LATENT VARIABLE MODELING FOR INTANGIBLE CONSTRUCTS IN SPORTS: THE CASE OF TEAM DEPTH IN THE NATIONAL HOCKEY LEAGUE *

Dylan Wiwad
Independent Data Scientist
Hockey Decoded
Chicago, IL, 60640
dwiwad@gmail.com

ABSTRACT

Sports analysts and fans frequently invoke intangible team qualities like "depth," "momentum," or "chemistry" to explain why teams win or lose, yet these constructs have infrequently been rigorously quantified. In this paper I demonstrate how latent variable modeling can formalize these intangible constructs in professional sports. Focusing on team depth in the National Hockey League, I construct the Total Depth Index using fifteen seasons of play-by-play (n = 6,049,797) and shift data (n = 15,021,146). I operationalize depth as the complement of the Gini coefficient applied to within-team distributions of shots-on-goal, expected goals, Corsi-For, and time-on-ice. Structural equation modeling confirms depth as a unidimensional latent construct with acceptable fit and measurement invariance across seasons and teams. While single-game depth fluctuates nearly randomly from one matchup to the next, aggregating across multi-game windows reveals a stable pattern. Teams with sustained depth advantages over recent games are significantly more likely to win their next matchup, outperforming a baseline prediction model in leave-one-season-out cross-validation. Critically, depth predicts winning not through shot volume but by creating higher-quality scoring chances. When controlling for shot quality, additional shot quantity becomes negatively associated with winning—a suppression effect demonstrating that depth's value lies in distributing dangerous opportunities across lines rather than simply flooding the net with low-percentage scoring attempts. This work moves depth from rhetorical shorthand to validated measurement and illustrates how psychometric methods can formalize other long-invoked but poorly measured constructs in sports analytics.

Keywords latent variable modeling · sports analytics · psychometrics · NHL · team depth · structural equation modeling · Gini coefficient · measurement invariance

1 Introduction

Trying to understand why one team wins and another loses is a perennial focus of sports analytics and punditry. When trying to explain or understand matchup outcomes commentators and fans alike often invoke intangible aspects of a game, constructs like "depth," "momentum," or "chemistry" that go beyond simple behavior like shooting, scoring, or passing (Noel et al., 2024; Taylor & Demick, 1994). Researchers have also demonstrated that such constructs can indeed influence matchup outcomes (e.g., Qiu et al., 2024). Intangible constructs, however, are often extremely difficult to define and measure (El-Den et al., 2020). This presents an interesting and worthwhile puzzle: how can we define, measure, and utilize unobservable constructs to further our understanding of team dynamics and the outcomes of athletic matchups? In this paper, I demonstrate how we can leverage the psychometric approach of latent variable modeling to define and measure intangible constructs in sports and help answer these questions.

* *Citation*: **Authors. Title. Pages.... DOI:000000/11111.**

The simple idea of measuring unobservable constructs is not a new idea in sports analytics. However, past explorations have tended to focus on single measurement indicators that serve as a proxy for the latent construct they are purported to measure. For example, Qiu et al. (2024) defined momentum conceptually as "positive or negative changes in cognition, physiology, emotions, and behavior caused by sudden or a series of continuous events" and determined that momentum in basketball can be measured as "achieving a net score difference of +6 points within 96s[econds]." Others have built econometric models quantifying team chemistry as a statistical enhancement of a player's individual performance due to the simultaneous on-court performance of their teammates, formally modeled as a significant positive within-team peer-effect (Horrace et al., 2020). While sophisticated, these models often rely on single indicators and assumptions that the underlying indicators capture the complete multi-faceted concepts they are trying to measure. However, the reality of constructs like momentum, chemistry, or depth is that they are multifaceted team capabilities that cannot be accurately reduced to a solitary indicator, no matter how complex its derivation.

Alternatively, we can consider these intangible constructs to be latent variables in a statistical sense. While such variables cannot be observed directly, they can be inferred by modeling the relationships between a series of other, conceptually similar but more observable, measurement indicators. This allows researchers to test whether the underlying indicators do indeed measure a single unidimensional (or multidimensional) latent construct with greater statistical certainty.

The technique of latent variable modeling is prominent in fields such as computational biology, computer science, and the social sciences more broadly (Blei, 2014). In particular, latent variable modeling is utilized heavily in the field of psychometrics, a subfield of psychology concerned specifically with the problem of measuring unobservable aspects of human psychology such as intelligence (Spearman, 1904), personality (Tupes & Christal, 1992), or psychopathology (Caspi et al., 2014). For example, in the study of personality, Tupes and Christal (1992) analyzed trait ratings and found that a consistent set of five latent factors (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect) captured much of the reliable variance in how people describe themselves and others (i.e., "personality"). Rather than treating each trait as independent, they showed that personality descriptors cluster systematically, reflecting deeper, underlying dimensions of human behavior. While psychometrics is predominantly concerned with the measurement of psychological traits, the underlying philosophy can be applied to any such unobservable variable.

Therefore, the approach of latent variable modeling gives us a meaningful structure through which we can create robust models to reliably measure, and validate our measurement of, unobservable dimensions of athletic competitions. In the present exploration, I have chosen to focus on the concept of "depth" as it pertains to hockey in the National Hockey League (NHL).

1.1 Depth in the National Hockey League

Depth in the NHL is often conceptualized as the extent to which a team's performance or production (e.g., goals, points, or shots on goal) are distributed across its players rather than concentrated within a few individuals. A team with high depth can maintain competitive performance even when star players are unavailable, lines are reshuffled, or game contexts shift because contributions to success are not driven entirely by the top lines or defensive pairings. In contrast, a team with low depth relies disproportionately on a small subset of players whose absence or underperformance can sharply diminish overall team performance. In the context of the NHL, this distributional quality manifests across observable dimensions such as ice time, shot generation, and point production. Each of these represents a measurable indicator of the latent construct of depth; that is, how evenly a team's capabilities are spread throughout its roster.

I have chosen to focus this exploration on depth in the NHL for three reasons. First, depth is a construct that hockey commentators, coaches, analysts, and fans invoke frequently. For example, Condor (2024) refers to depth with common phrases like "rolling four lines," saying "the Kraken's emphasis on depth scoring and rolling four lines consistently in games shows in league-wide rankings in which superstars like Colorado's Nathan MacKinnon and Mikko Rantanen and Edmonton's Connor McDavid and Leon Draisaitl log monster minutes." Yet despite its ubiquity in discourse, depth has rarely been defined or measured systematically. Traditional analytics tend to operationalize depth with behavioral proxies such as the average ice time of a team's "bottom six" forwards (Vollman, 2016), the number of players on a team with high versus low Game Score Value Added (GSVA, a measure of player productivity; Luszczyszyn, 2019, 2023), or the correlation between salary share and GSVA (Benson et al., 2024; Kerdman, 2016). While these indicators may hint at aspects of depth, they overlook its multifaceted nature, conflating player utilization, opportunity, and efficiency without distinguishing their conceptual contributions and, more importantly, systematically testing whether these metrics reliably measure a single latent factor.

The second reason I chose to focus on depth within the NHL is simply the availability and granularity of the necessary data. The NHL has, since the 2010-2011 season, tracked and made available detailed shift and play level information for every player in every game via its public API (See unofficial documentation at `https://github.com/Zmalski/NHL-API-Reference`). This comprehensive play-by-play data, coupled with shift charts, allows for the precise

calculation of time on ice (TOI) for every skater, as well as every single shot taken by every player. Without this granular data, it would be impossible to determine the true distribution of workload across the roster, which is a foundational component of the depth construct. Furthermore, the hockey analytics community has established robust secondary data sources, such as MoneyPuck's expected goals (xG) metric, which provide high-quality, standardized inputs for shot quality (MoneyPuck.Com - About and How It Works, n.d.). By leveraging this combination of official NHL shift-level data and robust community metrics, I was able to construct a game-level dataset spanning fifteen seasons, ensuring the consistency, reliability, and volume of data necessary to perform rigorous latent variable modeling and measurement invariance testing.

Finally, I focus on depth because it has real-world matchup and team performance implications. Teams with greater depth can likely absorb performance shocks (e.g., injuries or opponent mismatches) with less decline in effectiveness, distribute physical and cognitive load more evenly across players, and pursue more diverse offensive and defensive strategies. These properties mirror ideas from organizational behavior, where performance robustness emerges from redundancy and resource dispersion rather than centralized dependence (Argote & Ingram, 2000; March, 1991). By approaching depth through the lens of latent variable modeling, I treat it not as a collection of surface statistics but as a coherent, underlying dimension of team quality. My aim is to identify a set of observable indicators (i.e., the distribution of shots, expected goals, and time on ice) that collectively reflect an unobservable latent construct: team depth. Modeling these indicators together allows for the estimation of shared variance attributable to depth itself, distinct from the noise and idiosyncrasies of each individual measure. In doing so, I move analytics and punditry from rhetorical description of depth to its formal construct and measurement validation, providing a framework that allows depth, and other latent constructs, to be quantified, validated, and utilized to predict and understand outcomes such as win probability, consistency, and post-season success.

Additionally, while I define and validate depth at the team-game level, its performance implications are inherently comparative. Hockey outcomes are determined through head-to-head matchups, and thus the relevant empirical question is not simply whether a team is deep in absolute terms, but whether it is deeper than its opponent in a given contest. Accordingly, later analyses move from team-level descriptions of Depth to matchup-level tests that examine relative depth advantage between competing teams. This transition preserves Depth as the focal construct while aligning its use with the competitive structure of the sport.

2 Methods

2.1 Data Sources

In order to test and validate a measurement model capturing team depth, I created a game-level dataset encompassing all NHL regular and post-season playoff games (n = 19,197) from 2010 to 2025. I chose the 2010-2011 season as the cut-off date because this was the first season that the NHL tracked and hosted shift duration, a component of depth as I am conceptualizing it, in their API. Additionally, I opted to remove pre-season and international games because these games often have spotty or non-existent data tracking. For example, the play-by-play endpoint for the September 23, 2023 pre-season game between the St. Louis Blues and Arizona Coyotes contains only penalties and goals, a pattern typical of early historical NHL play-by-play or overall pre-season data (`https://api-web.nhle.com/v1/gamecenter/2023010003/play-by-play`). In order to build this dataset I utilized data from two primary sources. The first data source was raw data from the NHL API (`https://github.com/Zmalski/NHL-API-Reference`), from which I pulled all game play-by-play (n = 6,049,797) and shift data (n = 15,021,146) across all 15 seasons. The second data source was shot-level expected goals (xG) calculated and consolidated by MoneyPuck (n = 1,632,927; `https://moneypuck.com/data.htm`).

For each game, the NHL API play-by-play endpoint contains logs of many different types of events including game starts and stoppages, shots, and penalties as well as the player(s) involved in each event. Across all fifteen seasons, I pulled every shot-on-goal, goal, missed shot, or blocked shot (n = 2,210,691) that did not occur in a shootout. I chose to remove shootouts because they are not counted in NHL official shot counts and occur in a special context that does not reflect head-to-head on-ice competition. MoneyPuck's xG data contains, for every shot taken in a regular or post-season game, the predicted probability that that shot will be a goal. This xG metric is the output of a gradient boosted machine learning model that takes into account fifteen factors (e.g., distance from net, shot angle, shot type, etc). For more on xG and its calculation, see `https://moneypuck.com/data.htm`.

2.2 Data Processing and Cleaning

In order to calculate each component of depth, I began by creating a dataset containing every team roster by game. Importantly, these metrics only apply to skaters, and as such I removed all goalies from any further analysis and

calculation. For each skater in each game I then utilized my created play-by-play and shift datasets, as well as the MoneyPuck shot-level xG data, to calculate their total shots-on-goal (i.e., sum of shots-on-goal and goals as indicated in the NHL API), expected goals (i.e., sum of expected goals), corsi-for (i.e., sum of shots-on-goal, goals, missed shots, and blocked shots as indicated in the NHL API), assists, and time-on-ice (i.e., sum of all shift time in seconds). For each team within each game (team-game), I then calculated the Gini coefficient (See Construct Definition and Metric Operationalization) on the player-level totals of shots-on-goal, expected goals, corsi-for, assists, and time-on-ice. I then took the complement of each distribution so that the resulting coefficient reflects depth (i.e., equality) rather than inequality.

Given that the main focus of this investigation is the theoretical conception and construct validation for depth as it pertains to NHL hockey, I computed depth metrics across all game situations (even-strength, power play, penalty kill, overtime) without stratification. These contextual distinctions may be explored in future work.

2.3 Construct Definition

2.3.1 Gini as the Root of Depth

The core concept behind each component of the Total Depth Index (TDI) that I will outline below is the Gini coefficient (Dorfman, 1979). The Gini coefficient is a statistical measure of dispersion typically used to capture inequality in economic resources. For example, imagine a group of 10 people who collectively have $100,000. If one person has $100,000 and the other 9 have $0 that group will have a Gini coefficient of 1 representing perfect inequality. If each of the 10 people have $10,000 that group will have a Gini coefficient of 0 representing perfect equality.

While this metric is often used in economic resource distributions, it can be applied to any frequency distribution. For example, previous researchers have used the Gini coefficient in similar contexts to explore the effects of NBA salary inequality (Halevy et al., 2011) and organizational team status (Christie & Barling, 2010). Imagine the Edmonton Oilers play an NHL game with 18 rostered skaters, and take 36 shots over the course of that game. We can again see, theoretically, perfect inequality with 1 skater taking all 36 shots (Gini = 1) or perfect equality with each of the 18 skaters taking 2 shots (Gini = 0). In this paper I am using a standard definition and calculation of the Gini coefficient:

$$\text{Gini}(x) = \frac{\sum_{i=1}^{n}(2i - n - 1)a_i}{n\sum_{j=1}^{n} a_j} + \varepsilon \tag{1}$$

Where $x = (a_1, a_2, \ldots, a_n)$ is the vector of non-negative values for a metric of interest (e.g., shots-on-goal, xG, etc.) in ascending order, n is the number of observations (i.e., number of skaters on a team's roster for a given game), and ε is a small constant (i.e., 10^{-9}) to avoid division by zero (Guest, 2016/2025).

Given that Gini as defined above captures inequality in a distribution we can consider its complement to be a measure of depth (i.e., equality). Therefore:

$$\text{Depth}(x) = 1 - \text{Gini}(x) \tag{2}$$

Depth can thus be understood as the degree to which a given action (e.g., shots, xG, time on ice) is distributed across the rostered skaters. A value approaching 1 indicates perfect balance (every player contributing equally), whereas a value approaching 0 indicates extreme concentration among a few players.

2.3.2 Shot Depth Example

Between 2010 and 2025, the game with the most unevenly distributed shots (i.e., highest gini coefficient for shots on goal) was played by the New York Islanders in 2014 against the Washington Capitals. In this game, the New York Islanders rostered a full 18 skaters and took 11 shots. These 11 shots came from only six out of the eighteen rostered skaters, resulting in a shot-on-goal gini (SOG_G) coefficient of 0.783. Conversely, the game with the most evenly distributed shots was played by the Columbus Blue Jackets in 2016 against the Colorado Avalanche. In this game, the Columbus Blue Jackets also rostered a full 18 skaters and took 37 shots-on-goal. These 37 shots came from 17 out of the 18 rostered skaters, resulting in a SOG_G of 0.158. Thus, taking depth as the complement of these shot inequalities (i.e., $1 - SOG_G$), the New York Islanders had a shot depth (SOG_D) of 0.217, and the Columbus Blue Jackets had a SOG_D of 0.842, respectively, reflecting very shallow and very deep shooting distributions. Notably, SOG_D is not problematically correlated with volume metrics like total shot counts or total expected goals (See Construct Validation).

2.4 Metric Operationalization

Using this established depth-as-dispersion framework, one could consider various elements of a hockey game as important for quantifying depth. These elements could be offensive (e.g., such as shot or goal production), defensive (e.g., hits or blocked shots), or agnostic to position (e.g., ice time). In selecting the components of the TDI, I focused on measures that (a) reflect meaningful individual contributions to team performance across the entire roster, (b) are recorded consistently and reliably across all games and seasons, and (c) are broadly accessible via the NHL API.

I mainly focus on offensive indicators that capture puck possession and scoring opportunity creation, outcomes that matter heavily in determining the outcome of a game. I also included the depth of ice time because it captures overall depth in a team's roster contribution; did a team rely heavily on one line with a lot of ice time, or were they more evenly distributed across all four lines? Thus, in my initial conceptualization of depth as a latent construct I include the dispersion of: total shots on goal (i.e., shots on goal and goals), corsi-for (i.e., shots on goal, goals, missed shots, and blocked shots), total xG, assists, and ice time. I calculated each of the below metrics for every team, in every regular and post-season playoff game, across all 15 seasons ($n = 38,394$ team-games).

2.4.1 Shots on Goal

A shot on goal, as defined by the NHL, is any shot that "goes into the net, or would have gone in the net had the goaltender not stopped it" (Official Site of the National Hockey League | NHL.Com, n.d.). Practically, these are recorded in the NHL API play-by-play endpoint (`https://github.com/Zmalski/NHL-API-Reference`) as an event with type "shot-on-goal" ($n = 1,064,720$) or "goal" ($n = 113,708$). I calculated team-game shot depth ($M = 0.54$, $SD = 0.08$, $Min = 0.22$, $Max = 0.84$) according to the above definition of Depth.

2.4.2 Corsi-For (CF)

CF is an advanced metric that is not tracked as raw data by the NHL. As such, I calculated CF according to its standard definition of $CF =$ shots on goal $+$ goals $+$ missed shots $+$ blocked shots. While shots on goal capture scoring chances, CF is widely considered to be a proxy for offensive zone pressure and possession (Advanced Hockey Statistics, n.d.). Thus, CF theoretically captures how much time a team spends controlling the puck in the opposing team's zone. I calculated team-game CF depth ($M = 0.62$, $SD = 0.07$, $Min = 0.28$, $Max = 0.86$) according to the above definition of depth.

2.4.3 Expected Goals (xG)

xG is another advanced metric, and is the most complex shot metric included in this model. At its core, xG captures the probability of a shot being a goal. For instance, a shot with an xG value of 0.5 has a 50% chance of being a goal (MoneyPuck.Com - About and How It Works, n.d.). xG encompasses all shots taken by a player, including missed and blocked shots, and serves as a quantification of high quality scoring chances. I, again, calculated team-game xG depth ($M = 0.41$, $SD = 0.08$, $Min = 0.14$, $Max = 1.00$) according to the above definition of depth.

2.4.4 Assists

Moving away from shot-based metrics, assists as defined by the NHL are "awarded to the player or players (maximum of two) who touched the puck prior to the goal, provided no defender plays or possesses the puck in between" (Official Site of the National Hockey League | NHL.Com, n.d.). Assists are recorded in the NHL API play-by-play endpoint as the player ID for each of the two assisting players for every goal event, if there are assisting players. Any goals without assists are simply missing values. I, again, calculated team-game assist depth ($M = 0.28$, $SD = 0.22$, $Min = 0.06$, $Max = 1.00$) according to the above definition of depth.

2.4.5 Time-on-Ice (TOI)

Finally, I utilized data from the NHL API shift charts endpoint (`https://github.com/Zmalski/NHL-API-Reference`) to calculate depth in play time across the roster. The NHL records the length of time for every shift, for every player, in every game. Thus, within each team-game I summed every skater's total shift time to calculate the actual depth in roster play time ($M = 0.84$, $SD = 0.03$, $Min = 0.58$, $Max = 1.00$). Notably, ice time was the most equally distributed resource, suggesting fairly minimal ice time inequality across the 15 measured seasons.

2.5 Analytic Approach

Because this paper is focused on creating and validating a measurement model of an unobserved latent construct, I rely primarily on structural equation modeling. I first assess correlations among candidate depth indicators (e.g., Shot, CF, xG, assist, and TOI depth) and their relationships to total shot volume and xG to check for redundancy and multicollinearity. Next, I build a set of structural equation models to validate whether a single latent factor adequately represents Shot, CF, xG, assist, and TOI depth indicators. Additionally, I show that this model demonstrates metric invariance and partial scalar invariance across both season and team. Next, I explore how a rolling ten-game average of depth, and depth differential, can reliably predict matchup outcomes better than a baseline model. Finally, I use mediation models to explore whether, and why, model-implied factor scores of depth predict matchup outcomes. I conducted all data wrangling in python, and all analysis in R. All materials (codebook, preprocessing description, full final data sets, and sample raw data sets) are available on this project's Github repository (`https://github.com/dwiwad/tdi-peer-review`).

3 Construct Validation

3.1 Inter-indicator Correlations

I calculated pearson-product moment correlations at the game level between total shots on goal, xG, CF, shot depth, xG depth, CF depth, assist depth, and TOI depth (Table 1). There are four notable patterns here. First, overall correlations range from -.05 to .63, but correlations among the depth metrics range from .01 to .55. These moderate correlations among the depth indicators suggest that they are sufficiently distinct and capture overlapping, but not redundant aspects of depth. Second, the depth metrics are only moderately correlated with total shots, xG, and CF, ranging from .00 to .44. This, critically, suggests that the depth indicators are not simply redundant with volume of shot or point production.

Table 1: Pearson product–moment correlations among performance outcomes and depth indicators (team-game level).

	1	2	3	4	5	6	7	8
1. Total shots	—							
2. Total xG	.56***	—						
3. CF	.63***	.37***	—					
4. Shot depth	.42***	.19***	.33***	—				
5. xG depth	.34***	.03	.39***	.51***	—			
6. CF depth	.28***	.16***	.44***	.55***	.43***	—		
7. Assist depth	−.03	−.05	.00	.01	.04	.02	—	
8. TOI depth	−.04	.00	.07***	.16***	.11***	.21***	.08***	—

Note. $N = 38,394$. $\dagger p < .10$, $*p < .05$, $**p < .01$, $***p < .001$.

Third, assist depth is not significantly correlated with any other indicator. Because assist depth demonstrated uniformly weak relationships with both performance metrics and other depth indicators (all rs <= .05), I excluded it from the latent depth construct. Finally, TOI depth also shows some variance in how it correlates with the other indicators. Specifically, TOI depth is weakly correlated to total shots, xG and CF, but is still moderately correlated with shot, xG, and CF depth. As such, I will leave it in the measurement model but note that it is likely to contribute weakly to the latent depth factor.

3.2 Measurement Model

In order to test whether depth can be modeled as a single latent construct captured by the shot, xG, CF, and TOI depth indicators I specified a structural equation model with each indicator loading onto a single latent variable (Figure 1). I fixed the latent depth variance to be 1 because fixing the first loading to 1 already assumes some degree of invariance. This places the latent depth construct on a unit-variance scale and allows all loadings to be freely estimated. Additionally, in the rest of this paper I report and utilize the standardized factor loadings.

This model demonstrated acceptable fit ($\chi^2(2) = 377.331$, $p < .001$; CFI = .987, RMSEA = .070, SRMR = .022) according to modern benchmarks (Hu & Bentler, 1999), with all four indicators loading significantly onto the single latent construct. This suggests that depth is indeed a unidimensional factor that can be effectively measured as a composite of shot ($\beta = .800$, $SE = .005$, $p < .001$, 95% CI [.789, .811]), xG ($\beta = .634$, $SE = .005$, $p < .001$, 95% CI [.623, .644]), CF ($\beta = .685$, $SE = .005$, $p < .001$, 95% CI [.675, .696]), and TOI depth ($\beta = .221$, $SE = .006$, $p < .001$, 95% CI [.210, .232]). Notably, as foreshadowed by the correlations in Table 1, TOI depth is the weakest

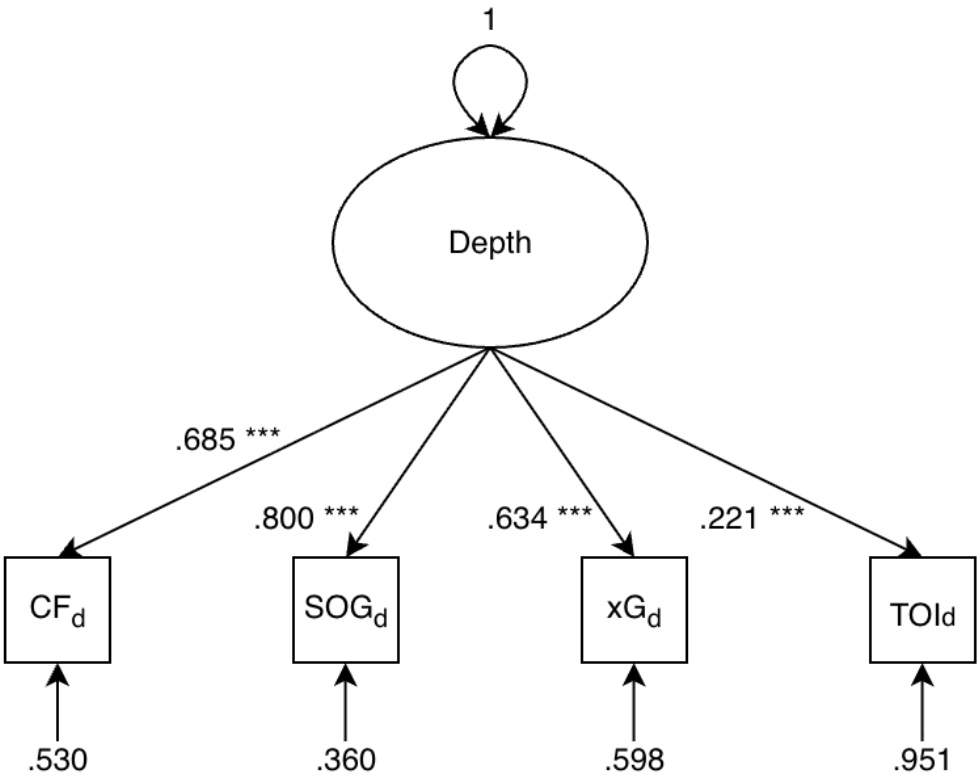

Figure 1: Measurement model for the TDI, computed across 15 seasons and 38,394 NHL regular and post season playoff team-games. CF_d is Corsi-For depth, SOG_d is shot depth, xG_d is xG depth, and TOI_d is time on ice depth.

contributor to depth; however, it still meaningfully contributes to the measurement model. This pattern supports the interpretation that depth is primarily driven by dispersion of shot-based metrics, with TOI contributing more modestly but still loading significantly on the construct.

3.3 Measurement Invariance

Measurement invariance refers to whether or not a measure is psychometrically equivalent across groups or time points. In other words, if we measure "depth" using the measurement model described above, does this latent construct represent the same underlying concept in the 2010–2011 season as it does in the 2024–2025 season, or within one NHL team compared to another? Without establishing invariance, it is unclear whether differences in the estimated factor scores reflect true differences in depth or simply differences in how the construct is measured across seasons or teams. Demonstrating measurement invariance is therefore essential for making valid comparisons of team depth across seasons and teams.

To assess measurement invariance, I conducted a standard sequence of increasingly strict tests. Configural invariance assesses whether the same factor structure holds across groups. Metric invariance tests whether each measurement indicator relates to the latent construct to the same degree across groups (i.e., the factor loadings). Finally, scalar invariance examines whether the baseline levels (i.e., the intercepts) of the measurement indicators are comparable, which is necessary to compare mean levels of depth.

Specifically, in each case I specify a set of three models. First, the configural model is the standard measurement model I specified above, except with a grouping factor (i.e., season or team). In this model, all loadings and intercepts are free to vary across seasons. Second, I fit the metric model in which all factor loadings are fixed to be equivalent across the grouping factors. Finally, I fit the scalar model in which all factor loadings and intercepts are fixed to be equivalent across the grouping factor. In each case, a null χ^2 difference test or minimal shifts in fit mean that the model has achieved the relevant sense of measurement invariance (Putnick & Bornstein, 2016).

3.3.1 Measurement Invariance Across Seasons

I first fit all three models with season as the grouping factor (Table 2). Notably, the configural model demonstrates acceptable fit, again suggesting that the unidimensional factor structure of depth is the same across seasons. Second, while the χ^2 difference test for the metric model is significant, this is to be expected with large sample sizes (Chen, 2007; Cheung & Rensvold, 2002; Putnick & Bornstein, 2016). Further, the changes in CFI (ΔCFI $= -.001$) and RMSEA (ΔRMSEA $= -.021$) are within conventional thresholds to consider metric invariance achieved (Chen, 2007; Putnick & Bornstein, 2016). This means that we can explore the relationship that depth has with other variables (e.g., win/loss matchup outcomes) across seasons. By contrast, the scalar model showed a substantial decline in fit (ΔCFI $= -.077$; ΔRMSEA $= .052$) relative to the metric model, suggesting that the model is not scalar invariant. This means mean-level comparisons of latent depth across seasons would not be valid if based on the full scalar model.

Table 2: Measurement Invariance Across Seasons.

Model	df	χ^2	$\Delta\chi^2$ (Δdf)	p-value	CFI	RMSEA	SRMR
Configural	30	327.89	—	—	0.990	0.062	0.017
Metric	72	379.06	51.17 (42)	0.003***	0.989	0.041	0.020
Scalar	114	2618.47	2239.41 (42)	$< .001$***	0.912	0.093	0.059
Partial Scalar	100	949.79	563.73 (28)	$< .001$***	0.970	0.057	0.030

Note: The partial scalar model allows TOI_D to vary across seasons. The $\Delta\chi^2$ for the partial scalar model are relative to the metric model.

To address this, I used partial invariance testing to identify problematic indicators. Results showed that intercepts for TOI_D (and, to a lesser degree, SOG_D and CF_D) varied substantially across seasons (See Table S1). By relaxing the equality constraint on the TOI_D intercept, I estimated a partial scalar model. This model fit well (CFI $= .970$, RMSEA $= .057$), just slightly worse than the metric model (ΔCFI $= -.019$; ΔRMSEA $= .016$) but still within acceptable bounds. These results suggest that while strict scalar invariance is untenable, a partial scalar solution where TOI_D is free to vary across seasons provides a compromise, permitting comparisons of mean depth across seasons.

3.3.2 Across Teams

I followed the same process as above to establish measurement invariance across NHL teams, finding a very similar pattern to cross-season measurement invariance. The configural model demonstrates acceptable fit, again suggesting that the unidimensional factor structure of depth is the same across teams. Second, while the χ^2 difference test for the metric model is significant, the changes in CFI (ΔCFI $= -.005$) and RMSEA (ΔRMSEA $= -.023$) are within conventional thresholds to consider metric invariance achieved (Chen, 2007; Putnick & Bornstein, 2016). By contrast, the scalar model again showed a substantial decline in fit (ΔCFI $= -.087$; ΔRMSEA $= .049$) relative to the metric model, suggesting that the model is not scalar invariant.

Table 3: Measurement Invariance Across NHL teams.

Model	df	χ^2	$\Delta\chi^2$ (Δdf)	p-value	CFI	RMSEA	SRMR
Configural	70	460.98	—	—	0.986	0.071	0.020
Metric	172	690.74	229.76 (102)	$< .001$***	0.982	0.052	0.031
Scalar	274	3266.97	2576.23 (102)	$< .001$***	0.895	0.100	0.070
Partial Scalar	240	1025.12	334.38 (68)	$< .001$***	0.972	0.055	0.036

Note: The partial scalar model allows TOI_D to vary across teams. The $\Delta\chi^2$ for the partial scalar model are relative to the metric model.

I again use partial invariance testing to identify problematic indicators (See Table S2). Results showed that intercepts for TOI_D (and, to a lesser degree, SOG_D and CF_D) varied substantially across teams. By relaxing the equality constraint on the TOI_D intercept, I estimated a partial scalar model. This model fit well (CFI $= .972$, RMSEA $= .055$), again just slightly lower than the metric model (ΔCFI $= -.010$; ΔRMSEA $= .003$) but still within acceptable bounds. These results suggest that while strict scalar invariance is untenable, a partial scalar solution where TOI_D is free to vary across teams provides a compromise, permitting comparisons of mean depth across teams.

3.3.3 Conclusion

Across a series of structural equation models, measurement invariance, and partial measurement invariance tests I find that the latent construct of "depth" in NHL hockey is a reliably unidimensional construct. In particular, depth as a construct can be accurately described as, and measured by, within-team within-game dispersion of shot production (i.e., shots-on-goal, xG), offensive pressure (i.e., CF), and actual roster depth (i.e., time on ice). Further, this model of depth demonstrates configural and metric invariance across both teams and seasons. This means that we can accurately compare the relationship depth has with other constructs (e.g., matchup outcomes) across teams and seasons. Further, while full scalar invariance was not achieved, relaxing the criteria and allowing time on ice depth to vary across seasons and teams achieves partial scalar invariance. Thus, moving forward, we can compare depth across seasons and teams utilizing this slightly altered model.

3.4 Depth in the NHL

Next, in line with previous research measuring latent constructs (Riedl et al., 2021; Woolley et al., 2010), I extract factor scores from the above specified models to explore the structure and predictive validity of depth in the NHL from 2010 to 2025. Specifically, I look at simple descriptive statistics of depth, how depth has changed across seasons, the game-to-game stability of depth, and how depth varies by team. Given metric invariance across seasons and teams, I examine the associations between Depth and performance outcomes across groups with the full model. Where mean comparisons are required (e.g., comparing average team depth), I use the partial-scalar solution (TOI_D intercept freed).

3.4.1 Distribution and Stability

To characterize the temporal and distributional properties of the TDI, I (a) visualized the distribution of team-game factor scores, (b) computed intraclass correlations to partition variance between teams and within teams, (c) estimated lag-1 autocorrelations within team-seasons, and (d) examined the predictive association between a 10-game rolling mean of depth and subsequent game-level depth.

Across all 15 seasons, the TDI follows an approximately normal distribution, as expected given its standardization ($M = 0$, $SD = 1$). Factor scores range from –4.61 to 3.68, indicating substantial variability in game-level depth. In other words, while most teams exhibit average depth from game to game, there are frequent instances of both highly top-heavy and highly balanced performances. Additionally, the intraclass correlation at the team level ($ICC = 0.019$) indicated that only about 1.9% of the total variance in game-level depth occurred between teams, with the vast majority (98.1%) reflecting within-team, game-to-game variability. This suggests that depth is primarily a situational property rather than a stable team characteristic.

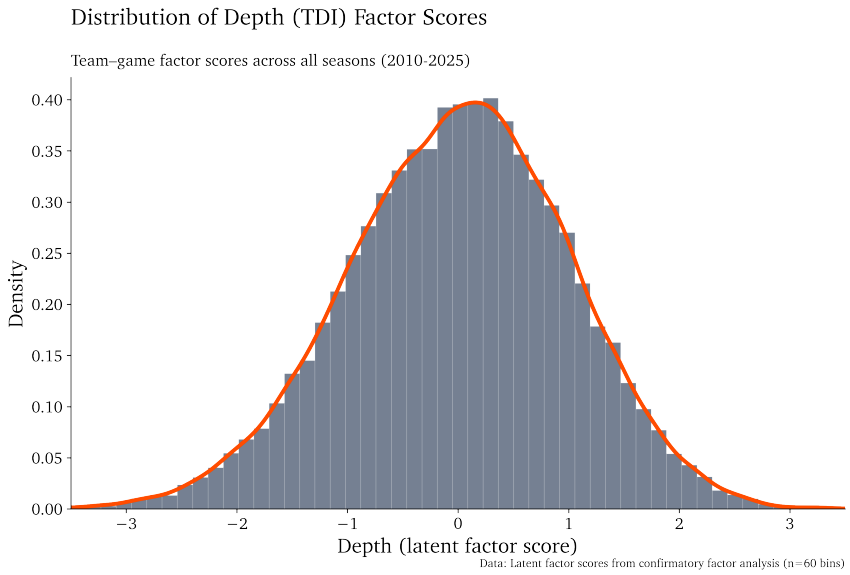

Figure 2: Distribution of Depth Across all Fifteen NHL Seasons from 2010 to 2025.

This assertion is supported by the fact that, across 462 team-seasons, the median lag-1 autocorrelation of depth was essentially zero ($r = .013$, IQR $= [-.060, .085]$). This indicates that single-game depth fluctuates almost randomly from one game to the next, with little short-term persistence. Although a handful of teams exhibited modest positive or negative autocorrelation (Figure 3), such patterns were rare, underscoring that depth is a highly situational game-level property rather than a stable team characteristic.

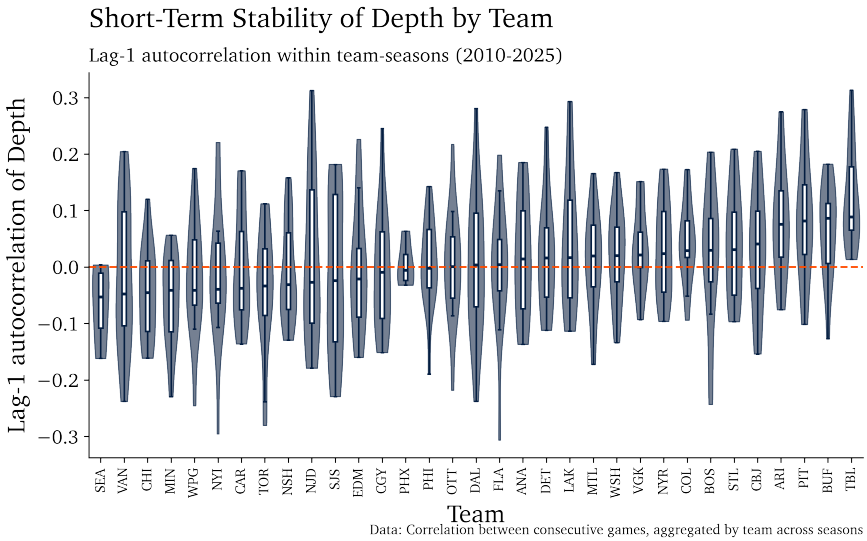

Figure 3: Distribution of lag-1 autocorrelation within each team, across all seasons. The 2010–2011 Atlanta Thrashers and 2024–2025 Utah Hockey Club were removed from this figure as they each only had one season and thus no variance in lag-1 autocorrelation.

Given this near-zero short-term stability, I next examined whether aggregating across a larger window yields more persistent patterns of depth. In order to establish how many games I should aggregate over, I estimated a series of multilevel models in which next-game depth was regressed onto the rolling mean of depth computed over the previous n games, with random intercepts for team and season. I evaluated window sizes ranging from 1 to 30 games, extracting the fixed-effect slope of rolling depth as an index of predictive stability (Figure S1). The fixed-effect slope for the rolling predictor increased monotonically with window length (e.g., $\beta = .37$ at $n = 10$, $\beta = .52$ at $n = 20$, and $\beta = .60$ at $n = 30$), indicating that broader aggregation systematically improves predictability of next-game depth. However, the incremental gain per additional game diminished beyond roughly 10 games (Table S3) and longer windows also represent an increasingly large fraction of an 82-game season and reduce the number of within-season observations available for rolling estimation. I therefore selected a 10-game rolling average in order to dampen the game-to-game noise but also remain responsive to meaningful within-season changes and to preserve adequate within-season coverage.

Next, I predicted next-game depth as a function of each team's previous 10-game rolling average, with random intercepts for team and season. As shown in Figure 4, the rolling 10-game average of depth predicts subsequent Depth ($\beta = 0.36$, 95% CI [0.33, 0.38], $p < .001$). This suggests that when a team has been deeper over its previous 10 games, it tends to remain deeper in its next matchup. This suggests that, while single-game Depth fluctuates substantially, averaging across several games reveals a more stable, enduring pattern.

3.5 Does depth help you win?

So far I have analyzed depth at the team-game level. That approach tells us whether depth is a stable team trait, and how it relates to an individual team's outcome of any given game. Of course, hockey is inherently head-to-head. So, in order to truly examine whether depth leads to more wins it is necessary to move from separate team level analysis to matchup level analysis. Only at this level can we ask whether being the deeper team on a given night translates into a higher probability of winning that specific matchup.

Even though I estimated and validated depth at the team-game level, the relevant causal/comparative question in a head-to-head sport is inherently relative: whether a team's Depth exceeds its opponent's in the same contest. Accordingly, I operationalize matchup-level depth advantage as a depth differential (ΔDepth = home depth − away depth). This does not introduce a new construct or change the measurement model; rather, it is a matchup-level re-expression of

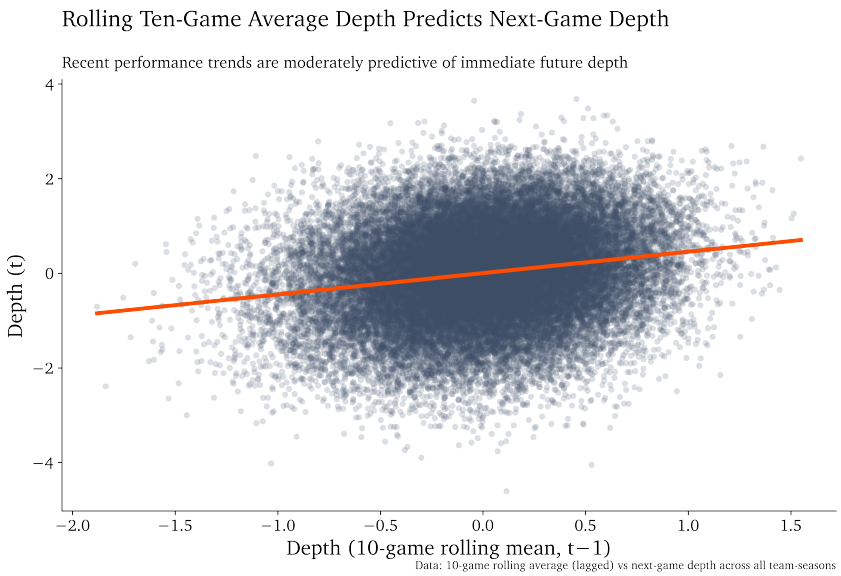

Figure 4: Relationship between 10-game rolling Depth and next-game Depth.

the same latent depth scores on a common scale. Modeling ΔDepth is equivalent to entering home and away depth simultaneously and testing whether the contrast between them predicts the win probability. This framing aligns the focal construct (team depth) with the unit of competition (the matchup), while keeping the interpretation directly tied to the question of whether being the deeper team provides an advantage on a given night.

So, to examine whether within-game differences in team depth relate to the outcome of that same game, I fit a mixed-effects logistic regression model predicting home-team win from the standardized home–away depth differential, with a random intercept by season. The model revealed that playing with more depth than your opponent in a given matchup actually predicted lower chances of winning ($\beta = -0.046$, $SE = 0.015$, 95% CI [-.075, -.018], $p = .001$). The random season intercept was near zero ($SD = 0.03$), indicating a stable baseline home-win rate across seasons. However, as we have seen already, within-game depth fluctuates substantially from one matchup to the next, and thus may capture short-term variance rather than a team's underlying level of play.

To explore a more stable indicator of team depth I again utilized the differential between the home and away team's 10-game rolling average of depth to predict matchup outcome in another mixed-effects logistic regression. This model revealed that, inclusive of the matchup in question, teams with higher depth over their ten most recent matchups were more likely to win the current game ($\beta = 0.096$, $SE = 0.015$, 95% CI [.067, .124], $p = .001$). Importantly, the effect of a 10-game rolling depth average on a games outcome (Figure 5) holds, and in fact is slightly stronger ($\beta = 0.121$, $SE = 0.016$, 95% CI [.089, .153], $p < .001$), when controlling for the differentials in Corsi-For ($\beta = -.714$, $SE = 0.020$, 95% CI [-.752, -.675], $p < .001$) and xG ($\beta = 1.122$, $SE = 0.022$, 95% CI [1.079, 1.165], $p < .001$).

The models above suggest that longer-run deeper teams are more likely to win, specifically when depth is measured over a sustained stretch of games. But does this relationship actually generalize to out-of-sample data? To test whether depth provides predictive power rather than a season-specific correlation, I next used a leave-one-season-out cross-validation approach, evaluating how well a depth-based model (home win as a function of lagged 10-game rolling depth, controlling for 10-game rolling xG_D and CF_D) predict game outcomes across held-out seasons. I also compare this to a simple baseline "home team wins" model, simply predicting that the home team will win the game. Historically, home teams win about 54% of the time. Additionally, I compare this to the MoneyPuck pre-game prediction model as a gold standard complex prediction model. I would expect this more simplistic depth model to perform better than baseline, but worse than the MoneyPuck model.

Overall, the depth based model ($M_{acc} = 56.2\%$, AUC = .571, Brier = .245, Log Loss = .682) performs slightly above the home team wins baseline ($M_{acc} = 54.2\%$, AUC = .571, Brier = .252, Log Loss = .689). Across all 15 seasons, as expected, the average accuracy of predicting the home team to win ranges between 52.0% and 58.1%. The average accuracy of the model with TDI as well as xG_D and CF_D ranges between 53.9% and 60.6%. Additionally, the depth model outpredicts the home team wins baseline in 14 out of 15 seasons, notably performing better in the post-covid years (Figure 6) and showing consistently better log loss. The accuracy and log loss for the MoneyPuck model are

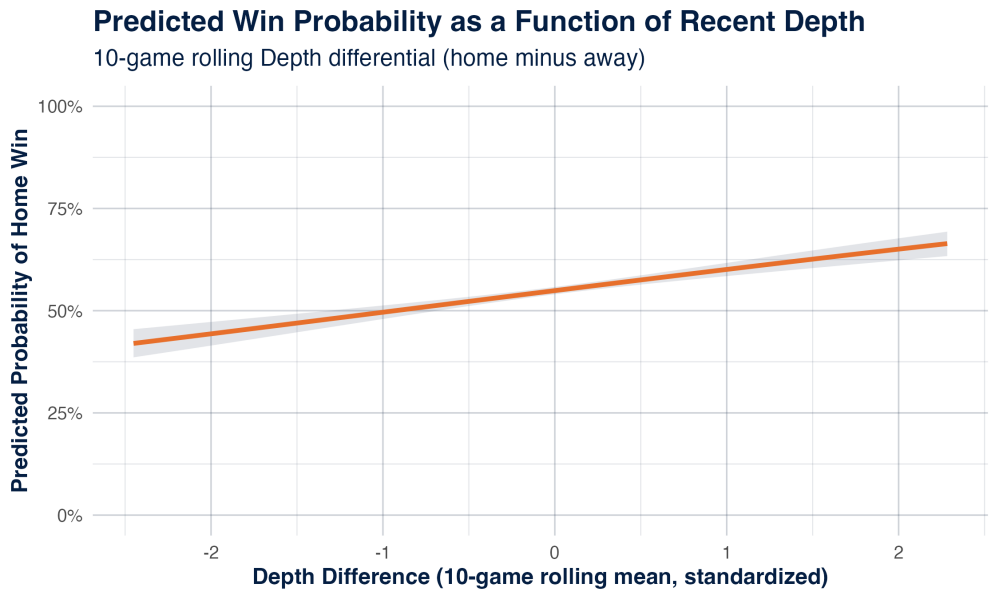

Figure 5: Relationship between 10-game rolling depth differential and probability of winning.

provided from 2020 until present. As we would expect to see, the depth based model under-performs the considerably more sophisticated MoneyPuck prediction model, a model with many indicators designed explicitly to predict matchup outcomes.

3.5.1 Why does depth help you win?

The models above establish that, within a season, a team's 10-game rolling average of depth significantly predicts game outcomes. Specifically, teams with greater depth recently are more likely to win their next matchup. One key open question, however, is why depth predicts winning. A plausible mechanism is that greater depth leads to higher offensive activity (e.g., more shots, more puck possession, and higher-quality scoring chances) which in turn increase a team's likelihood of winning. To test this explanation, I estimated a mediation model in which lagged 10-game depth (Depth_{t-1}) predicts expected goals (xG), shots-on-goal (SOG), and Corsi-For (CF), which in turn predict the outcome of the next matchup. As expected, teams that played with greater depth over the previous ten games generated more total shots ($\beta = 0.314$, $SE = 0.013$, $p < .001$, 95% CI [0.288, 0.340]), more offensive zone possession (CF; $\beta = 0.344$, $SE = 0.013$, $p < .001$, 95% CI [0.317, 0.370]), and more high-quality scoring chances (xG; $\beta = 0.244$, $SE = 0.013$, $p < .001$, 95% CI [0.218, 0.270]) in their next matchup.

When predicting the outcome of this matchup, only xG positively predicted the likelihood of winning ($\beta = 0.163$, $SE = 0.003$, $p < .001$, 95% CI [0.157, 0.169]). In contrast, shots on goal ($\beta = -0.035$, $SE = 0.004$, $p < .001$, 95% CI [–0.043, –0.028]) and Corsi-For ($\beta = -0.067$, $SE = 0.003$, $p < .001$, 95% CI [–0.074, –0.061]) were negatively associated with winning once xG was accounted for. This pattern indicates a suppression effect such that, when controlling for the quality of shots (xG), shot volume and possession no longer predict success and may even be counterproductive. Thus, we see that depth contributes to winning primarily through creating higher-quality scoring opportunities (xG; $\beta_{\text{indirect}} = 0.040$, $SE = 0.002$, $p < .001$, 95% CI [0.035, 0.044]), not merely increasing the quantity of attempts (SOGs; $\beta_{\text{indirect}} = -0.011$, $SE = 0.001$, $p < .001$, 95% CI [–0.014, –0.009]) or offensive pressure (CF; $\beta_{\text{indirect}} = -0.023$, $SE = 0.001$, $p < .001$, 95% CI [–0.026, –0.020]).

3.5.2 Conclusion

Taken together, these findings demonstrate that depth in the NHL is a highly situational but consequential property of team performance. While single-game depth fluctuates substantially from one matchup to the next, averaging across several games reveals a more stable pattern that carries meaningful predictive power. Teams that maintain greater depth relative to their opponents over a sustained window of play are reliably more likely to win their next matchup, even when controlling for conventional predictors such as shot-share and xG differentials.

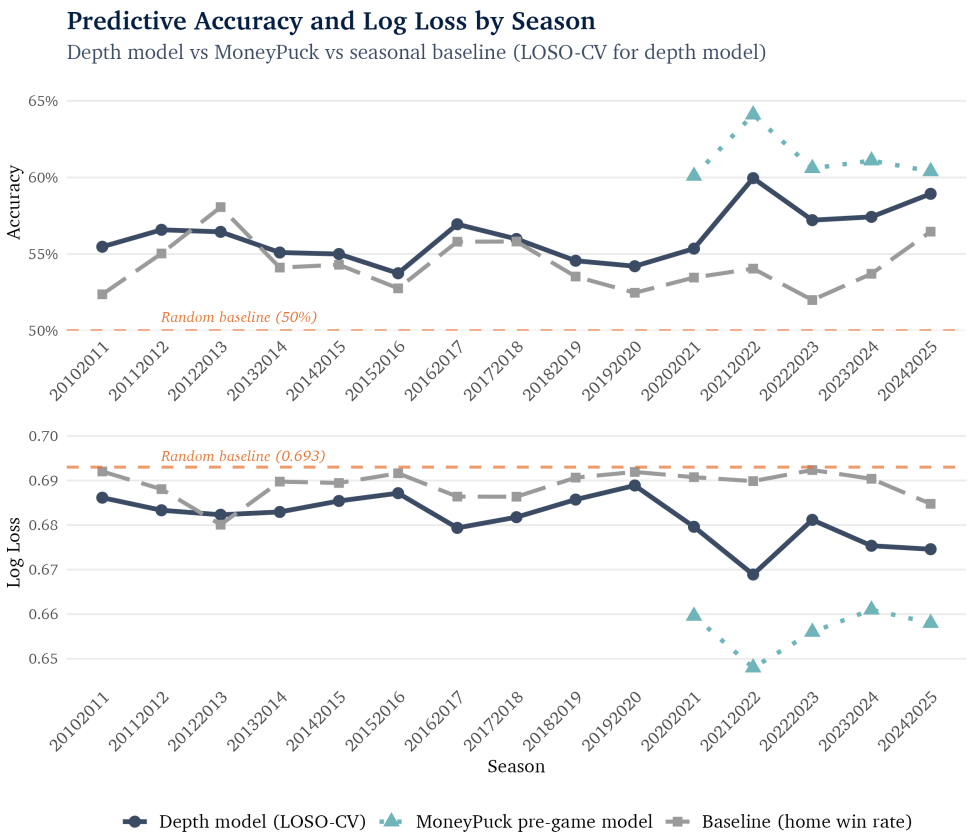

Figure 6: Accuracy and Log Loss season by season of the depth model compared to the home team wins baseline and MoneyPuck pre-game prediction model.

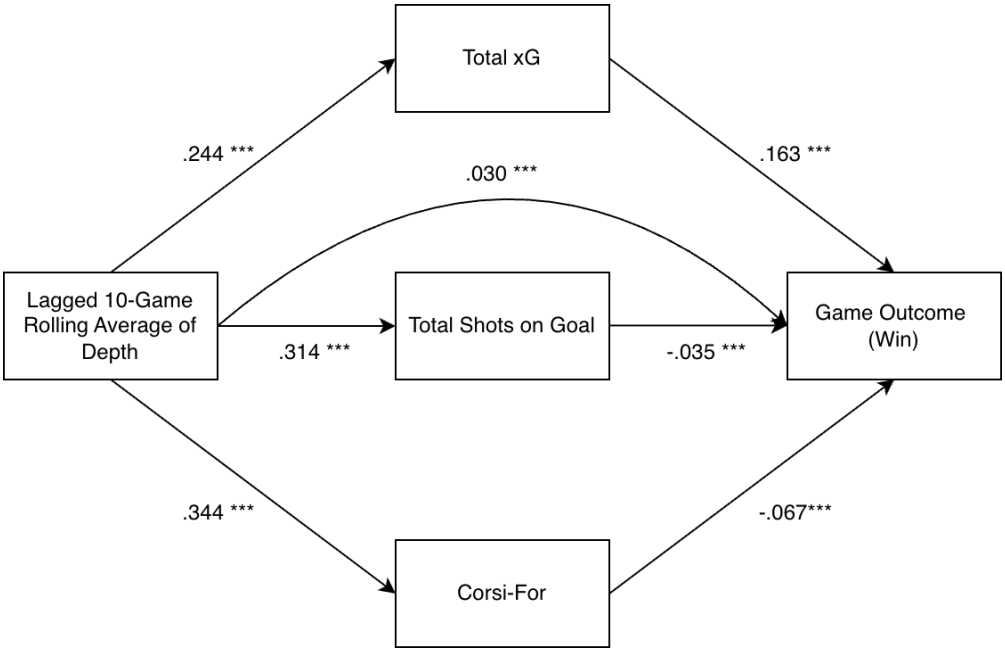

Figure 7: Mediation model of 10-game rolling average of depth predicting next-game outcome via higher high-quality scoring chances, total shots on goal, and offensive pressure.

Importantly, the mechanism through which depth contributes to winning is not volume, but quality. Depth predicts sustained offensive pressure and shot production, but its real advantage lies in creating higher-quality scoring opportunities. Teams whose 10-game rolling depth differential favors them (e.g., they have been deeper than their opponents over recent games) are reliably more likely to win their next matchup, even after accounting for differences in shot share and expected goals.

In this way, the Total Depth Index captures a dimension of performance that conventional metrics like xG or CF alone cannot: the distributional structure of contribution within teams. Depth reflects how evenly teams are able to spread offensive impact across players, and its predictive value suggests that balance plays a key role in success over the course of a season.

4 General Discussion

Although team depth is often discussed in the context of professional sports, especially Hockey, this work reveals that it is a statistically robust, latent construct that can modestly predict team success. Across fifteen seasons of National Hockey League data and millions of individual plays, I mapped the distribution of player contributions within a roster, examined the stability of these contributions over time, and assessed the specific mechanisms through which depth influences game outcomes to develop the Total Depth Index (TDI).

In this work, I offer two primary contributions to the fields of sports analytics and statistics. First, I demonstrate that the TDI serves as a robust predictor of game outcomes, demonstrating independent predictive power above the traditional "home-team wins" baseline model. Critically, I also clarify the mechanism behind this predictive success: depth drives winning not by increasing the sheer volume of shots or offensive pressure, but by specifically enhancing the quality of scoring opportunities, as measured by expected goals (xG). Beyond these practical findings, I make a methodological contribution by demonstrating how latent variable modeling and psychometric techniques (e.g., construct validation, structural equation modeling, and measurement invariance testing) can be leveraged to formally define and validate intangible constructs that have long eluded systematic measurement in sports analytics.

One primary finding of this research concerns the stability of depth. One could hypothesize that depth is (a) a stable team level trait that remains consistent within a season, or (b) a state that fluctuates to some degree. This work shows that depth is less of a stable organizational trait and more of a dynamic game-to-game construct that demonstrates stability when measured over time. Single game-to-game depth is marked by high levels of situational volatility. That is, depth in game g does not predict depth in game g+1; observed depth varies considerably within a team game-to-game. However, depth stabilizes when observed over longer time horizons. Specifically, a team's ten-game rolling average of depth measured from game g-9 to game g is indicative of how much depth they will demonstrate at game g+1. This distinction is critical, as it suggests that the "depth" of a team is best understood as a sustained semi-stable performance state rather than a static trait that is inherent to an organization.

Secondly, I demonstrate that depth is a reliable indicator of winning matchups in the NHL. Within a single matchup, depth differential did not predict matchup outcomes. That is, playing with more depth than your opponent on a given night did not reliably predict a win. However, teams that have been recently playing with more depth than their opponent (as measured by the ten-game rolling average depth differential) were more likely to win. Critically, leave-one-season-out cross-validation demonstrated that the differential in ten game rolling depth averages predicts matchup outcomes with more accuracy than the "home team wins" baseline prediction model, but less than Moneypuck's game prediction model. This demonstrates that sustained game-over-game depth is an important and reliable indicator of team success.

Finally, I uncover at least one mechanism through which depth facilitates matchup success. The model I ran simultaneously tests two hypotheses: volume versus quality. On one hand, depth could drive success simply through more offensive production. When you have more players contributing, a team takes more shots, makes more shot attempts, and thus scores more goals. On the other hand, depth could drive success through higher quality offensive production. That is, when you have more players contributing they may produce more high-quality scoring chances. This work clearly demonstrates that depth does not function through volume, but instead through quality. That is, while depth does predict higher shots on goal and Corsi-For, those metrics alone do not predict winning. However, depth predicts more high quality shots (as measured by expected goals), which does in turn predict winning. This suggests that the presence of a high-depth roster forces a distributional pressure on opponents that opens up high-danger opportunities, rather than simply increasing the number of attempts on net.

All of these findings together suggest that it is important for team General Managers not to overlook the importance of depth. The present results indicate that higher concentration in a roster can come at the expense of sustained team performance. Teams that consistently operate with greater depth are better positioned to create high-quality scoring opportunities and withstand matchup variability.

4.1 Limitations

One of the main limitations of this work is naturally the fact that all of my empirical explorations of depth as a construct were within the confines of the NHL. While the concept of depth as I have defined it could indeed carry predictive or explanatory weight in other sports (or other hockey leagues besides the NHL), the entire statistical justification for this construct in this work relies on NHL data. Therefore, future work should explore the psychometric properties of a similar dispersion-based metric of depth in other sports. For example, one might explore how depth in metrics such as batting average, on-base percentage, or slugging percentage predict the outcome of major league baseball matchups.

An additional limitation is that, in exploring this in the NHL, I was restricted to data that was widely available and easily accessible. It is possible that other metrics could meaningfully contribute to the construct of depth in NHL hockey but are difficult to impossible to access. For example, I utilized Corsi-For as a shot-based proxy for time-based offensive zone possession time. The greater depth a team has, the more time they should spend controlling the puck in the opposing team's zone. While a more robust time-based metric exists, such as offensive zone possession time, it is cost prohibitive to access. Thus, this work is limited, both in time and scope, by the data that is provided by the NHL and other advanced hockey stats community members.

Finally, as noted in the measurement invariance section, the final model achieved only partial scalar invariance across seasons and teams. While this level of invariance is generally sufficient to support comparisons of latent relationships and regression paths over time, it limits the strength of conclusions regarding absolute differences in latent mean levels of depth across seasons and teams. This suggests that although the underlying structure of depth is largely stable, some indicators may shift in their intercepts as the league evolves, reflecting changes in playing style, roster construction, or broader structural features of the NHL. Future work could address this limitation by adopting more flexible longitudinal approaches (e.g., latent growth curve models or models with time-varying covariates), which would allow depth to be modeled not only as a stable construct but also as one that changes over time or across teams.

4.2 Future Directions

Future work could examine whether depth relates to outcomes beyond win probabilities. For example, depth may plausibly buffer teams against performance shocks such as injuries or fatigue by distributing the physical and cognitive load more evenly and reducing dependence on a small subset of superstar players. This can be tested directly by modeling depth as a mediator between roster characteristics like injury burden and downstream outcomes such as expected goals, goal differential, or matchup outcome. That is, does depth actually function as roster shock resilience, or is depth primarily just a proxy for overall team quality?

A second direction is to establish external validity by testing whether the depth aligns with the most common non-analytic proxy for roster depth: salary balance. In hockey discourse, teams are often described as "deep" when payroll is less concentrated among stars, implicitly assuming that salary dispersion maps onto on-ice dispersion of contribution and skill. Future work should test whether salary inequality (i.e., roster salary Gini) correlates with performance depth as captured by the TDI. A strong association would link economic roster construction to functional depth; a weak association would clarify that a balanced salary roster and distributed shot and goal production are distinct constructs, and that depth is not something that can be inferred reliably from cap structure alone.

Although the present work operationalizes depth via the complement of the Gini coefficient, future work could further evaluate whether the construct is robust to alternative dispersion metrics. Measures such as the Theil Index or entropy-based metrics differ in their sensitivity to tail behavior and extreme concentration, and thus provide a useful stress test against the concern that the TDI's predictive results are an artifact of a specific inequality index. Formal convergent validity across dispersion measures would strengthen the claim that depth reflects a fundamental distributional property of team performance rather than a modeling choice.

Finally, the current model is deliberately restricted to skater depth and does not incorporate goaltending, despite the fact that goalie performance can dominate matchup outcomes. As this work shows, depth is not the be-all end-all for predicting wins. One likely reason for instances in which a team with higher depth loses a given matchup is differentials in goalie performance. Future work could explore increased predictive power of a depth-based model by factoring in rolling averages of each team's goals saved above average (e.g., a metric of goaltending performance). It is possible that skater depth and goalie performance (a) interact, (b) compensate for one another, or (c) operate independently in predicting outcomes. This would extend the latent-construct approach to a fuller roster-level account of how teams win.

5 Conclusion

Sports discourse is full of references to intangible constructs like depth, momentum, or team chemistry. Yet, sports analytics often reduces these ideas to single-statistic proxies or leaves them as metaphor. In this work, I rely on the concept of "team depth" in NHL hockey to show that we can instead treat these intangibles as latent variables and use a psychometric approach to define, validate, and measure them. Across fifteen seasons and millions of play-by-play and shift data, I show that depth behaves as a coherent unidimensional construct reflected by the within-team dispersion of shot production, expected goals, and offensive pressure, with time on ice contributing more modestly. Empirically, depth is not a stable game-to-game trait (i.e., single-game depth fluctuates nearly at random) but it stabilizes as a sustained state when measured over multi-game windows. At the matchup level, observed within-game depth differential does not reliably predict who wins on a given night, but teams that have been playing with greater depth over the prior ten games are modestly more likely to win a matchup, outperforming a simple home-win baseline model. Critically, depth does not win by sheer shot volume; instead it produces higher high-quality scoring chances. Taken together, the Total Depth Index moves depth from rhetoric to a validated construct and illustrates how latent variable modeling can formalize other long-invoked but weakly measured concepts in professional sports.

Acknowledgments

I thank Christopher To, Shawn Schwartz, and Brett Mercier for their input and thoughts on an earlier version of this paper.

References

[1] Hockey-Reference.Com. Advanced hockey statistics, n.d. Retrieved December 28, 2025.

[2] L. Argote and P. Ingram. Knowledge transfer: A basis for competitive advantage in firms. *Organizational Behavior and Human Decision Processes*, 82(1):150–169, 2000.

[3] A. Benson, T. Brown, and E. Zuckerman. An analysis of nhl salary optimization. Northwestern Sports Analytics Group, 2024.

[4] D. M. Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232, 2014.

[5] A. Caspi, R. M. Houts, D. W. Belsky, S. J. Goldman-Mellor, H. Harrington, S. Israel, M. H. Meier, S. Ramrakha, I. Shalev, R. Poulton, and T. E. Moffitt. The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, 2(2):119–137, 2014.

[6] F. F. Chen. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3):464–504, 2007.

[7] G. W. Cheung and R. B. Rensvold. Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2):233–255, 2002.

[8] A. M. Christie and J. Barling. Beyond status: Relating status inequality to performance and health in teams. *Journal of Applied Psychology*, 2010.

[9] B. Condor. The fyi on toi: Shorter shifts win. Seattle Kraken, January 2024.

[10] R. Dorfman. A formula for the gini coefficient. *The Review of Economics and Statistics*, 61(1):146–149, 1979.

[11] S. El-Den, C. Schneider, A. Mirzaei, and S. Carter. How to measure a latent construct: Psychometric principles for the development and validation of measurement instruments. *International Journal of Pharmacy Practice*, 28(4):326–336, 2020.

[12] O. Guest. Oliviaguest/gini, 2025. Original work published 2016. Python.

[13] N. Halevy, E. Y. Chou, A. D. Galinsky, and J. K. Murnighan. When hierarchy wins. *Social Psychological and Personality Science*, 2011.

[14] W. C. Horrace, H. Jung, and S. Sanders. Network competition and team chemistry in the nba. *Journal of Business & Economic Statistics*, 40(1):35–49, 2020.

[15] L. Hu and P. M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, 1999.

[16] T. Kerdman. The nhl salary cap allocation theory. Medium, October 2016.

[17] D. Luszczyszyn. Improving our nhl projection model ahead of the 2019-20 season. *The New York Times*, June 2019.

[18] D. Luszczyszyn. In the nhl playoffs, what's more valuable: Star power or depth? *The New York Times*, April 2023.

[19] J. G. March. Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87, 1991.

[20] Moneypuck.com—about and how it works, n.d. Retrieved December 27, 2025.

[21] J. T. P. Noel, V. Prado da Fonseca, and A. Soares. A comprehensive data pipeline for comparing the effects of momentum on sports leagues. *Data*, 9(2):29, 2024.

[22] NHL.com. Official site of the national hockey league, n.d. Retrieved December 28, 2025.

[23] D. L. Putnick and M. H. Bornstein. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41:71–90, 2016.

[24] M. Qiu, S. Zhang, Q. Yi, C. Zhou, and M. Zhang. The influence of "momentum" on the game outcome while controlling for game types in basketball. *Frontiers in Psychology*, 15:1412840, 2024.

[25] C. Riedl, Y. J. Kim, P. Gupta, T. W. Malone, and A. W. Woolley. Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences*, 118(21):e2005737118, 2021.

[26] C. Spearman. "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201, 1904.

[27] J. Taylor and A. Demick. A multidimensional model of momentum in sports. *Journal of Applied Sport Psychology*, 6(1):51–70, 1994.

[28] E. C. Tupes and R. E. Christal. Recurrent personality factors based on trait ratings. *Journal of Personality*, 60(2):225–251, 1992.

[29] R. Vollman. *Stat Shot: A Fan's Guide to Hockey Analytics*. ECW Press, 2016.

[30] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 2010.