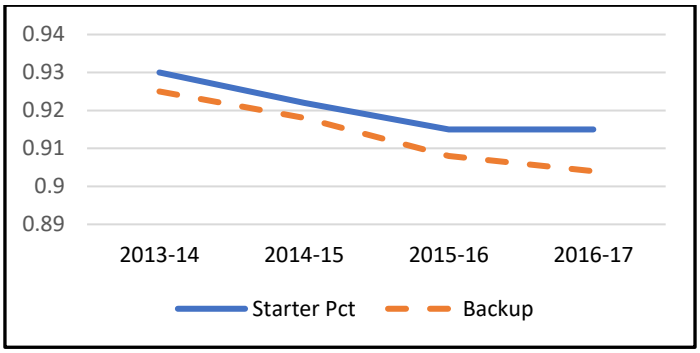

Fig. 12. Season to season save percentage of Boston Bruins starting goalie, Tuuka Rask, versus a backup goalie that changed each of the 4 seasons charted.

We can also consider the Boston Bruins. The Bruins, in fact, changed the backup goalie every single year. In this case, the high paid starter, Tuuka Rask, outperformed the cheap backup every season, but only by a few thousandths of a percent. Regardless of who the Bruins used as a backup goalie in any season, they could anticipate his save percentage would mirror the starter's save percentage.

This presents fairly strong evidence of save percentage as a team measure. Namely, teams' save percentage holds much more constant year-over-year than individuals, and goalies that change teams are more likely to inherit their new teams' save percentage than carry over they save percentage they saw with their older team. Lastly, even if a team has a carousel of backup goalies, their performance as a group is likely to be steady.

5.5 Goalie Hot Zone

In light of the team effects demonstrated in the previous session, we explored diving deeper into breaking down individual goalie's performance against different shot types. Looking outside of limited hockey analytics to practices in other sports, one of the best resources for a baseball pitcher is a hitter's hot zone. This zone is a construct of several hitting statistics that result in a checkerboard pattern image with different colored sections illustrating the hitter's performance. These hot zones are tailored to a specific hitter and are invaluable to pitching against them. Performing a stepwise logistic regression method on our data produced probability curves that reveal individual goalie measures analogous to baseball hitter hot zones.

Upon testing, our stepwise regression boiled down to two major predictor variables: shotDistance (measured in feet) and shotType (consisting of seven different types: BACK, DEFL, SLAP, SNAP, TIP, WRAP and WRIST). Each shot type was individually used against shot distance in the regression models. This created seven different probability curves for the individual goalie. These probability curves, when

put together, give a visual comparison for the performance of a goalie against the different shot types.

The results against goalie Henrik Lundqvist are illustrated in Figure 13. Over the course of the seasons in our dataset he encountered thousands of shots on goal. This chart shows Lundqvists's performance against all classifications of shots and corresponding distances. He performs the best against BACK type shots and the worst again DEFL type shots. We expect the WRAP shot type probability curve does not conform to the others due to only 122 events over the 10,442 shots on goal Lundqvist encountered. In comparison, Figure 14 shows how Kari Lehtonen performed. Lehtonen performed the worst against SLAP type shots and the grouping of shot types where closers together compared to Lundqvist.

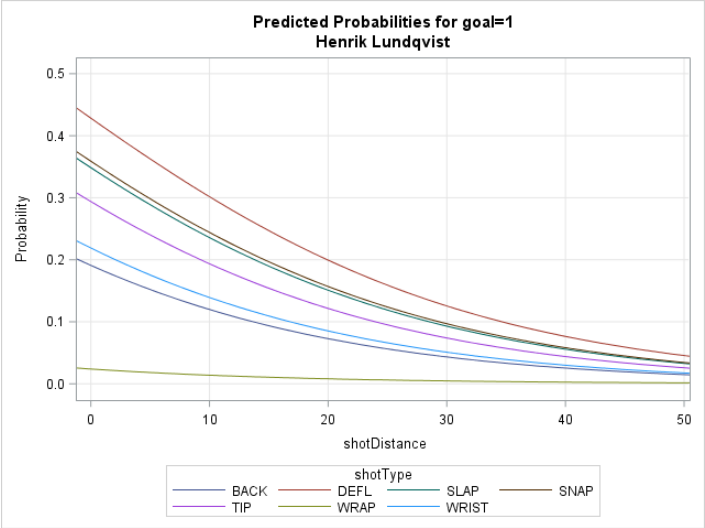

Fig. 13. Predicted Probabilities for Henrik Lundqvist for all shots on goal grouped by type of shot.

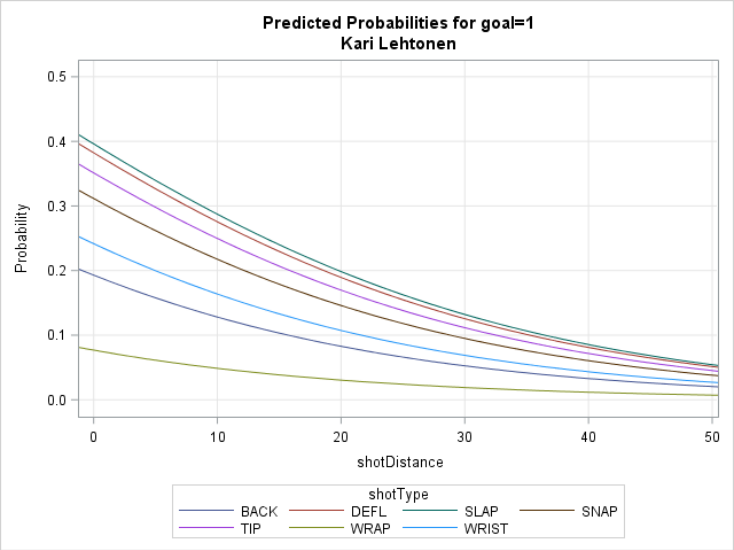

Fig. 14. Predicted Probabilities for Kari Lehtonen for all shots on goal grouped by type of shot.

With these probabilities, we can construct a hot zone type of image for the goalie. Figure 15 shows a representation of the successful shots on goal and differentiated by shot type. Each section represents a 5% chance of making a goal. Here a deflection shot has a greater range than a backhand shot at the same probability level.

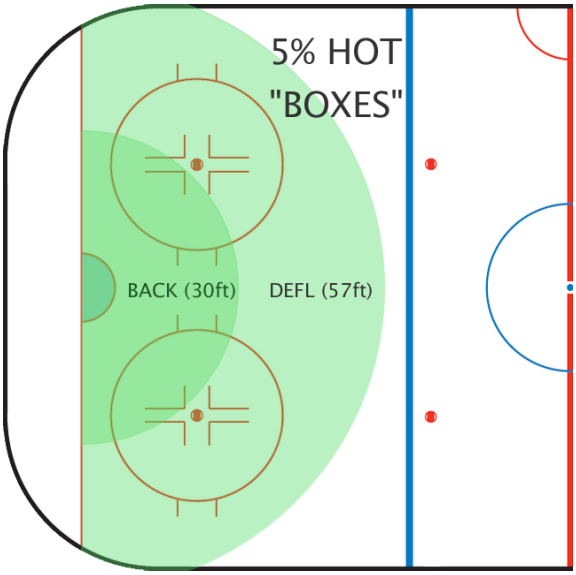

Fig. 15. Hot Zone for Henrik Lundqvist showing backhand and deflection type shots only.

With this hot zone data an opposing team could plan their strategy around what type of shots the goalie has the most difficult time defending and knowing at what range those shots are the most effective. For Henrik Lundqvist, opposing players should attempt more deflecting type shots and not attempt backhand shots when they are at greater range. This process could expose some goalies' weaknesses and give an advantage to other teams, but this also would provide the goalie and his team with information to improve their skills and maximize the synergy between team save percentage and individual save percentage. Lastly, intelligently comparing the respective five percent curves for different goalies side-by-side may peel back another layer of the onion that is team effect on save percentage.

6 Ethics

In completing this study, there are few ethical concerns in the data itself. All statistical events are public spectacles viewed by thousands, and intensely tracked and released to the public. In using the data, however, it raises question about the interplay between data and people.

This paper has already discussed the resistance and slow adoption of the hockey old guard to in-depth statistics. One reason for this may be that NHL teams are overwhelming run by former players, who are men who generally do not have higher degrees. For instance, a recent study found that 74% of NHL general managers are former professional players. Contrast that to baseball, where only 3% of general managers are former professional players, and 100% of them have post-secondary or higher degrees.[15] This simple fact may go a long way towards the league collectively casting a cold eye towards conclusions based on spreadsheets.

This gulf between the stat-heads and team management trickles down to the players as well. If the boss of the team is a former star player who publicly questions the value of "fancy stats," it is unlikely the players will bother to learn the logic of the number crunchers. This creates the current situation where you have a minority of analytically inclined people creating a body of new wisdom about understanding and quantifying the sport of hockey, trying to apply that knowledge, but the players themselves don't understand the first thing about it.

As an example, one of the authors was at hockey analytics panel when the topic of players fluency with these stats arose. A panelist then related a story where a former NHL player, currently working as a color commentator on an NHL broadcast, was asked about advanced stats. He said he "didn't buy into any of that Corsi stuff," but admitted shot count differentials have some value. The punchline here is that Corsi IS shot count differentials. This is a testament to the presumably common attitude of skepticism towards analytics from hockey management and players.

This dynamic creates an untenable environment that may leave important analysis misunderstood or ignored, while the players, whose livelihood is playing the game, cannot even understand the conversation. Of course, it is not necessary for players to understand p-values or programming code, but stats people can't just talk to themselves.

[15] Jason Paul, "Who's Running the Show?" https://www.waveintel.org/single-post/2018/02/06/Whos-Running-the-Show, (February 6, 2018).

Players and coaches have to be on board and understand the conclusions of the stats, and respond to them.

Advanced statistical analysis is coming to hockey teams. In a business that is all about the bottom line of dollars and wins, it is practically malpractice for a team to be ignorant of statistics and leave a powerful tool in the shed, unused. Going forward, it makes much more sense for analysts to play an active role in team management, which role should include give and take with the coaches and players such that players understand the what and why of the analysis. Perhaps it is just a matter of time until some NHL finds this structure.

7 Conclusions

Our study of the shot data over the course of six seasons does not yield a simple yes or no answer, yet it does reveal several patterns to advance understanding of classic save percentage in the NHL. Overall, we confirm that classic save percentage is a weak calculation, identify alternative ways to calculate save percentage that make it more a reliable indicator of performance, and pursue wrinkles in the data regarding team effects of all data based on saves.

Beginning with the original assumption of classic save percentage being an insufficient calculation, our tests largely confirm this. As measured by intra-season consistency, classic save percentage proves to be weakly statistically significant (p-value 0.0425), with a mediocre magnitude of correlation (0.1458). Furthermore, it is not a reliable, robust indicator. Some season some goalies will show high consistency of save percentage, whereas other seasons they will demonstrate almost no consistency and we cannot identify why.

Alternatively, we see a few ways to increase the reliability of save statistics. Controlling shots per our Clean shots formulation notably increases intra-season random correlation by about 15% (from 0.1458 to 0.1705). Even better, calculating "save percentage" by counting all shot attempts, whether on net or not, improves that correlation all the way up to 0.2656, and gets the p-value all the way down to 0.0001. While there is still plenty of space to improve upon this measure in terms of robustness and randomness, this is a strong finding.

Additionally, we can improve affairs even more calculating performance against expected goals instead of a save percentage. The intra-season repeatability of All Attempts performance against expected goals (per the provided moneypuck.com model) is 0.3123, more than double the respective value for classic save percentage. Although expected goals is a black box, subjective algorithm, we know there are gains to be had from this exercise.

Regarding our second test, intra-season predictability from first half performance to second half performance, we find only minor improvement by using All Attempts instead classic save percentage. The best group to use here is our Clean shots classification. Again, the power and robustness of these results is not dramatic, but again we find improvement from classic save percentage by switching to all attempts or Clean shot restrictions.

Lastly, we find uniformly bad inter-season consistency when judging individuals. Most goalies simply do not maintain consistent save percentage from season to season, although using expected goals on all shot attempts will get us to statistical significance

(0.0319) and a modest correlation (0.1503). More interesting is contrasting these values against how consistent team save percentage is across seasons.

Team save percentage inter-season correlation is wildly higher than individual goalies. Consistency of performance against expected goals on a team level is also significantly better than at an individual level, although less sharply. From this we must conclude that save percentage is very much a team statistic. This also may relate back to the individual consistency of using all shots instead of shots on goal, in that both the team and the individual goalie combine to perform a repeatable skill in forcing shots wide. This conclusion is buttressed by our supplemental finding that for goalies that switch teams in the offseason, their save percentage with their new team is closer to that team's save percentage from the previous season with a different goalie than it is to the goalie's own save percentage last year.

The Hot Zone analysis further reveals distinct separation between the goalies as it pertains to how they perform against shot types and shot distance. These predictive factors give the ability to determine how a goalie responds to different shot types and provides analysts more information about goalie performance. Teams may alter their strategy to deliver or prevent specific types of shots with greater frequency, while also presenting additional information that may explain parts of a goalie's performance within the umbrella of team save percentage performance.

While we believe these conclusions are important clues to improving upon classic NHL save percentage, we stress that none of formulations are truly robust, or isolate goalie performance away from team effects. Any way we look at the shot data, the results remain messy in terms of consistency or predictability. If better statistics of goalie performance are to be built in the future, there is still much to learn.

New data is probably needed to improve on quantifying goaltending performance. Specifically, what happens before any shots are taken need to be tracked. There are currently attempts to track "royal-road passes" that cross the center of the ice and require the goalie to move and re-position on the fly before stopping a shot.[16] Even better is full real-time tracking of puck and player position, which the NHL used at the World Cup of Hockey 2016 and may soon be fully implemented for all NHL games.[17] Other useful data points would be information of where shots are beating the goalie (above the glove, through the legs, etc), and more information about visual screens and deflected shots. Our hot zone analysis is a step in this direction.

Ultimately, the defeatist expression "goalies are voodoo" persists. We inch towards answers, but new questions arrive that may or may not be solved with more data. Maybe a bit of mystery will always remain around goalies, the cloistered individual apart from his 18 teammate skaters.

It's a profession where the difference between leading the league and being a fringe player is four shots in a hundred (.940 save percentage versus .900). Humans make mistakes, and they will appear in random patterns. And while there approximately 650 skaters in the NHL at any given moment, there are only 62 goalies. That is simply a

[16] Kevin Woodley, "Unmasked: Analytics provide new evaluation tools," https://www.nhl.com/news/unmasked-analytics-provide-new-evaluation-tools/c-744483, (December 18, 2014).

[17] Dan Rosen, "NHL pursuing revolutionary player-tracking systems," https://www.nhl.com/news/camera-player-tracking-2020-hockey-season/c-293815836, (December 8, 2017).

small group of athletes, and the distribution of their performance may not normalize. Here and now, we see the improvement in statistical quality by using All Attempts or Clean shots, utilizing expected goal models, understanding how much the team factors into an individual goalie's save percentage, and beginning to visualize specific aspects of goalie performance with hot zones.