

# Predicting the NHL Draft with Rank-Ordered Logit Models

Brendan Kumagai, Ryker Moreau, and Kimberly Kroetch

Simon Fraser University, Burnaby BC, Canada

brendan\_kumagai@sfu.ca, ryker\_moreau@sfu.ca, kimberly\_kroetch@sfu.ca

**Abstract.** The National Hockey League Entry Draft has been an active area of research in hockey analytics over the past decade. Prior research has explored predictive modelling for draft results using player information and statistics as well as ranking data from draft experts. In this paper, we develop a new modelling framework for this problem using a Bayesian rank-ordered logit model based on draft ranking data obtained from scouting sites and media outlets. Rank-ordered logit models are designed to model multicompetitor contests such as triathlons, sprints, or golf through a sequence of conditionally dependent multinomial logit models. We apply this model to a set of draft ranking data from the 2021 NHL draft and use it to provide a consolidated ranking for the draft and estimate the probability that any given player will be selected at any given pick.

## 1 Background and Motivation

Over the past two decades, the National Hockey League (NHL) has imposed a hard salary cap to limit player salaries and control a team’s ability to retain and add talented players in an effort to enforce competitive balance throughout the league. This has forced teams to become increasingly savvy in how they allocate resources. The NHL has three main outlets where a team can add, lose or maintain talent: free agency, trades, and the entry draft. Acquiring players through free agency or trades can often be an expensive endeavour costing valuable cap dollars or assets. On the other hand, the draft is a low-risk, high-reward way to find and develop NHL-level talent.

Every NHL team employs a department of scouts to identify and evaluate the top draft-eligible players throughout the season and inform the team’s draft selections each year. To strategize and obtain the players they desire, teams make assumptions on how long a player will last before being selected in the draft. Previous research has explored predictive modelling approaches for the outcome of the entry draft in both hockey [1] and other sports [2,3].

In this paper, we take a new approach to this problem by building a rank-ordered logit (ROL) model to estimate the probability that any given draft-eligible player will be selected at any given pick in the NHL draft. ROL models are typically used in sports that involve multicompetitor contests such as sprinting, triathlons, or golf. Primarily, our work was inspired by a discussion with Tyrel Stokes on this topic and his work with ROL models in the 100m dash [4].

In multicompetitor sports, there are generally dozens of major events per year that can be used to fit the ROL model and predict the outcome of future events. However, the NHL draft only occurs once a year and has a completely different crop of players each year. To address this issue we scrape draft rankings from various draft experts that provide ranking lists on scouting sites (i.e., Elite Prospects, Dobber Prospects, etc.) and media outlets (i.e., TSN, Sportsnet). We will refer to these media outlets and scouting sites hereinafter as ‘agencies’. Additionally, we will refer to each ranking list from an agency hereinafter as a ‘ranking set’. These ranking sets from various agencies are used as input into our model.

## 2 Methods

### 2.1 Multinomial Logit Models

We begin with a brief review of multinomial logit models. A multinomial logit (MNL) model is a method used in statistics to classify observations into one of two or more discrete outcome categories.

In particular, we are concerned with a special case of the MNL where we consider one trial (draft pick) being taken from  $c$  categories (available draft-eligible players). The goal of this model is to predict probabilities that each player is selected with a particular draft pick. In other words, we wish to estimate probabilities,  $[\pi_1, \pi_2, \dots, \pi_c]$

such that  $\pi_k$  is the probability of player  $k$  being selected with the draft pick of interest out of the  $c$  available draft-eligible players.

In the MNL model, these probabilities are derived as

$$\pi_k = \frac{\exp(\theta_k)}{\sum_{j=1}^c \exp(\theta_j)} \quad (1)$$

where  $\theta_k$  is an ‘ability’ parameter for player  $k$  that we wish to estimate by fitting this model [5].

As an example, suppose we wish to model the outcome of the 1st overall pick in the 2021 NHL draft given draft rankings from various agencies. By using the 1st overall ranked player from these ranking sets, we can estimate the values of  $\theta_k$  for all available players  $k = 1, \dots, c$ , and consequently, obtain estimates for the probability that player  $k$  is selected 1st overall,  $\pi_k$  from (1), for all available players.

### 2.2 Rank-Ordered Logit Model

The MNL provides us with a simple framework for estimating the probability that a player is selected with the first pick in the draft, but there are still questions that this model cannot answer alone such as: What is the probability of a player being drafted 2nd, 3rd or beyond? How would these probabilities differ depending on which player(s) were selected prior? If a player is consistently ranked top 5 but is never ranked 1st, would his probability,  $\pi_k$ , of being selected 1st be the same as a player rarely ranked in the top 200?

These questions can be addressed using a rank-ordered logit model. A ROL model can be thought of as a series of conditional multinomial logit models where the 1st overall pick is modelled as a MNL model with a single pick from the pool of all draft-eligible players, then the 2nd pick is modelled as a MNL model with a single pick from all draft-eligible players excluding the player selected 1st, and so on until the  $n$ th player, who is modelled using the MNL model with a single pick from all draft-eligible players excluding the  $n - 1$  players that have already been selected.

To define this model, let  $\theta_i$  be the underlying ability parameter for player  $i$  and let  $Y_i$  be the latent evaluation of player  $i$ ’s ability by the agency that developed the ranking set.

A key assumption in this model is that the latent evaluation by the agency is a realization from a Gumbel distribution with a location parameter of  $\theta_i$  and a scale parameter of 1. That is,  $Y_i|\theta_i \sim \text{Gumbel}(\theta_i, 1)$  [6]. If we let the true performance  $Y_i$  equal  $\theta_i + \epsilon_i$ , where  $\epsilon_i$  is an error term, then this assumption is equivalent to assuming that the distribution of the error is Gumbel with  $\mu = 0, \beta = 1$  where  $\mu$  and  $\beta$  are the location and scale parameters of the Gumbel distribution, respectively. The convenience of this assumption is made clear by Luce and Suppes [7], who show that a Gumbel

assumption of the errors implies a logit formula for the choice probabilities; furthermore a logit formula for the choice probabilities implies a Gumbel distribution for the errors [8]. In practice, this assumption is almost identical to an assumption of independent, normal errors, although extreme value distributions have fatter tails [9]. This assumption allows us to define the likelihood for a single draft ranking set in this model as

$$P(Y_1 > Y_2 > \dots > Y_n \mid \theta_1, \dots, \theta_n) = \prod_{i=1}^{n-1} \frac{\exp(\theta_i)}{\sum_{j=i}^n \exp(\theta_j)} \quad (2)$$

For example, consider a ranking set by TSN. Suppose TSN ranks Shane Wright 1st, Logan Cooley 2nd, and Juraj Slafkovsky 3rd, and  $\theta_1, \theta_2$ , and  $\theta_3$  correspond to Wright, Cooley and Slafkovsky's underlying abilities, respectively. This implies that  $Y_1, Y_2$ , and  $Y_3$  correspond to the TSN evaluation of Wright, Cooley and Slafkovsky's abilities, respectively, where  $Y_1 > Y_2 > Y_3$ .

We do not observe these scores directly from any ranking sets. However, we operate under the assumption that some sort of rating scale exists for each ranking set. To add some intuition behind the latent  $Y_i$ 's, imagine that the scouting team at Elite Prospects gets together and collaboratively comes up with a player grading scheme with scores ranging from 0-100. They may have scored Wright as 93/100, Cooley as 89/100, Slafkovsky as 88/100, and everyone else as 86/100 or below.

### 2.3 Accounting for Unranked Players

We can improve on the basic rank-ordered logit model specified in Section 2.2 by accounting for unranked players in our model likelihood.

Consider two ranking sets. In ranking set A there are 32 players ranked; Aatu Rätty is ranked 8th while Fyodor Svechkov is ranked 20th. In ranking set B there are also 32 players ranked; Aatu Rätty is not ranked in the top 32 while Fyodor Svechkov is ranked 22nd.

When we attempt to fit this model and estimate the  $\theta_i$ 's, the likelihood from the base ROL model as defined in Section 2.2 will take into account that Rätty ranked 8th in set A but will not penalize Rätty for being unranked all together in set B. On the other hand, the likelihood will take into account that Svechkov was ranked 20th and 22nd in sets A and B, respectively.

This example highlights an issue with the basic ROL model in the NHL draft setting. Players with more volatile rankings (i.e., players that are ranked highly by some agencies and are left unranked entirely by others) will have overestimated ability parameters because the cases where they are left entirely unranked do not factor into the likelihood at all.

To address this, we leverage the extension to the rank-ordered logit model for ranking the top  $m$  competitors out of a pool of  $M$  total competitors as outlined by Fok et al. [10]. The likelihood for a single draft ranking set in this case is expressed as follows:

$$P(Y_1 > Y_2 > \dots > Y_m > \max(Y_{m+1}, \dots, Y_M) \mid \theta_1, \theta_2, \dots, \theta_M) = \prod_{i=1}^m \frac{\exp(\theta_i)}{\sum_{j=i}^M \exp(\theta_j)} \quad (3)$$

Here we assume that a ranking set ranks  $m$  players out of a pool of  $M$  total players available. Referring back to the above example, this would now account for the fact that Aatu Rätty was unranked in ranking set B and adjust his  $\theta_i$  estimate accordingly.

## 2.4 Considering Changes in Rankings Over Time

At the beginning of the 2020-21 season, Aatu Rätty was ranked as a likely candidate for the 1st overall pick. However, Rätty struggled to perform well in his draft year and as the season wore on, he rapidly fell down every agency’s draft rankings until he was eventually selected 52nd overall in the 2021 NHL draft.

Suppose we were in the days leading up to the draft in June 2021, and ranking set A from September 2020 had Rätty ranked 1st overall, while ranking set B from May 2021 had Rätty ranked 45th overall. Using the ROL model as we have defined it so far would allow both ranking sets to influence the  $\theta_i$  estimates equally. However, ranking set B is likely more relevant to how the draft will play out in reality since it was built with an entire season of information that ranking set A did not observe.

This can be addressed by allowing player abilities to vary over time by assuming that the  $\theta_i$ ’s follow an autoregressive process through the season as done in Glickman and Hennessey [11]. To do so, we divide the season into time periods. Typically, this could be done according to key dates throughout the season, but the 2020-21 season had inconsistent scheduling across leagues due to COVID-19. We thus split the season into four three-month time periods as follows:

$$t = \begin{cases} 1, & \text{if between 2021-11-01 and 2021-02-01} \\ 2, & \text{if between 2021-02-01 and 2021-05-01} \\ 3, & \text{if between 2021-05-01 and 2021-07-23.} \end{cases}$$

We define  $\theta_t$  as the ability parameters for all players in time period  $t$ . Recall that the autoregressive process assumes that

$$\theta_{t+1} = \nu\theta_t + \delta_{t+1}$$

$$\delta_{t+1} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}).$$

Essentially, the ability parameter from the previous time period,  $\theta_{it}$ , is regressed towards zero by the autoregressive parameter  $\nu \in [0, 1]$  while varying by the random  $\delta_{t+1} \sim N(0, \tau^2)$  component to obtain the updated  $\theta_{i(t+1)}$ .

## 3 Model Setup

Now that we have laid out a ROL model for the NHL draft, we can move on to implementing the model in R [12] and Stan [13]. We opted to use Bayesian inference to fit this model as it involves a complex autoregressive hierarchical structure that is beyond the scope of any current ROL model packages available in R. The computation time for this model took approximately 55 minutes to run using the ‘sampling’ function from the ‘rstan’ package in R [14].

The likelihood used in our ROL model is simply the product of (3) from Section 2.3 over all draft ranking sets in all time periods as defined below. Here,  $K_t$  represents the number of draft ranking sets from time period  $t$  with  $m_{kt}$  and  $M_{kt}$  representing the total number of players ranked and the total number of draft-eligible players available to be ranked from our database, respectively, in the  $k$ th ranking set of the  $t$ th time period.

$$L(\theta, \nu, \tau) = \prod_{t=1}^3 \prod_{k=1}^{K_t} P(Y_1 > \dots > Y_{m_{kt}} > \max(Y_{m_{kt}+1}, \dots, Y_{M_{kt}}) \mid \theta_t) \quad (4)$$

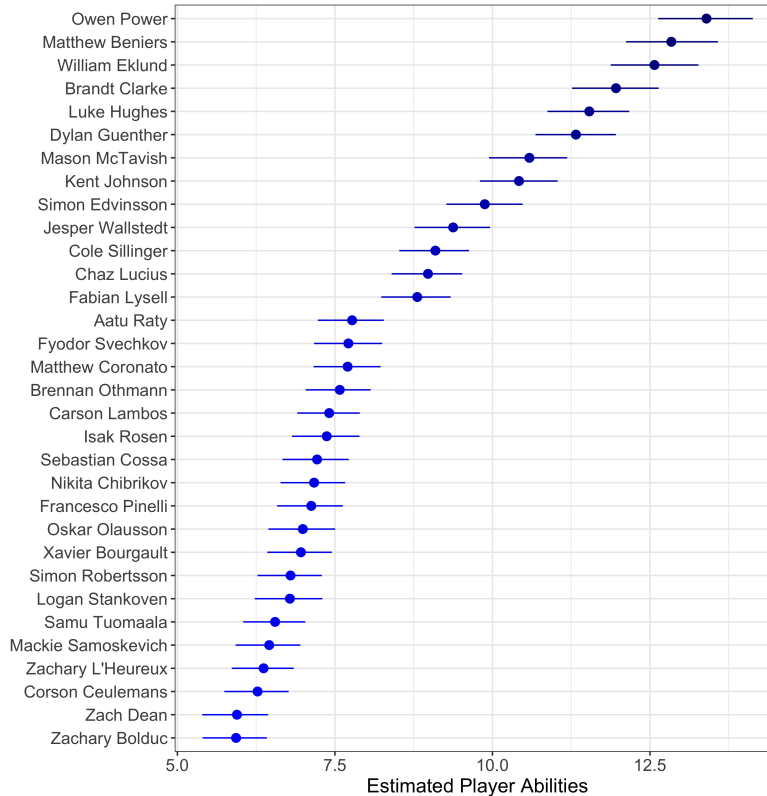
We assume a simple multivariate normal prior on the ability parameters in the first time period,  $\theta_1$ . Each subsequent time period leverages the autoregressive process described in Section 2.4 to set a prior on  $\theta_t$ ,  $t = 2, 3$ . Additionally, we assume hyperpriors on  $\nu$  and  $\tau$  of  $\text{Unif}(0,1)$  and  $\text{Inv-Gamma}(2,1)$ , respectively.

Since the variance of  $Y_i|\theta_{it} \sim \text{Gumbel}(\theta_{it}, 1)$  will remain constant at  $\frac{\pi^2}{6}$  for any value of  $\theta_{it}$  [15],  $\theta_t$  is only identifiable up to an additive constant. To address this, we impose a constraint on the model that all player ability parameters in a given time period must sum to zero. As a result, the ability parameters should be interpreted as ability relative to the other players being considered.

## 4 Results

### 4.1 Parameter Estimates

We obtain estimates for the player ability parameters,  $\theta_{it}$ , in each time period via posterior distributions from our Bayesian ROL model. Figure 1 displays the top 32 players based on their posterior means of  $\theta_{i3}$ . These ability estimates allow us to get a consolidated draft ranking based on our input data and determine the most likely draft outcome (by ordering abilities from greatest to least).



**Fig. 1.** Top 32 players in the NHL draft based on ability parameter posterior estimates from our rank-ordered logit model in time period 3 (2021-05-01 to 2021-07-23). Points represent the posterior means of  $\theta_{i3}$  for player  $i$ ; lines represent the corresponding 95% credible intervals of the posterior.

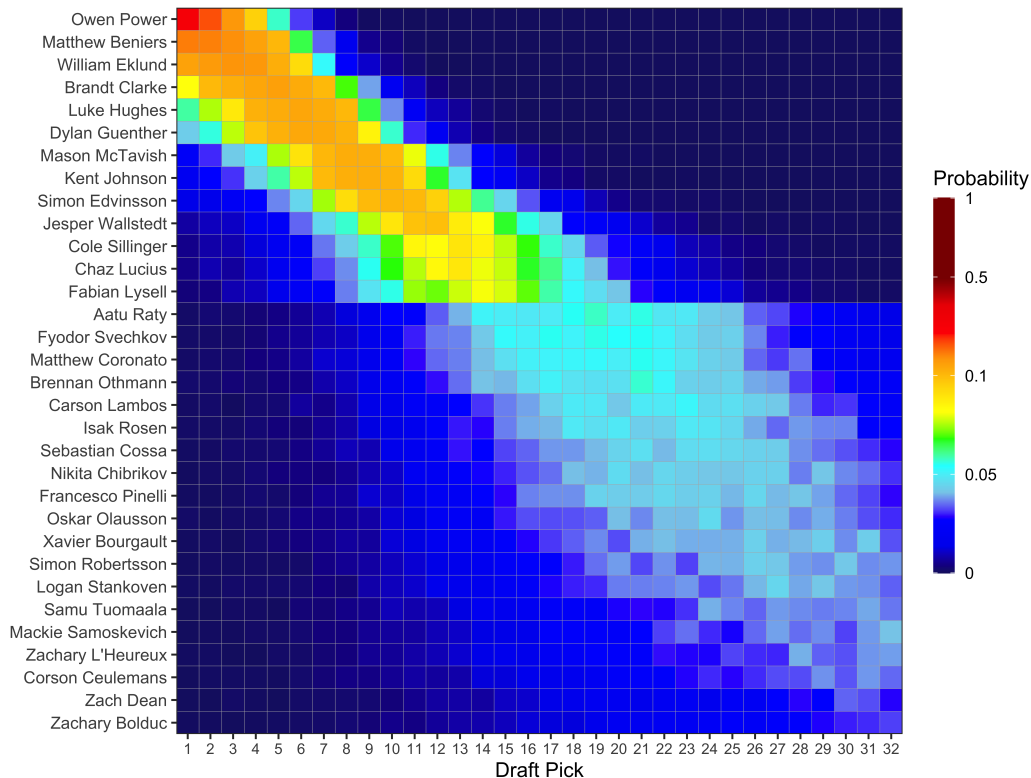
## 4.2 Draft Simulations

These player ability parameter estimates are much more powerful than a tool for basic comparison between players. We can also use these abilities to estimate the probability that player  $i$  will be selected with the next pick given the remaining pool of players  $i + 1, \dots, M$  available at that pick. This probability can be expressed as the following equation:

$$P(Y_i > \max(Y_{i+1}, \dots, Y_M) \mid \theta_t, Y_1 > Y_2 > \dots > Y_{i-1}) = \frac{\exp(\theta_{it})}{\sum_{j=i}^M \exp(\theta_{jt})} \quad (5)$$

With the player ability parameters estimated, we can now use (5) to simulate entire drafts. At each pick we use (5) to calculate the probability of each remaining player being selected at the pick of interest, then use these probabilities to take a multinomial draw of size 1 from the remaining players to simulate the next player selected.

Figure 2 provides an illustration of the probability distribution of pick/player combinations in the 2021 NHL draft as determined by these draft simulations. The probabilities are determined by taking the total number of cases where a player was drafted in a certain position and dividing it by the total number of simulations. We ran 10,000 simulations of the NHL draft based on posterior draws to produce this visual.



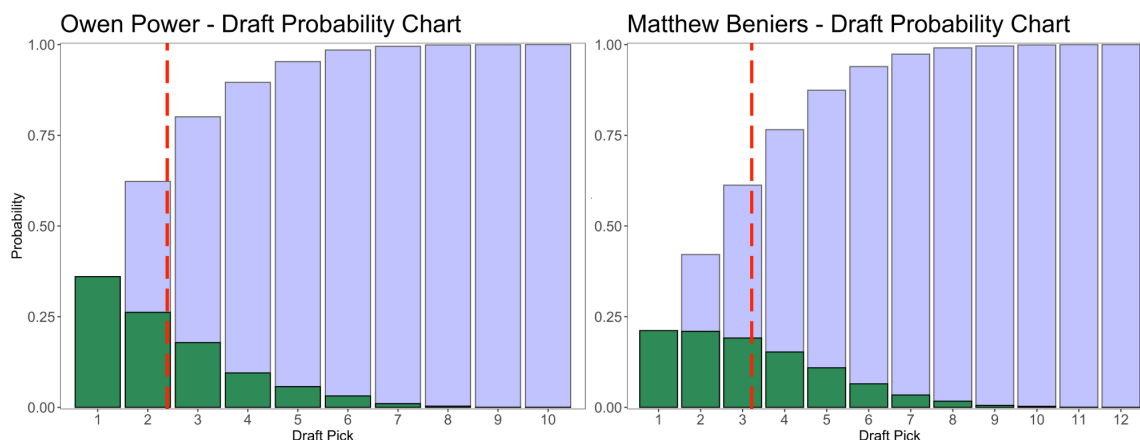
**Fig. 2.** A visualization of the probability that any of the top 32 players are selected with any of the top 32 picks; the colour of a square indicates the probability that the corresponding player will be picked with the corresponding pick.

The purpose of these draft simulations is to estimate the probability that a player is selected at any given draft pick. Ideally, we would compute this directly by calculating the probability for every possible permutation of the draft then summing up the total probability that player  $i$  is selected at pick  $j$  for all pairs of  $i, j$ ; however, this is computationally infeasible. Assuming we consider 400 draft-eligible players and select 224 (7 picks for each of 32 teams), there are  ${}_{400}P_{224} = 3.23565 \times 10^{548}$  possible draft outcomes. By simulating the NHL draft 10,000 times we can gain estimates of these probabilities without as much of a computational burden.

### 4.3 Player Ranking Distributions

Upon simulating the NHL draft using the posterior estimates of the player ability parameters, we can obtain discrete probability distributions for the pick number at which a player will be selected, which we call a ‘player ranking distribution’. For example, Figure 3 displays the player ranking distributions for Owen Power and Matthew Beniers.

To provide an example of how this model can be used by a team, consider a team with the 7th pick in the draft. Lets suppose they believe Matthew Beniers is going to be a superstar. From the cumulative distribution function (blue) provided in Figure 3, we can see that the probability that he is selected prior to the 7th pick is roughly 90%. Thus, to have a better shot at selecting Beniers, the team would have to consider trading their 7th overall pick plus additional assets in order to acquire a higher pick in the draft where Beniers will have a higher probability of being available.



**Fig. 3.** Player ranking distributions for Owen Power (left) and Matthew Beniers (right) based on 10,000 NHL draft simulations. Red, dashed lines represent expected draft pick based on the distribution.

## 5 Concluding Remarks

In summary, we built a rank-ordered logit model based on NHL draft ranking data. This model allows us to estimate the ability of draft-eligible players relative to their peers, simulate draft outcomes, and estimate a probability distribution for the pick at which each player will be selected.

This model is still a work in progress and we feel there are many different routes that we can take to improve its performance and accuracy. Primarily, we intend to model the ability parameter  $\theta$

by a linear predictor of player covariates with coefficients that assume a hierarchical structure to allow the model to adjust for team and agency tendencies. We expect that both agencies and teams will value particular traits (such as skating, shooting, passing, grit, etc.) differently and teams may draft players to address certain team needs (e.g., draft a defenceman when their roster and prospect pipeline are lacking talent on defence).

Additionally, we do not directly address the between-ranking correlation due to communication/collaboration between agencies. Two agencies may share thoughts amongst each other and, as a consequence, bias each other's evaluation of certain players. This has not been acknowledged directly in our paper and is an area that we hope to address with future work.

## Acknowledgements

We would like to thank Tyrel Stokes and Dr. Rachel Altman for their help with this project.

Tyrel was instrumental in helping us develop this idea, find resources to improve our model, and implement the model in Stan. We appreciate him taking the time to help us debug various issues with the multivariate priors in Stan, among other things.

Dr. Altman provided us with thought-provoking insights and questions regarding our model and was very generous with her time when discussing our ideas.

## References

1. Springings, D.: NHL Draft Probability Tool. (2016). Retrieved from <https://hockey-graphs.com/2016/06/08/nhl-draft-probability-tool/#more-9010>
2. Robinson, B.: Grinding the Bayes: A Hierarchical Modeling Approach to Predicting the NFL Draft. (2020). In proceedings of the 2020 Carnegie Mellon Sports Analytics Conference. <https://www.stat.cmu.edu/cmsac/conference/2020/assets/pdf/Robinson.pdf>.
3. ESPN. *Draft Day Predictor Tool*. ESPN Analytics. (n.d.). <http://espnstatsanalytics.com/about.html>.
4. Stokes, T. Monaco\_ranking [github repository]. (2022). Retrieved from [https://github.com/tyrelstokes/Monaco\\_ranking](https://github.com/tyrelstokes/Monaco_ranking).
5. Agresti, A.: An Introduction to Categorical data analysis. 3rd edn. New York: John Wiley (2019).
6. Train, K.: Discrete Choice Methods with Simulation. 2nd edn. Cambridge University Press (2009).
7. Luce, D., Suppes, P.: Preferences, utility and subjective probability. In: Luce, R. Bush, R., Galanter, E. (eds.) Handbook of Mathematical Psychology, pp. 249–410. John Wiley and Sons, New York (1965).
8. McFadden, D.: Conditional logit analysis of qualitative choice behavior. In: Zarembka, P (ed.) Frontiers in Econometrics, pp. 105– 142. Academic Press, New York (1974).
9. Train, K.: Discrete Choice Methods with Simulation. 2nd edn. Cambridge University Press (2009).
10. Fok, D., Paap, R., Van Dijk, B.: A rank-ordered logit model with unobserved heterogeneity in ranking capabilities. Journal of Applied Econometrics, **27** (5), 831–846. (2010). <https://doi.org/10.1002/jae.1223>.
11. Glickman, M., Hennessy, J.: A stochastic rank ordered logit model for rating multi-competitor games and sports. Journal of Quantitative Analysis in Sports, **11**(3), 131-144. (2015). <https://doi.org/10.1515/jqas-2015-0012>.
12. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2022). URL <https://www.R-project.org/>.
13. Stan Development Team: Stan Modeling Language Users Guide and Reference Manual, Version 2.21.0. (2022). <https://mc-stan.org>



14. Stan Development Team: RStan: the R interface to Stan. R package version 2.21.5. (2022). <https://mc-stan.org/>
15. Gumbel.: E.J.: Statistics of Extremes. Columbia University Press, New York (1958).