Valuation of NHL Draft Picks using Functional Data Analysis

by

Ryker Moreau

B.Sc., Simon Fraser University, 2020

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2022

Declaration of Committee

Name:	Ryker Moreau
Degree:	Master of Science
Project title:	Valuation of NHL Draft Picks using Functional Data Analysis
Committee:	Chair: Yi Lu Professor, Statistics and Actuarial Science Harsha Perera Supervisor Lecturer, Statistics and Actuarial Science Tim B. Swartz Committee Member Professor, Statistics and Actuarial Science Jiguo Cao Examiner Professor, Statistics and Actuarial Science

Abstract

Evaluation of player value in sport can be measured in several ways. These measures, when captured over an entire career, provide insights concerning player contributions. Professional sports teams select young talent through a draft process with the goal of acquiring a player that will provide maximum value, but these expectations diminish as the pool of players grows smaller. In this project, we develop valuation measures for draft picks in the National Hockey League (NHL) and analyze the value of each pick number with these measures. Specifically, we use different measures of player value to provide an expected value of that measure for each pick number in the draft. Our approach uses functional data analysis (FDA) to find a mean value curve from many observed functions in a nonparametric fashion. These functions are defined by each separate year of draft data. The resulting FDA model follows the assumption of monotonicity, ensuring that a smaller pick number always provides more expected value than any larger pick number. Based on a cross-validation approach, measuring value on annual salary provides the best predictive results. The proposed approach can be extended to sports in which an entry draft occurs and player career data are available.

Keywords: Functional data analysis; sports analytics; National Hockey League; entry draft; draft pick charts

Acknowledgements

First, I'd like to thank my family for all of their support through my journey at Simon Fraser University (SFU). During my time at SFU I was fortunate enough to follow my dreams of working in professional sports, and I couldn't have done it without the work ethic that you instilled in me. I'd also like to thank my fiancé, Jacey, for putting up with all the late nights and stressful deadlines. Your love and support helped push my across the finish line.

Thank you to my supervisors, Dr. Harsha Perera and Dr. Tim Swartz, for your guidance and generosity during my time as a M.Sc. student. Your passion for sports analytics has been a driving force in helping me get to where I am today. Your support through this project is greatly appreciated.

Lastly, I'd like to thank all the wonderful people I have met during my time at SFU. Thank you Brendan for inspiring me to take my visuals and presentations to the next level; Elijah for your endless humour, and never ending supply of interesting papers and articles; Robyn for your leadership during our Big Data Bowl championship; Kim for helping me get through more than a few homework assignments; and SFU alumni Dani, Matt, Lucas, and Peter for your mentorship and the foundation you have laid for SFU to become a powerhouse in sports analytics.

Table of Contents

List of Tables

draft, Mike McCoy of the Dallas Cowboys developed a chart that assigned perceived value to draft picks. The first draft pick was worth 3000 points, the second draft pick was worth 2600 points, and so on, in an exponentially decreasing order. With values assigned to draft picks, the Cowboys were able to trade picks to other teams, and accumulate value. Over the years, with better knowledge of the value of draft picks, the Cowboys had “fleeced” other teams, and set themselves up for a prolonged period of excellence that lasted for at least five years.

Since that time, many pick value charts have been created across professional sports. A sample of such contributions includes the National Basketball Association (Pelton 2017), Major League Baseball (Cacchione 2018), Major League Soccer (Swartz, Arce and Parameswaran 2013), and the National Football League (Massey and Thaler 2013, Schuckers 2011a).

In the NHL, there have been alternative constructions of pick value charts. Schuckers (2011b) used nonparametric regression on career games-played as a measure of player value. A review article on various issues associated with drafting in the National Hockey League was provided by Tingling (2017).

A major challenge in the construction of pick value charts is the determination of player value. Player value is an ambiguous term as it has different interpretations over the short

In Chapter 2, we introduce the data where four metrics are proposed to assess player value. Importantly, we take into account the importance of considering performance metrics. The FDA model is described in Chapter 3 where its advantages are discussed with respect to pick value charts. The resultant FDA pick value charts are presented in Chapter 4 where

Chapter 2

Background

2.1 Data

In this project, we use three data sources. Sports Reference LLC (2022) data provides us with draft data from 1982-2016. This data are the full set of players that will be included in our analysis, even if players did not play in the NHL. The total number of players in this data is 8,613.

Vollman (2018) data provides us with player statistics from 1982-2006, including games-played (GP), goals (G), and assists (A), and advanced metrics such as point-share (PS). This data contains the players who played at least one game in the NHL. Of the drafted players, we have player statistics for 2,803 players, including 72 players who were still active in the 2021-2022 season. Although more years of player statistics are available, we want to introduce player performance measures based on career contributions. Therefore, we truncate this data at 2006, allowing most players to complete their playing careers by 2022.

Lastly, Sportrac (2022) data provides contract information from 2001-2022 for players from the 2001-2016 drafts. This data includes the contract length, total value of the contract, and annual average value (AV) of the contract. We calculate the percentile of each player's AV for all years under contract to normalize data in an effort to deal with inflation. In total, data was available for 1,144 players.

Data were cleaned as to match names across data sets. This included removing foreign letters, using full first names, and dealing with players with the same names to ensure consistency. A verification process was done to ensure all players were matched properly and no players were missed.

The NHL entry draft has evolved over the years as the number of teams in the league has increased. From 1982 to 1991 the draft consisted of 12-rounds, then fell to 9-rounds in 1995 and 7-rounds in 2005 which is now the present day amount. The current state of the NHL draft in 2022 involves 32 teams selecting for 7-rounds, for a total of 224 picks. There are rare cases of compensatory picks being awarded or picks being taken from teams due to violations of league rules, but those do not impact our analysis. To obtain a valuation of NHL draft picks, we truncate our data for past drafts to include only the first 224 picks. This applies to all drafts except 2006-2016 where less than 224 players were selected.

For our analysis, we only require a single explanatory variable, consistent with the proposed FDA framework. This variable will be the overall pick of the draft. Overall pick consists of the natural numbers $1, 2, ..., n$, where $n = 224$ is the total number of picks.

2.2 Measures of Player Value

This project considers valuations of NHL draft picks using player statistics and salary information. A valuation will be obtained through four measures of player value. Measuring a player's value is a difficult task as there are many ways to measure the quality of a career. A logical way is to use total games-played, as done in Schuckers (2011b), as this describes the longevity of a player.

Another way to view the problem is to use performance metrics as measures of player value. An advanced statistic, point-share, is what we will focus on. For this metric, a player's value will be measured by both career average, and career total.

Although statistics are unbiased measures of assessment, they are often confounded by other factors including the contribution of teammates. The use of salary data is another approach for player evaluation where we take the point of view that team executives (who determine salary) are knowledgeable about player value. This approach was utilized by Swartz, Arce, and Parameswaran (2013) in the context of soccer.

2.2.1 Games-Played (Y_1)

The first measure of player value, Y_1, is gTf0 1.637 Td4alue, ers play very few games, and only a small number of players reach the 1000-game milestone.

Figure 2.1: Histogram of games-played by all drafted players (left) and drafted players who played at least one NHL game (right).

2.2.2 Average Point Share (Y_2)

The second measure of player value, Y_2, is average-point-share (APS). Point-share is an advanced statistic based on a similar baseball statistic called win-share created by Bill James. Win-share is calculated for each player based on their contribution (or lack thereof) to a win. NHL teams are awarded points for each game: two for a win, one for an overtime loss, and zero for a regulation loss. Therefore, APS is calculated to closely match the team's point total for a season. The APS for goalies is not calculated in our dataset and therefore we remove goalies for this valuation method. For drafted players that did not play in the NHL, we set their APS value to the minimum APS in the dataset. The calculations for APS are complicated and will not be discussed fully in this project. The full details are provided in Kubatko (2010).

A benefit of APS is that it uses both offensive and defensive contributions, something that most statistics don't consider. Since the goal of the NHL is to win games, evaluating a player's contribution to these wins is a reasonable measure of player value. However, measuring player value through team success may not fully capture a player's individual value. In many cases there are high level players on low level teams that are undervalued in this metric. For example, during the 2007-08 season Marian Hossa was traded from the Atlanta Thrashers to the Pittsburgh Penguins. Hossa, who had 66 points in 72 games that season, went from the 28th ranked team to the 4th ranked team, increasing his PS even though his contribution was similar for both teams.

In addition, APS measures a player's excellence, without examining their longevity. This will increase the value of players who had their career cut short by injury or other factors.

Figure 2.2 provides histograms of $Y_2 = APS = \frac{1}{S}\sum_s APS_s$, where S is the total number of seasons a player played in the NHL, for all drafted players in our dataset and APS_s is the player's APS in season s. The histogram on the left includes every drafted player, so we see the majority of our data at the minimum value due to the large amount of drafted players who never made it to the NHL. On the right shows Y_2 for all players who played at least one game in the NHL, which still has the majority of players near zero. The exponential shape shows that elite talent is rare and most drafted players will have minimal contribution in comparison.

2.2.3 Total Point Share (Y_3)

The third measure of player value, Y_3, is total-point-share (TPS). A description of APS has been provided in Chapter 2.2.2. The difference between TPS and APS is that TPS accounts for longevity and excellence. This measure benefits players with high APS, who had more games-played. This not only allows us to evaluate player value in different ways, but also compare how we view PS. The TPS for goalies is not calculated in our dataset and therefore

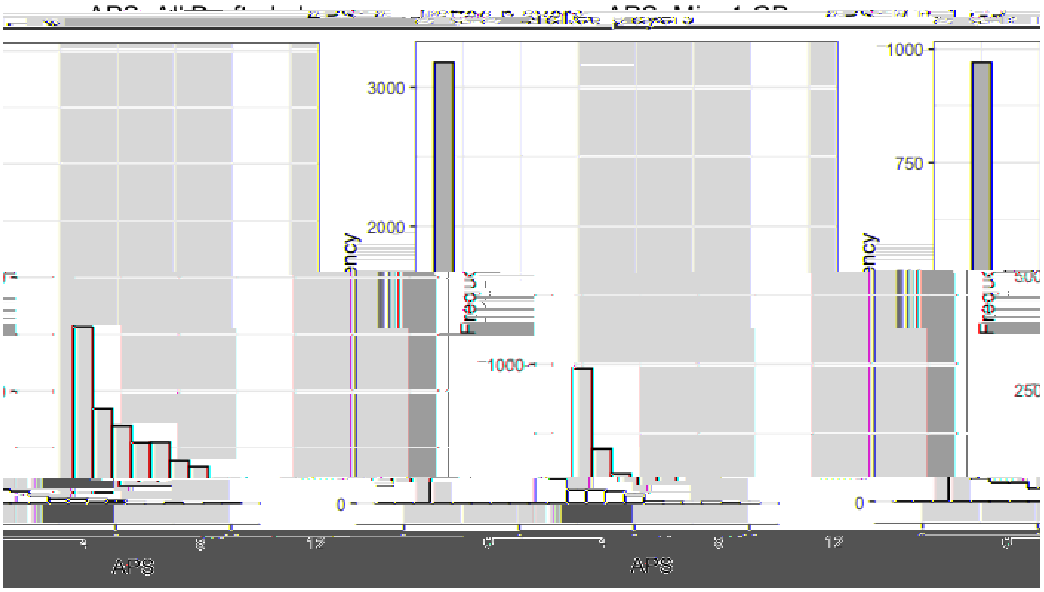


Figure 2.2: Histogram of average-point-share by all drafted players (left) and drafted players who played at least one NHL game (right).

butions of APS. Most of the players in the dataset are around zero, with a small number of elite players with high TPS.

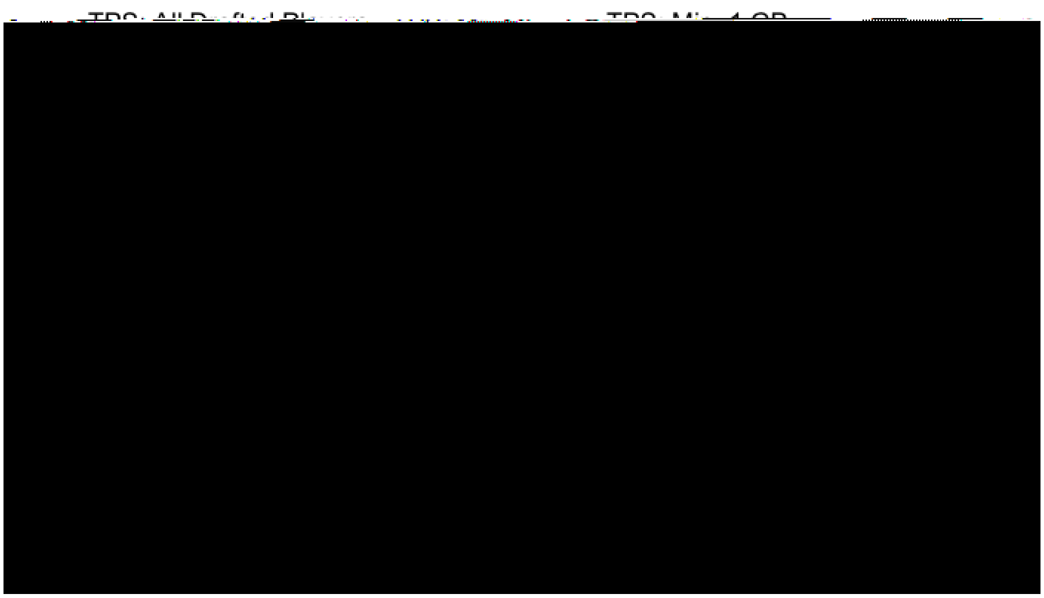


Figure 2.3: Histogram of total-point-share by all drafted players (left) and drafted players who played at least one NHL game (right).

2.2.4 Standardized Salary Ranking (Y_4)

The fourth measure of player value, Y_4, is standardized-salary-ranking. We have specified this ranking as the percentile of the annual average value (AV) of contracts for each player within each year. We will denote this metric as PAV. By using the percentiles instead of raw values of AV we standardize the data since salaries will change over time with inflation.

This method of player value gives a different perspective than performance measures

where the $_{ij}$ terms are assumed independent and arise from a normal distribution with zero mean and constant but unknown variance.

In this application, there are some immediate questions related to the assumptions of the linear regression model (3.1). First, the linearity assumption relating player value and draft position is highly questionable. In fact, all previous work has indicated that the value of draft picks has an exponential shape which tends to flatten towards the end of the draft. Second, the distributional assumptions associated with $_{ij}$ appear strong. For example, one might expect skewness in $_{ij}$ for players selected late in the draft. The reason for this is that they either do not produce (zero value as expected for a late round pick) or they surprise positively. Third, model (3.1) is a regression of points and does not take into account the model is written as

$$y_{ij} = \alpha_i + \beta_i \int_1^j exp[z_i(u)]du + \epsilon_{ij} \tag{3.3}$$

3.2.2 Basis Functions

As stated in Chapter 3.1, the linearity assumption relating to player value is highly questionable. In our proposed model we address this issue and introduce nonlinearity through the use of basis-functions. Basis-functions, denoted as φ_k where $k = 1, ..., K$, are functional building blocks that allow a curve to act differently between break-points. These basis-functions make up each $z_i(u)$ in model (3.3), and are defined by basis-function-expansion in the form $z_i(u) = \sum_{k=1}^{K} c_{ik}\varphi_k(u)$, where the c_{ik} are expansion coefficients. Inserting our basis-functions into model (3.3) results in

$$y_{ij} = \alpha_i + \beta_i \int_1^j exp\left[\sum_{k=1}^{K} c_{ik}\varphi_k(u)\right] du + \epsilon_{ij} \tag{3.4}$$

Our proposed model uses B-spline basis-functions. B-splines are piece-wise polynomials that are flexible. To produce a B-spline basis, we need to specify knots and break-points in our functions. We will only use a single knot at each break point. As the number of knots grows, so does the variability of the resulting curves. Our goal is to find the optimal amount of knots that produces a smooth curve without losing information about variability. For our model, we will use an order four cubic-spline basis with a single knot at each break point. This method ensures all neighboring splines will have matching first and second derivatives, resulting in a smooth continuous curve.

The number of knots specified determines the number of K basis-functions. Too many basis-functions will result in over-fitting, and a high measurement error, whereas too few basis-functions will fail to capture the features of the curve. This is also know as bias-variance trade-off. Typically an optimal K is chosen by minimizing mean-squared-error (MSE), which is common way to deal with the bias-variance trade-off. However, since we want the resulting mean curve to be smooth, we use fewer knots than optimal.

3.2.3 Smoothing

Now that we have specified our model (3.4), we can define the smoothing technique. Our linear smoother is obtained by determining the coe□cients of the basis-expansion, c_{ik}. We estimate these coe□cients with a penalized weighted least squares criterion (PENSSE). This criterion includes a smoothing parameter, λ, where as λ increases, the roughness penalty is larger and y_{ij} becomes more linear and as λ decreases, the penalty is reduced and y_{ij} fits the data better. The details of the PENSSE criterion are provided in Ramsay and Silverman (2002).

3.2.4 Final Model

Estimates for $\hat{\alpha}_i$, $\hat{\beta}_i$, and $\hat{c}_{ik}$ for each $\hat{y}_{ij}$ from model (3.4) are computed with the R package (Ramsay, Hooker, and Graves 2022). These estimates are derived through an iterative process that minimize the PENSSE as described in Chapter 3.2.3. After determining $\hat{y}_{ij}$ for all draft years, i, the resulting mean curve for draft position j is defined as

$$\hat{\mu}_j = \frac{1}{M}\sum_i \hat{y}_{ij} = \frac{1}{M}\sum_i \left[\hat{\alpha}_i + \hat{\beta}_i \int_1^j exp[\hat{c}_{ik}\varphi_k(u)]du\right] \tag{3.5}$$

where M is the total number of draft years, or functions, in the analysis. Model (3.4) is perhaps one of the most basic models that has been developed in FDA, but is of great practical importance.

Chapter 4

Results

We now explore the results of our proposed FDA model for the four measures of player value described in Chapter 2.2.1-2.2.4. Computing was done in R and followed the methods provided in Ramsay et al. (2009).

4.1 Pick Value Charts

We define each draft year, i, as a function of the overall pick in the draft, $j = 1, ..., 224$. As discussed in Chapter 2.1 we truncate any draft with more than 224-picks to represent the current state of the NHL draft. For each measure of player value we will show the mean curve and 95% pointwise confidence interval. The inclusion of confidence intervals have not been included in other pick value charts but we believe it is a valuable addition to our charts. We then validate our model with a cross-validation method and compare our results.

4.1.1 Games-Played (Y_1)

Figure 4.1 provides the predicted curves, $\hat{y}_{ij}$, for draft years $i = 1982, 1983, ..., 2006$ and the mean curve, $\hat{\mu}_j$, for the first measure of player value GP (Y_1). The mean curve represents the predicted number of GP we would expect from a player drafted at pick j. For instance, we predict that a player drafted in 1st-overall should play 866 games in the NHL, compared to a player drafted 33rd-overall (1st pick of the 2nd-round) who we predict should play 252 games. This sharp decline in value is apparent for the first-round (picks 1-32), followed by the curve flattening out with a much slower decrease.

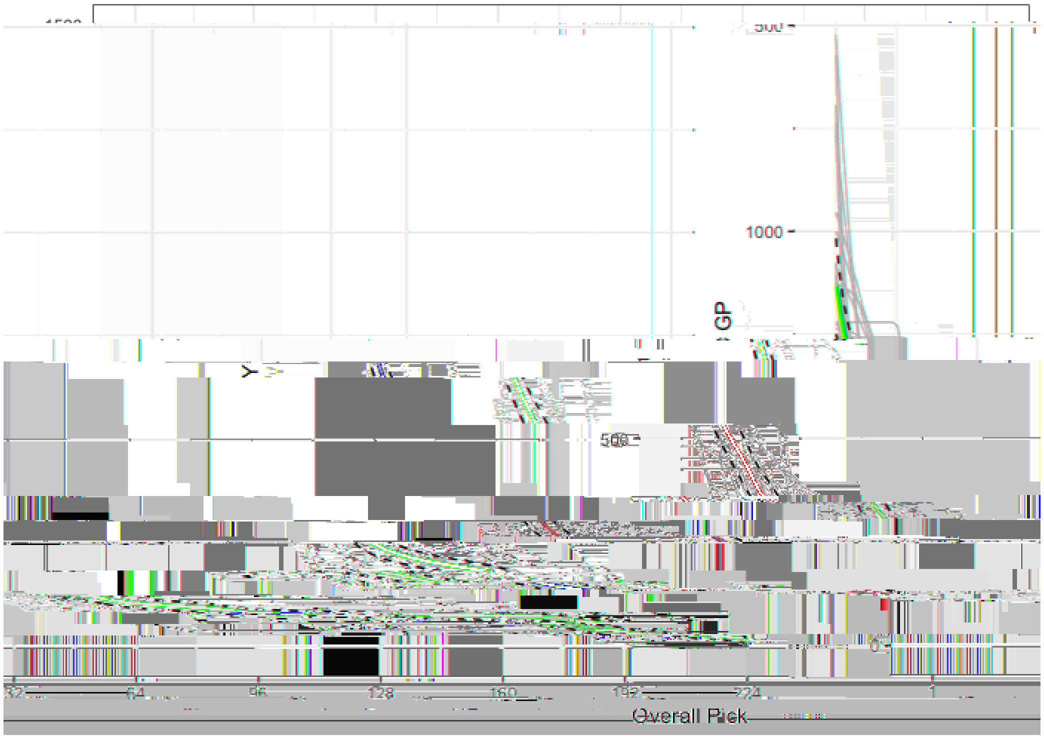


Figure 4.1: Plot of the measure of player value, GP. Shown are predicted curves, $\hat{y}_{ij}$, for draft years 1982-2006 (grey), the resulting mean curve, $\hat{\mu}_j$ (red), and pointwise confidence

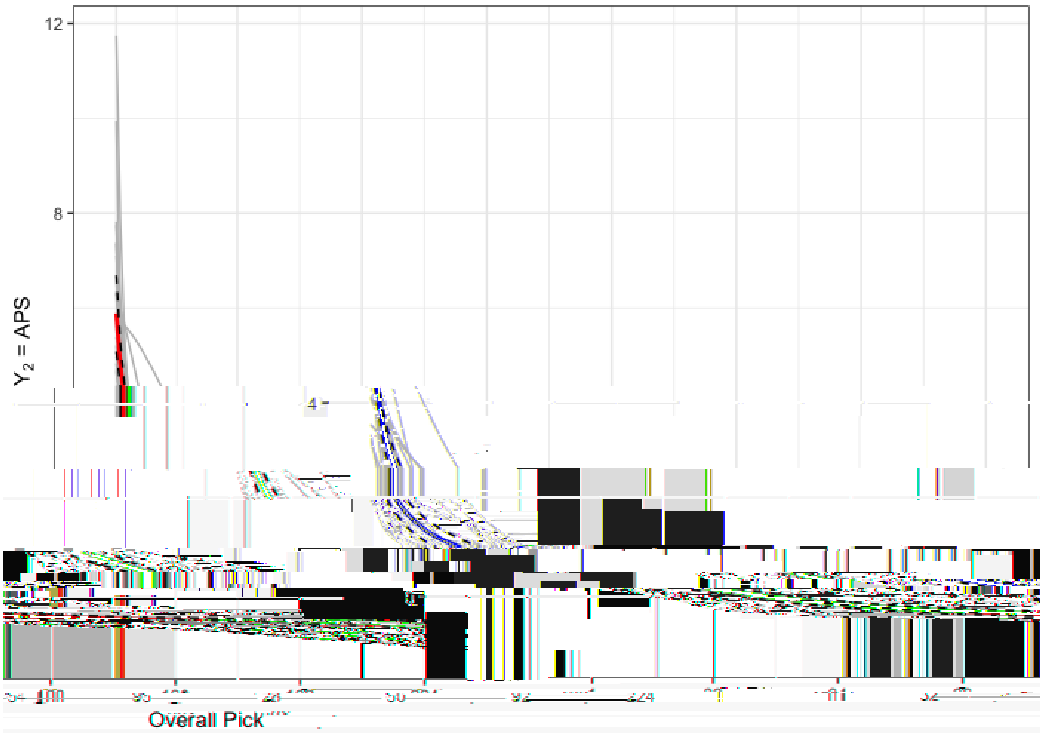


Figure 4.2: Plot of the measure of player value, APS. Shown are predicted curves, $\hat{y}_{ij}$, for draft years 1982-2006 (grey), the resulting mean curve, $\hat{\mu}_j$ (red), and pointwise confidence

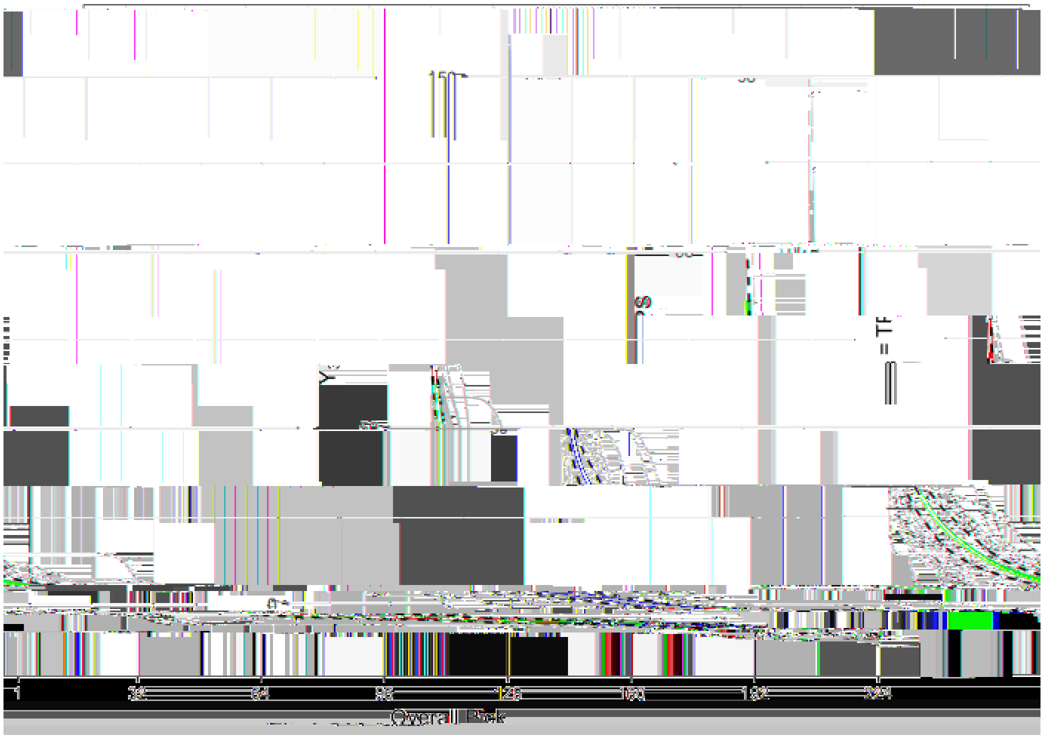


Figure 4.3: Plot of the measure of player value, TPS. Shown are predicted curves, $\hat{y}_{ij}$, for draft years 1982-2006 (grey), the resulting mean curve, $\hat{\mu}_j$ (red), and pointwise confidence interval (black) for overall picks 1-224.

TPS than the player with fewer games.

4.1.4 Standardized Salary Ranking (Y_4)

Figure 4.4 provides the predicted curves, $\hat{y}_{ij}$, for draft years $i = 2001, 2002, ..., 2016$ and the mean curve, $\hat{\mu}_j$, for the third measure of player value PAV (Y_4). The mean curve represents the predicted percentile of AV that a player drafted at pick j should average over their career. We predict that the 1st overall pick should average being in the 82nd percentile of AV over their career, and players drafted with the 10th pick or later are predicted to be in the 50th percentile or lower.

The 95% pointwise confidence interval in Figure 4.4 shows relatively high variation up to pick 100, and very high variation for early picks.

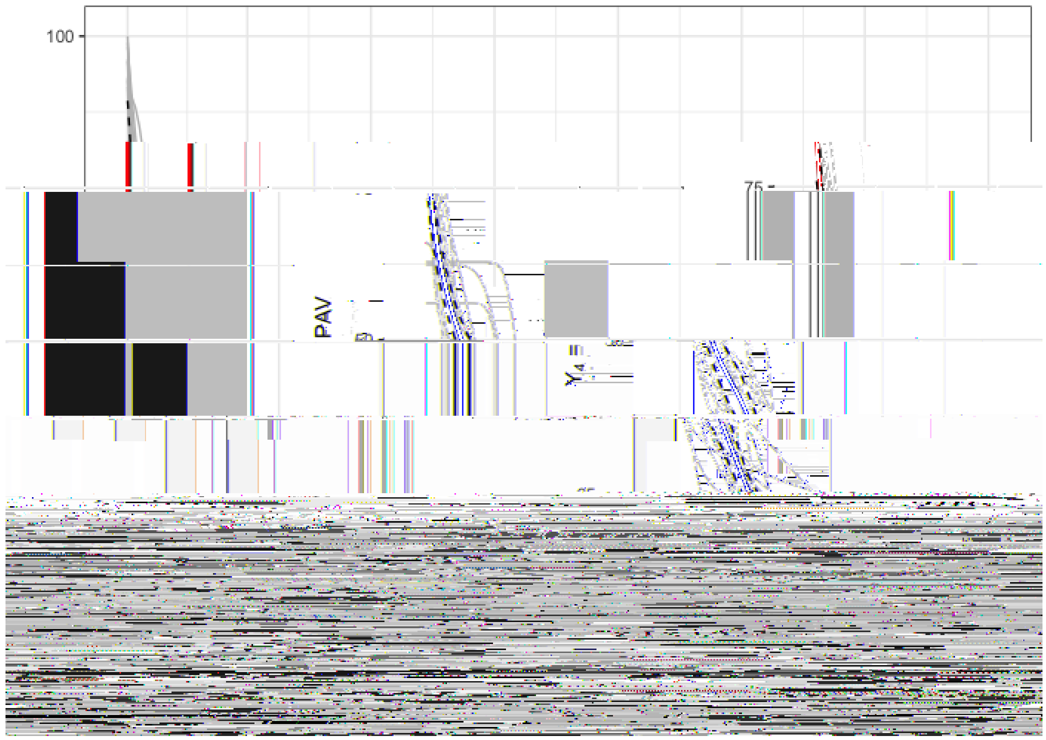


Figure 4.4: Plot of the measure of player value, PAV. Shown are predicted curves, $\hat{y}_{ij}$, for draft years 2001-2016 (grey), the resulting mean curve, $\hat{\mu}_j$ (red), and pointwise confidence interval (black) for overall picks 1-224.

4.2 Comparing Value Charts

To compare the four measures of player value, we can visualize the corresponding mean curves on one plot. Since all the measures of player value are in different scales, we first ference to make conclusions.

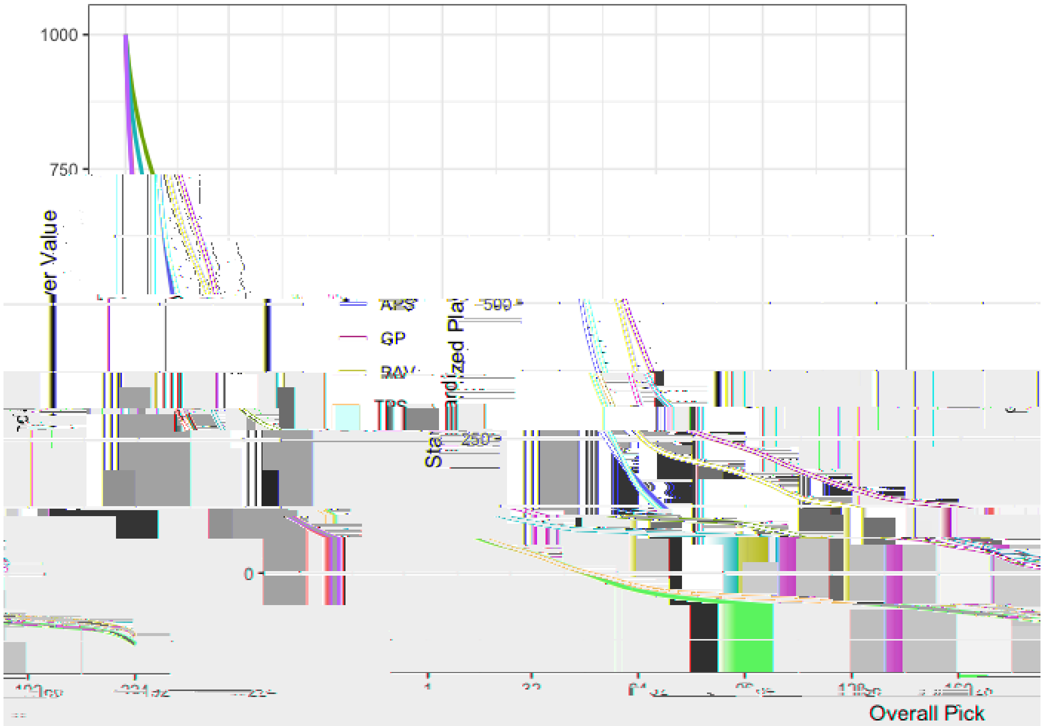


Figure 4.5: Plot of mean curves from all measures of player value, Y_1, Y_2, Y_3, Y_4. Values are standardized so that the value for the 1st-overall pick $Y_{i1} = 1000$.

4.3 Validation

We validate our model with a leave-one-out-cross-validation (LOOCV) for each measure of player value. This method involves taking one draft year out as a test year and completing the analysis, then repeating the process for all draft years in the dataset. We calculate the root-mean-square-error (RMSE) for each test year with the corresponding predicted values and take the average RMSE from all iterations. The RMSE for year i is defined as $\frac{1}{n_j}\sum_{j=1}^{n_j}(y_{ij} - \hat{y}_{ij})^2$. The standardized values described in Chapter 4.2 are used for this process so the RMSE for each measure has the same scale. Table 4.1 contains the resulting RMSE from LOOCV. Our results show that PAV performed best while GP was the worst in terms of predictive error. The high RMSE for GP may be due to the issues described in Chapter 2.2.1 about injuries, and that many players who do not have a high APS, TPS, or PAV may still be able to have a long career if they are valuable in other aspects. The low RMSE for PAV is likely because this measure is not affected by injury or decrease in production during a contract. Top draft picks that sign big contracts who do not meet expectations are not penalized in this measure until they sign a new contract.

Measure of Player Value (Y)	RMSE
GP	306.17
APS	258.13
TPS	238.91
PAV	226.38

Table 4.1: Root-mean-square-error (RMSE) results from the leave-one-out-cross-validation (LOOCV) performed on all measures of player value GP, APS, TPS, and PAV.

We also investigate our model by comparing the most widely used measure of player value, GP, to the results from Schuckers (2011b), which includes predictions for the number of career GP we expect for draft picks from $j = 1, 2, ..., 210$. Figure 4.6 shows the mean curves for each method.

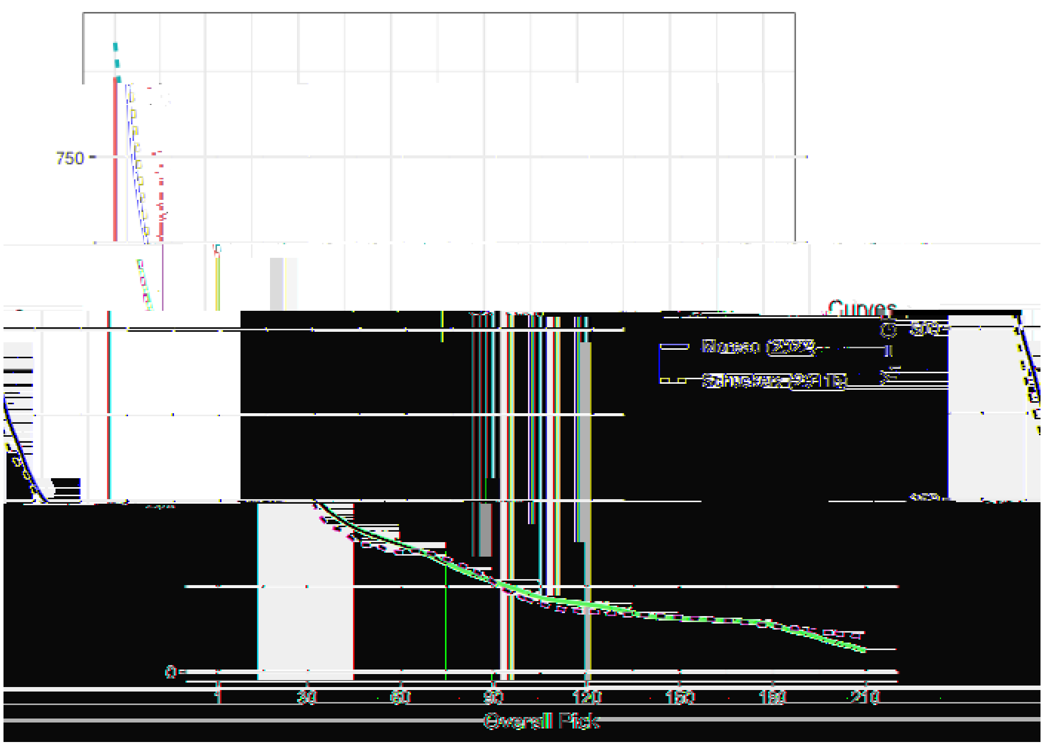


Figure 4.6: Comparison of predicted mean curves for GP from our derived model (3.5) and Schuckers (2011b) for draft picks 1-210.

We see the curve from Schuckers (2011b) is closely related to the one derived in this project. Schuckers curve is not strictly monotonically decreasing, which was stressed as an important property of our proposed model. Besides that, it does not seem that FDA has provided much of a difference.

Chapter 5

Discussion

In this project we develop valuation measures for draft picks in the NHL entry draft and analyze the value of each pick with these measures. We use Functional Data Analysis (FDA) to find a mean value curve from many observed functions using a nonparametric approach. This approach allows us to use a functional linear regression framework that introduces nonlinearity to account for the differences of players across draft years. The four measures of player value used are games-played (GP), average-point-share (APS), total-point-share (TPS), and a standardized salary ranking (PAV). The resulting mean value curve for each measure is the predicted value of a player drafted at a certain overall draft pick.

There are limitations to the proposed methods. Limitations of the measures of player value differ by measure, but include accounting for injuries, draft pick bias, and confounding variables such as teammates. Also, for GP, APS, and TPS, we are limited to using data only up until 2006. Surely the NHL has developed and changed over time, so not being able to use recent data may limit our accuracy when comparing our results to the present day. There is also a lack of reported salary data before 2001, and including more data may be beneficial to our results.

Although comparing many measures of player value is useful to determine differences, having a mean value curve from an all encompassing metric would be very useful. Developing one metric from a combination of many metrics and then using our proposed model would be an area of future work which we would like to explore.

Our draft value charts can be useful tools for NHL teams. They can be used as a trade tool, to determine a fair value when trading draft picks with another team. They can also be used as a salary-cap tool, to predict the contract value of a drafted player years in advance of signing.

Bibliography

Ramsay, J.O. and Silverman, B.W. (2005). *A* *, 2* , Springer Series in Statistics, Springer: New York.

Schuckers, M. (2011a). An alternative to the NFL draft pick value chart based upon player performance. *A* , 7(2), Article 10.

Schuckers, M. (2011b). What's an NHL draft pick worth? A value pick chart for the or y005 (uc)28. (h)

Appendix A

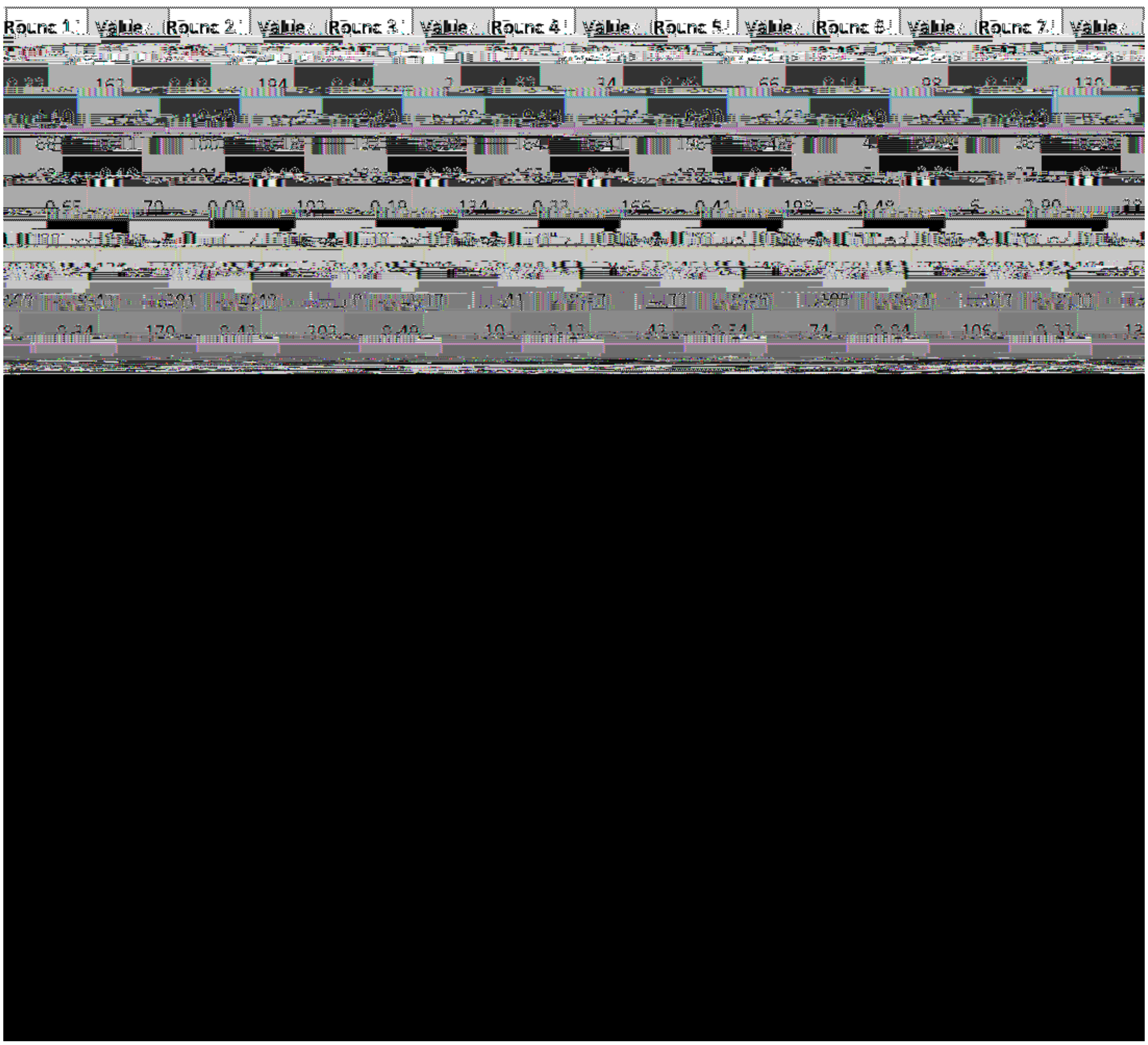
Pick Value Chart: GP (Y_1)

Round 1	Value	Round 2	Value	Round 3	Value	Round 4	Value	Round 5	Value	Round 6	Value	Round 7	Value
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
751	35	242	67	174	39	118	131	90	163	76	135	50	3
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

[illegible]

Appendix B

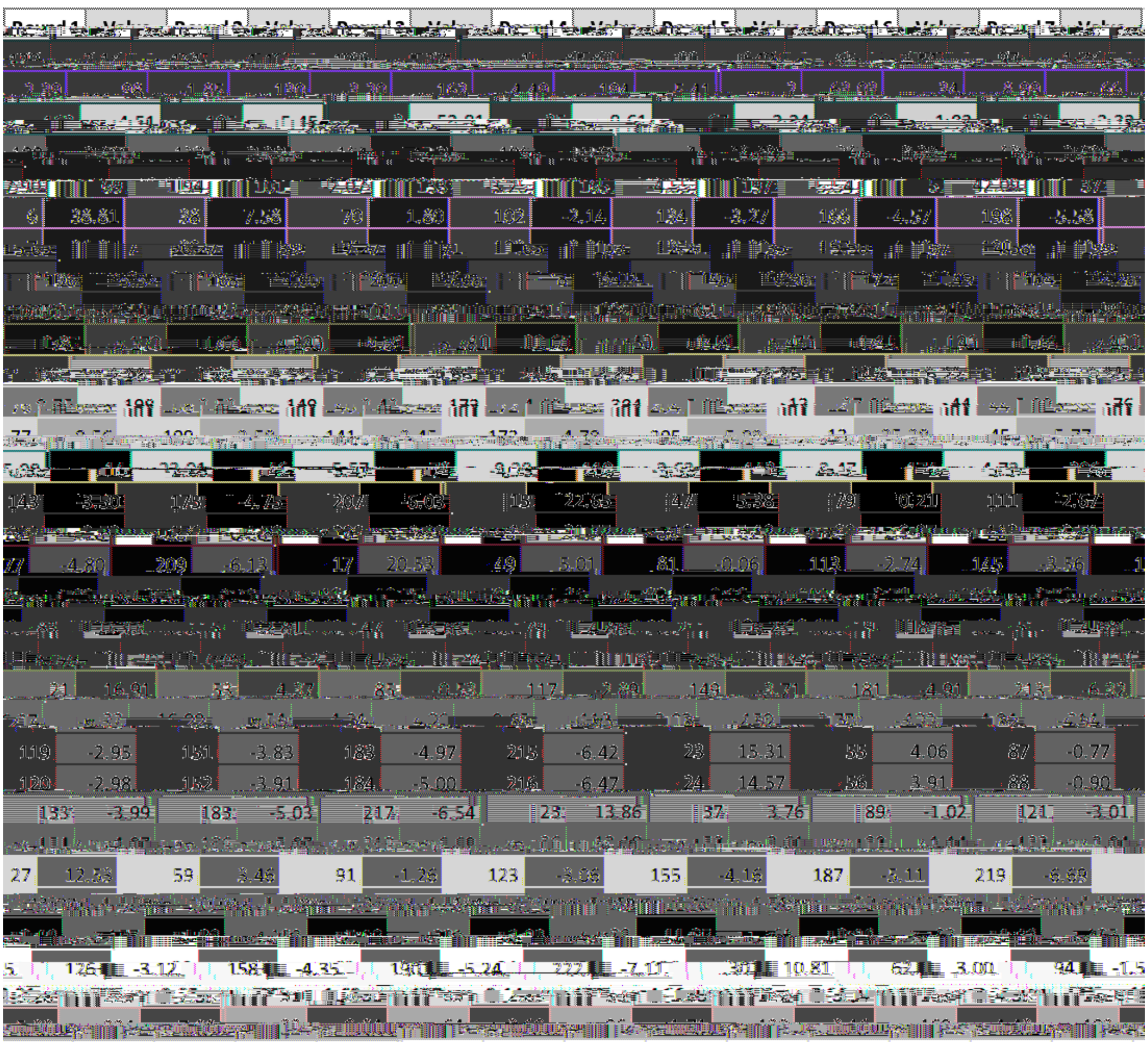
Pick Value Chart: APS (Y_2)



Round 1	Value	Round 2	Value	Round 3	Value	Round 4	Value	Round 5	Value	Round 6	Value	Round 7	Value
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Appendix C

Pick Value Chart: TPS (Y_3)



Appendix D

Pick Value Chart: PAV (Y_4)

Round 1	Value	Round 2	Value	Round 3	Value	Round 4	Value	Round 5	Value	Round 6	Value	Round 7	Value
[illegible]	17.3	73	12.4	105	8.3	137	6.1	169	3.4	201	3.2	9	31.8