Panel Data Analysis of Interactive Effects between High-Danger Shots, Rebounds, and Secondary Skills on NHL Goal Production

1. Introduction

Modern hockey analytics has moved beyond traditional statistics to examine the complex factors driving goal production. High-danger shots, defined as shots taken from within 29 feet of the center of the goal and bounded by imaginary lines drawn from the face-off dots to 2 feet outside the goalpost [1] are strong predictors of goal scoring. However, secondary skills like rebound generation, physical play, and puck management may amplify or diminish their effectiveness [2]. Understanding these interactive relationships is crucial for player evaluation and team strategy.

Despite advances in hockey analytics, limited research has systematically examined how high-danger shooting combines with complementary skills to influence goal production. Previous research predominantly treats these factors independently, possibly missing critical interactive relationships. This study addresses this gap by analyzing five seasons of NHL data (2020-21 through 2024-25) using fixed-effects panel regression to control for unobserved player characteristics and examine skill complementarity in offensive production.

1.1 Research Question

How do high-danger shooting opportunities and rebound generation jointly influence goal production in hockey, and how do secondary skills (hits, takeaways, giveaways) interact with this relationship?

2. Methodology

2.1 Data Sources

This study utilizes comprehensive player-level performance data [3], covering five complete NHL regular seasons from 2020-21 through 2024-25. The dataset includes game-by-game statistics for all NHL players across multiple performance dimensions, providing detailed metrics on shooting, defensive actions, and situational factors.

```r
#load the libraries
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
library(plm)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plm':
##
##     between, lag, lead
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(sandwich)

#load files and add season indicator
data_24_25 <- read.csv("D:/hockey-skill-interactions-analysis/dataset/2024_2025.csv") %>% mutate(season = "2024_2
5")
data_23_24 <- read.csv("D:/hockey-skill-interactions-analysis/dataset/2023_2024.csv") %>% mutate(season = "2023_2
4")
data_22_23 <- read.csv("D:/hockey-skill-interactions-analysis/dataset/2022_2023.csv") %>% mutate(season = "2022_2
3")
data_21_22 <- read.csv("D:/hockey-skill-interactions-analysis/dataset/2021_2022.csv") %>% mutate(season = "2021_2
2")
data_20_21 <- read.csv("D:/hockey-skill-interactions-analysis/dataset/2020_2021.csv") %>% mutate(season = "2020_2
1")

#merge all seasons
data_all <- bind_rows(data_24_25, data_23_24, data_22_23, data_21_22, data_20_21)
```

2.2 Variable Construction

All performance metrics were standardized as per-minute rates to ensure comparability across players with varying ice time allocations. The dependent variable, goal production rate (`goals_rate`), represents goals scored per minute of ice time. Primary independent variables include high-danger shot rate (`hd_rate`) and rebound rate (`reb_rate`), while secondary skills encompass physical engagement (`phys_rate`), defensive takeaways (`take_rate`), and puck management (`give_rate`).

```r
#create rate variables per minutes
data_all <- data_all %>%
  mutate(
    hd_rate    = I_F_highDangerShots             / icetime,
    goals_rate = I_F_goals                       / icetime,
    mom_diff   = (xGoalsForAfterShifts - xGoalsAgainstAfterShifts) / icetime,
    reb_rate   = I_F_xGoals_with_earned_rebounds / icetime,
    phys_rate  = I_F_hits                        / icetime,
    take_rate  = I_F_takeaways                   / icetime,
    give_rate  = I_F_giveaways                   / icetime,
    o_d_start_diff = I_F_oZoneShiftStarts - I_F_dZoneShiftStarts)

#remove missing values
data_clean <- na.omit(data_all)
```

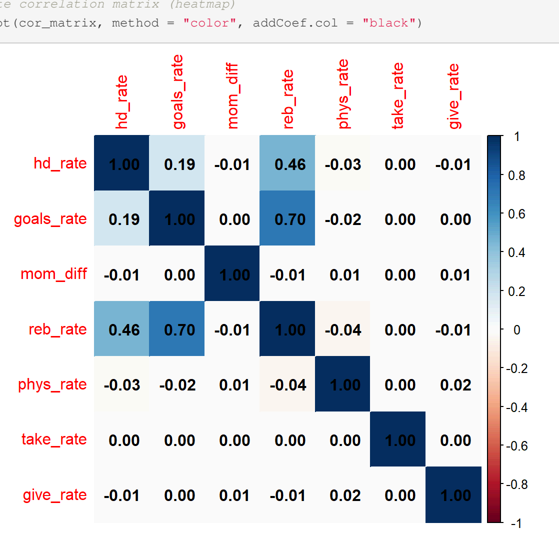
2.3 Exploratory Data Analysis

Prior to model estimation, correlation analysis was conducted to examine the relationships between key variables and identify potential multicollinearity concerns.

```r
#compute correlation matrix
cor_vars <- c("hd_rate","goals_rate","mom_diff",
              "reb_rate","phys_rate","take_rate","give_rate")
cor_matrix <- cor(data_clean[, cor_vars])
round(cor_matrix, 3)
```

```
##            hd_rate goals_rate mom_diff reb_rate phys_rate take_rate give_rate
## hd_rate      1.000      0.186   -0.006    0.459    -0.029    -0.003    -0.010
## goals_rate   0.186      1.000   -0.005    0.704    -0.019    -0.001    -0.002
## mom_diff    -0.006     -0.005    1.000   -0.010     0.006    -0.001     0.011
## reb_rate     0.459      0.704   -0.010    1.000    -0.043     0.000    -0.009
## phys_rate   -0.029     -0.019    0.006   -0.043     1.000    -0.004     0.018
## take_rate   -0.003     -0.001   -0.001    0.000    -0.004     1.000     0.002
## give_rate   -0.010     -0.002    0.011   -0.009     0.018     0.002     1.000
```

```r
#compute correlation matrix (heatmap)
corrplot(cor_matrix, method = "color", addCoef.col = "black")
```



2.4 Panel Data Structure

The analysis utilized a balanced panel design where individual players observed across multiple games within each season. To create unique time identifiers, games were sequentially numbered within each season for each player, then combined with season indicators to generate unique `season_game` identifiers.

```r
#create an unique identifier for each game within each player for each season
df_games <- data_clean %>%
  arrange(playerId, season) %>%
  group_by(playerId, season) %>%
  mutate(game_row = row_number()) %>%
  ungroup()

#create a new columns that merge season and game_row to become season_game that uniquely identifies each game ins
ide a season
df_games <- df_games %>%
  mutate(season_game = paste(season, game_row, sep = "_"))

#build panel data frame on playerId and season_game so that the fixed effects with plm can be conducted
df_panel_games <- pdata.frame(
  df_games,
  index = c("playerId", "season_game"),
  drop.index = TRUE
)

#check for duplicate pairs which should be zero
df_check <- df_games %>% count(playerId, season_game) %>% filter(n > 1)
df_check
```

```
## # A tibble: 0 × 3
## # ℹ 3 variables: playerId <int>, season_game <chr>, n <int>
```

2.5 Model Specification

The primary approach utilizes fixed-effects panel regression to investigate both main effects and interaction patterns between high-danger shooting and complementary skills. The fixed-effects ("within") estimator controls for unobserved player-specific characteristics that remain constant over time, such as innate shooting ability or positional tendencies.

```r
#built model with interaction terms
fe_goals_model <- plm(goals_rate ~ hd_rate * reb_rate + hd_rate * phys_rate + hd_rate * take_rate + hd_rate * giv
e_rate, data = df_panel_games, model = "within")

#look at the summary
summary(fe_goals_model)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = goals_rate ~ hd_rate * reb_rate + hd_rate * phys_rate +
##     hd_rate * take_rate + hd_rate * give_rate, data = df_panel_games,
##     model = "within")
##
## Unbalanced Panel: n = 1460, T = 2-25, N = 22117
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -1.9073e-02 -7.8727e-05  8.0314e-05  2.3081e-04  1.1962e-01
##
## Coefficients:
##                       Estimate  Std. Error  t-value  Pr(>|t|)
## hd_rate            -1.5518e-02  7.5151e-03  -2.0649   0.03894 *
## reb_rate            2.7941e+00  1.8134e-02 154.0833 < 2.2e-16 ***
## phys_rate           2.7689e-02  4.7661e-03   5.8096 6.354e-09 ***
## take_rate           3.1708e-03  3.4303e-03   0.9244   0.35531
## give_rate           1.8637e-02  1.3265e-02   1.4050   0.16004
## hd_rate:reb_rate   -4.6847e+01  9.7533e-01 -48.0322 < 2.2e-16 ***
## hd_rate:phys_rate -1.8026e+02  1.0019e+01 -17.9920 < 2.2e-16 ***
## hd_rate:take_rate -3.2542e+02  2.1708e+01 -14.9909 < 2.2e-16 ***
## hd_rate:give_rate -5.7993e+01  9.8285e+00  -5.9005 3.680e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    0.072433
## Residual Sum of Squares: 0.030644
## R-Squared:      0.57694
## Adj. R-Squared: 0.54686
## F-statistic: 3128.67 on 9 and 20648 DF, p-value: < 2.22e-16
```

2.6 Statistical Inference

Standard errors were adjusted for heteroscedasticity and within-player correlation using the HC1 robust variance estimator with clustering at the player level. This approach accounts for potential correlation of error terms within players across games while allowing for heteroscedastic residuals.

```r
#computes robust standards error that adjust for clustering by the player and run t-test
coeftest(fe_goals_model, vcov = vcovHC(fe_goals_model, type = "HC1", cluster = "group"))
```

```
##
## t test of coefficients:
##
##                       Estimate  Std. Error t value Pr(>|t|)
## hd_rate            -1.5518e-02  1.3153e-01 -0.1180  0.90608
## reb_rate            2.7941e+00  1.1962e+00  2.3358  0.01951 *
## phys_rate           2.7689e-02  1.7879e-02  1.5487  0.12148
## take_rate           3.1708e-03  4.0254e-03  0.7877  0.43088
## give_rate           1.8637e-02  1.6248e-02  1.1471  0.25137
## hd_rate:reb_rate   -4.6847e+01  3.8458e+01 -1.2182  0.22318
## hd_rate:phys_rate -1.8026e+02  1.1909e+02 -1.5137  0.13012
## hd_rate:take_rate -3.2542e+02  2.3096e+02 -1.4090  0.15885
## hd_rate:give_rate -5.7993e+01  8.8481e+01 -0.6554  0.51220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.7 Model Simplification

Following estimation of the full interactive model, a simplified specification was developed focusing on statistically significant relationships, retaining the theoretically important high-danger shot and rebound interaction while including secondary skills as direct effects.

```r
#simplified model focusing on significant relationships
fe_goals_simple <- plm(goals_rate ~ hd_rate * reb_rate + phys_rate + take_rate + give_rate, data = df_panel_game
s, model = "within")
summary(fe_goals_simple)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = goals_rate ~ hd_rate * reb_rate + phys_rate + take_rate +
##     give_rate, data = df_panel_games, model = "within")
##
## Unbalanced Panel: n = 1460, T = 2-25, N = 22117
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -1.8200e-02 -6.9390e-05  8.1712e-05  2.3740e-04  1.2508e-01
##
## Coefficients:
##                      Estimate  Std. Error  t-value  Pr(>|t|)
## hd_rate           -5.8885e-02  7.4299e-03  -7.9254 2.388e-15 ***
## reb_rate           2.6679e+00  1.7719e-02 150.5689 < 2.2e-16 ***
## phys_rate          1.4487e-02  4.7863e-03   3.0267  0.002475 **
## take_rate          2.2228e-04  3.4799e-03   0.0639  0.949070
## give_rate          2.6385e-03  1.3230e-02   0.1994  0.841922
## hd_rate:reb_rate  -4.0512e+01  9.5795e-01 -42.2909 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    0.072433
## Residual Sum of Squares: 0.031631
## R-Squared:      0.5633
## Adj. R-Squared: 0.53232
## F-statistic: 4439.68 on 6 and 20651 DF, p-value: < 2.22e-16
```

3. Results & Discussion

3.1 Correlation Analysis

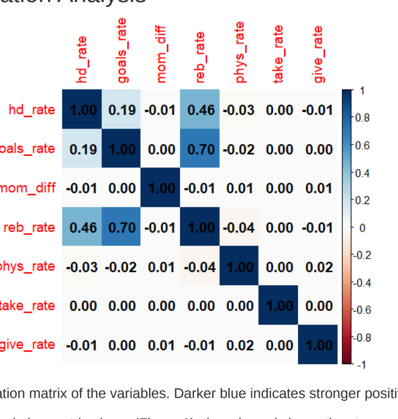


Figure 1. Correlation matrix of the variables. Darker blue indicates stronger positive relationships and red indicates negative relationships.

Based on the correlation matrix above (Figure 1), the rebound shows the strongest association with goal production ($r = 0.70$), while high-danger shots demonstrate a moderate correlation with goal production ($r = 0.19$). The two primary skills are moderately linked ($r = 0.46$), suggesting skilled shooters often generate rebounds. Furthermore, the secondary skills exhibit negligible correlations with goals, where hits, takeaways, and giveaways all show $|r| < 0.03$. These near-zero relationships indicate that physical engagement and puck management do not directly influence scoring output.

3.2 Fixed-effects Panel Regression Results

3.2.1 Full Interactive Model

```
Oneway (individual) effect Within Model

Call:
plm(formula = goals_rate ~ hd_rate * reb_rate + hd_rate * phys_rate +
    hd_rate * take_rate + hd_rate * give_rate, data = df_panel_games,
    model = "within")

Unbalanced Panel: n = 1460, T = 2-25, N = 22117

Residuals:
       Min.     1st Qu.      Median     3rd Qu.        Max.
-1.9073e-02 -7.8727e-05  8.0314e-05  2.3081e-04  1.1962e-01

Coefficients:
                     Estimate  Std. Error  t-value  Pr(>|t|)
hd_rate           -1.5518e-02  7.5151e-03  -2.0649   0.03894 *
reb_rate           2.7941e+00  1.8134e-02 154.0833 < 2.2e-16 ***
phys_rate          2.7689e-02  4.7661e-03   5.8096 6.354e-09 ***
take_rate          3.1708e-03  3.4303e-03   0.9244   0.35531
give_rate          1.8637e-02  1.3265e-02   1.4050   0.16004
hd_rate:reb_rate  -4.6847e+01  9.7533e-01 -48.0322 < 2.2e-16 ***
hd_rate:phys_rate -1.8026e+02  1.0019e+01 -17.9920 < 2.2e-16 ***
hd_rate:take_rate -3.2542e+02  2.1708e+01 -14.9909 < 2.2e-16 ***
hd_rate:give_rate -5.7993e+01  9.8285e+00  -5.9005 3.680e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    0.072433
Residual Sum of Squares: 0.030644
R-Squared:      0.57694
Adj. R-Squared: 0.54686
F-statistic: 3128.67 on 9 and 20648 DF, p-value: < 2.22e-16
```

Figure 2. Fixed-effects model with full interaction terms.

The full fixed-effects model provides definitive answers to both components of the research question (Figure 2). Regarding how high-danger shots and rebounds jointly influence goal production, the strongly negative interaction coefficient ($\beta = -46.847$, $p < 0.001$) reveals that these skills operate as substitutes rather than complements. Players achieve optimal goal production through specialization in one primary skill rather than maximizing both simultaneously, indicating distinct pathways to of

Addressing how secondary skills interact with this relationship, all interaction terms are statistically significant: physical engagement ($\beta = -180.26$, $p < 0.001$), takeaways ($\beta = -325.42$, $p < 0.001$), and giveaways ($\beta = -57.993$, $p < 0.001$). These universally negative coefficients reveal that secondary skills diminish rather than amplify high-danger shooting effectiveness, suggesting resource allocation trade-offs where focusing on secondary activities reduces high-danger shooting efficiency. The model explains 57.7% of within-player variation in goal production, with minimal difference from adjusted R^2 (0.547) indicating robust explanatory power.

3.2.2 Robust Standard Error Analysis

```
t test of coefficients:

                     Estimate  Std. Error t value Pr(>|t|)
hd_rate           -1.5518e-02  1.3153e-01 -0.1180  0.90608
reb_rate           2.7941e+00  1.1962e+00  2.3358  0.01951 *
phys_rate          2.7689e-02  1.7879e-02  1.5487  0.12148
take_rate          3.1708e-03  4.0254e-03  0.7877  0.43088
give_rate          1.8637e-02  1.6248e-02  1.1471  0.25137
hd_rate:reb_rate  -4.6847e+01  3.8458e+01 -1.2182  0.22318
hd_rate:phys_rate -1.8026e+02  1.1909e+02 -1.5137  0.13012
hd_rate:take_rate -3.2542e+02  2.3096e+02 -1.4090  0.15885
hd_rate:give_rate -5.7993e+01  8.8481e+01 -0.6554  0.51220
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3. Fixed-effects coefficients with robust standard errors.

The robust standard error analysis reveals important nuances in the reliability of these findings (Figure 3). While rebound rate maintains strong significance ($p < 0.02$), the high-danger shot and rebound interaction loses statistical significance ($p = 0.22$) under heteroscedasticity-adjusted standard errors. Similarly, all secondary skill interactions become non-significant when accounting for within-player clustering: physical engagement ($p = 0.12$), takeaways ($p = 0.43$), and giveaways ($p = 0.25$).

This pattern suggests that rebound generation represents the most reliable predictor of goal production, while the evidence for interactive effects is more fragile than initially indicated. The loss of interaction significance under robust standard errors indicates that conclusions about skill substitutability and secondary skill moderation should be interpreted with appropriate caution, highlighting the importance of statistical robustness in sports analytics.

3.2.3 Simplified Model

```
Oneway (individual) effect Within Model

Call:
plm(formula = goals_rate ~ hd_rate * reb_rate + phys_rate + take_rate +
    give_rate, data = df_panel_games, model = "within")

Unbalanced Panel: n = 1460, T = 2-25, N = 22117

Residuals:
       Min.     1st Qu.      Median     3rd Qu.        Max.
-1.8200e-02 -6.9390e-05  8.1712e-05  2.3740e-04  1.2508e-01

Coefficients:
                    Estimate  Std. Error  t-value  Pr(>|t|)
hd_rate          -5.8885e-02  7.4299e-03  -7.9254 2.388e-15 ***
reb_rate          2.6679e+00  1.7719e-02 150.5689 < 2.2e-16 ***
phys_rate         1.4487e-02  4.7863e-03   3.0267  0.002475 **
take_rate         2.2228e-04  3.4799e-03   0.0639  0.949070
give_rate         2.6385e-03  1.3230e-02   0.1994  0.841922
hd_rate:reb_rate -4.0512e+01  9.5795e-01 -42.2909 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    0.072433
Residual Sum of Squares: 0.031631
R-Squared:      0.5633
Adj. R-Squared: 0.53232
F-statistic: 4439.68 on 6 and 20651 DF, p-value: < 2.22e-16
```

Figure 4. Simplified fixed-effects model retaining only significant interaction.

Given the robust standard error results, a simplified model focusing on reliable relationships was estimated (Figure 4). The model retains the high-danger shot and rebound interaction ($\beta = -40.512$, $p < 0.001$) while treating secondary skills as direct effects. Rebound rate emerges as the dominant predictor ($\beta = 2.6679$, $p < 0.001$), while physical engagement shows a positive direct effect ($\beta = 0.01$, $p < 0.01$).

These findings fundamentally address the research question by revealing that goal production operates through a skill specialization framework rather than skill complementarity. The dominance of rebound generation over high-danger shooting challenges conventional assumptions about offensive skill hierarchy, while the fragility of interaction effects under robust standard errors suggests that individual skill excellence may be more important than skill combinations. This has significant implications for player evaluation and development, supporting focused skill specialization rather than attempting to maximize all abilities simultaneously.

4. Conclusion

4.1 Implications

These findings challenge conventional assumptions about skill complementarity in hockey analytics and support a skill specialization system for player evaluation. Teams should prioritize players who excel in rebound generation and consider that developing multiple offensive skills simultaneously may yield diminishing returns. The results suggest that effective player development strategies should focus on maximizing specific abilities rather than attempting to enhance all skills equally.

4.2 Limitations

The analysis focuses on individual-level performance without accounting for team context, linemate effects, or defensive systems that may moderate skill interactions. The loss of interaction significance under robust standard errors highlights the sensitivity of these relationships to statistical assumptions. Additionally, the study examines only offensive skills, excluding defensive contributions that may influence overall player value.

4.3 Future Research

Future studies should explore positional differences in skill complementarity, examine how team systems moderate individual skill effectiveness, and investigate the temporal dynamics of skill development. Research incorporating defensive context and team-level factors would provide a more comprehensive understanding of skill interactions in hockey performance.

Resources

1. EDGE.NHL.com. NHL edge puck and player tracking statistics - glossary [Internet]. [cited 2025 Jun 27]. Available from: https://edge.nhl.com/en/glossary
2. Cusimano MD, Nastis S, Zuccaro L. Effectiveness of interventions to reduce aggression and injuries among ice hockey players: A systematic review. Canadian Medical Association Journal. 2012 Dec 3;185(1). doi:10.1503/cmaj.112017
3. MoneyPuck.com. Player and Team Data [Internet]. [cited 2025 Jun 27]. Available from: https://moneypuck.com/data.htm