Microti data analysis

# Material and methods

### Statistics Analysis

Two outcome variables of interest were identified

* **Micro** microbiological status of the lymph nodes (dichotomous variable Pos / Neg)
* **Histo** histological grade of granulomas (ordinal variable with four levels: g1 / g2 / g3 / g4)

On the basis of the nature of the outcome variable, we adapted specific Bayesian regression models with the aim of identifying which variables among those measured significantly influence the outcomes. We intend this models as descriptive models (Berk [2010](#ref-berk2010you)) not aimed at prediction, but for capturing the associations between the dependent and independent variables rather for causal inferences or for prediction.

**Bayesian statistics** focuses on the posterior distribution where θ are the model parameters (unknown quantities) and Y are the data (known quantities) to condition on. The posterior distribution is generally computed as

In the above equation is the likelihood, is the prior distribution and is the marginal likelihood. The likelihood is the distribution of the data given the parameters and thus relates the the data to the parameters. The prior distribution describes the uncertainty in the parameters before having seen the data. It thus allows to explicitely incorporate prior knowledge into the model. The marginal likelihood serves as a normalizing constant so that the posterior is an actual probability distribution. In classical frequentist statistics, parameter estimates are obtained by finding those parameter values that maximise the likelihood. In contrast, Bayesian statistics estimate the full (joint) posterior distribution of the parameters. This is not only fully consistent with probability theory, but also much more informative than a single point estimate (and an approximate measure of uncertainty commonly known as ’standard error’). Obtaining the posterior distribution analytically is rarely possible and thus Bayesian statistics relies on Markov-Chain Monte-Carlo (MCMC) methods to obtain samples (i.e., random values) from the posterior distribution. Such sampling algorithms are computationally very intensive and thus fitting models using Bayesian statistics is usually much slower than in frequentist statistics. However. advantages of Bayesian statistics—such as greater modeling flexibility, prior distributions, and more informative results—are often worth the increased computational cost (Bürkner and Vuorre [2019](#ref-burkner2019ordinal); Gelman, Hill, and Vehtari [2020](#ref-gelman2020regression); McElreath [2020](#ref-mcelreath2020statistical))

## Micro

For this analysis, the statistical unit is the single lymph node and predictors used were:

* **NAF**: presence of acid fast bacteria in lymph node granulomas. This variable was obtained by calculating for each single lymph node the average number of bacteria present in the various granulomas and then categorizing the lymph node itself as NAF = 1 when the average was greater than 0 and NAF = 0 (i.e. absence of acid-resistant bacteria) in the case of averages equal to 0.
* **MNC**: similarly to what was done for NAF was done for the categorization of lymph nodes on the basis of the presence or absence of cells in granulomas.
* **G1, G2, G3, G4**: to grasp the influence that the histological grade of granulomas has on the microbiological state of a lymph node, we constructed four new variables, one for each histological grade, which identify the proportions of granulomas of the different grades for each single lymph node . For each lymph node we calculated the number of granulomas of the different histological grades and then we divided it by the total number of granulomas (**NGR**) per lymph node, thus obtaining a value between 0 and 1 which corresponds to the proportion of granulomas of the different grades. For example, 16 granulomas were observed in lymph node # 8, of which 0 of Grade 1 and 2, 13 of grade 3 and 3 of grade 4, so for lymph nodes 8 we will have this profile G1 = 0, G2 = 0, G3 = 13/16 (0.81), G4 = 3/16 (0.19). By way of example we report in the following table lymph nodes and their profile in relation to the microbiological status to the proportion of granulomas of different histological grades and the total number of granulomas.

| Idlinf | Micro | G1 | G2 | G3 | G4 | NGR |
| --- | --- | --- | --- | --- | --- | --- |
| 76 | 1 | 0 | 0.05 | 0.74 | 0.21 | 19 |
| 598 | 0 | 0 | 0.48 | 0.36 | 0.15 | 33 |
| 601 | 0 | 0 | 0.29 | 0.29 | 0.43 | 7 |
| 619 | 0 | 0 | 0.00 | 0.50 | 0.50 | 6 |
| 620 | 0 | 0 | 0.00 | 0.70 | 0.30 | 10 |

**Micro** was modeled using a Bayesian General Linear Model (McElreath [2020](#ref-mcelreath2020statistical)) . We consider the observed positive microbiological status M as a Bernoulli realizations of a random process with probability **p** to be positive:

Then **p** was modeled using **logit link** function :

Where X is the predictors design matrix; is the intercept parameter and is the matrix of predictors regression coefficients.  
The **logit link** maps a parameter that is defined as a probability mass, and therefore constrained to lie between zero and one, onto a linear model that can take on any real value. The logit function itself is defined as the *log-odds* :

The “odds” of an event are just the probability it happens divided by the probability it does not happen. So we can write:

from which it can be derived :

The above function is usually called **LOGISTIC**. In this context, it is also called the **INVERESE-LOGIT**, because it inverts the logti transform.

For both parameters ( and ) weakly uninformative priors were selected according to McElreath ([2020](#ref-mcelreath2020statistical)) :

The model parameter estimates were obtained by sampling with the Hamiltonian Monte Carlo (HMC) algorithm using four chains, with 4000 iterations of which 1000 were warmup (excluded after sampling). For this model, the brms (Bürkner [2017](#ref-burkner2017brms))package was used as the interface of the STAN language (Gelman, Lee, and Guo [2015](#ref-gelman2015stan)) implemented in R in the rstan (Team and others [2016](#ref-team2016rstan)) package.

Bayesian posterior estimates of the parameters were then summarized in table using the median of the posterior distribution, the 95% credibility intervals, the **Probability Direction (PD)**, as a measure of the importance of the effects of the parameters, and the **Region of Practical Equivalence (ROPE)** as an index of significance as reported by (Makowski et al. [2019](#ref-makowski2019indices))

**Probability Direction (PD)** can be interpreted as the probability that a parameter (described by its posterior distribution) is strictly positive (whichever is the most probable) and it varies between 50% and 100%. PD is defined as the proportion of the posterior distribution that is of the median’s sign (Makowski et al. [2019](#ref-makowski2019indices))

**Region of Practical Equivalence (ROPE)** is defined as the percentage of the whole posterior distribution that lies within a Region of Practical Equivalence (ROPE) defined as range from -0.1 to 0.1 for linear regression and its equivalent, -0.18 to 0.18 for logistic models (based on the formula to convert log odds ratios to standardized differences (Makowski et al. [2019](#ref-makowski2019indices); Kruschke [2014](#ref-kruschke2014doing); Cohen [1992](#ref-cohen1992statistical) )

# Histo

For this analysis, the statistical unit is the single granuloma and predictors used were:

* **lnGrArea**: log of the granuloma area
* **lnaf**: log of the number of acid fast bacteria per granuloma
* **lmnc**: log of the number of MNCs per granuloma
* **micro**: microbiological state of the lymph nodes (Pos / Neg)
* **Grcompl**: status of complete / incomplete granuloma
* **lngr**: number of granulomas per lymph fund
* **IdLinf**: code identifier of lymph (used as a “random” variable in the model)

In this case, a Bayesian ordinal regression model was fitted according to Bürkner and Vuorre ([2019](#ref-burkner2019ordinal)). The ordinal regression models are of three different types:

* **Cumulative Model (CM)**
* **Sequential Model (SM)**
* **Adjacent Model (AM)**

The **CM** assume that the observed ordinal variable **Y**, the histopatologic grade of granulomas in our case, originates from the categorization of a latent (not observed) continuous variable . To model this categorization process, the **CM** assumes that there are *K* theresholds . In our study there are response categoriee, and therefore thresholds

For many ordinal variable, the assumption of a single underlying continuos variable may not be appropriate. If the response can be understood as being the result of a sequential process, such that a higher response category is possible only after all lower categories are achieved, the **SM** model as proposed by (Tutz [1997](#ref-tutz1997sequential)) is usually appropriate. In our study this type of model assumes that the different classes of the outcome variable (in this case the histological grade) are the expression of a sequential process for which, for example, a granuloma is classified as grade 4, after having “evolved” from the previous grades. The **AC** is a widely used ordinal model in item-response theory and is applied in many large scale assessment studies. AC is very different to the CM and SM because it is difficult to think of a natural process leading to it. AC can be chosen for its mathematical convenience rather than any quality of interpretation.

We choosed **SM** because the assumption of the model with respect to the variable outcome is consistent with the histological development of the granulomas. The model is further complicated by the hierarchical structure of the data, whereby observations made at the granuloma level are nested under the lymph nodes. So we fitted a multilevel ordinal sequential model with **Idlymph** as random effect.

In **SM** the dependend variable Y results from a counting process and is truly ordinal in the sense that in order to achieve a category , one has to first achieve all lower categories 1 to . For every category there is a latent continuos variable determining the transition between the . We assume that depends on the predictor term and error :

where is the linear predictor: where is the matrix of regression coefficients and is the matrix of the values of predictors.

So we can formulate the bayesian model as:

with this priors :

The model parameter estimates were obtained by sampling with the Hamiltonian Monte Carlo (HMC) algorithm using four chains, with 4000 iterations of which 1000 were warmup (excluded after sampling). For this model, the brms (Bürkner [2017](#ref-burkner2017brms))package was used as the interface of the STAN language (Gelman, Lee, and Guo [2015](#ref-gelman2015stan)) implemented in R in the rstan (Team and others [2016](#ref-team2016rstan)) package.

Bayesian posterior estimates of the parameters were then summarized in table using the median of the posterior distribution, the 95% credibility intervals, the **Probability Direction (PD)**, as a measure of the importance of the effects of the parameters, and the **Region of Practical Equivalence (ROPE)** as an index of significance as reported by (Makowski et al. [2019](#ref-makowski2019indices))

In order to show how a particular predictor statistically influences the distribution of Histological grade of granulomas while holding the other predictors costant, we plot the conditional effects of relevant predictors, as described by (Bürkner [2017](#ref-burkner2017brms)).

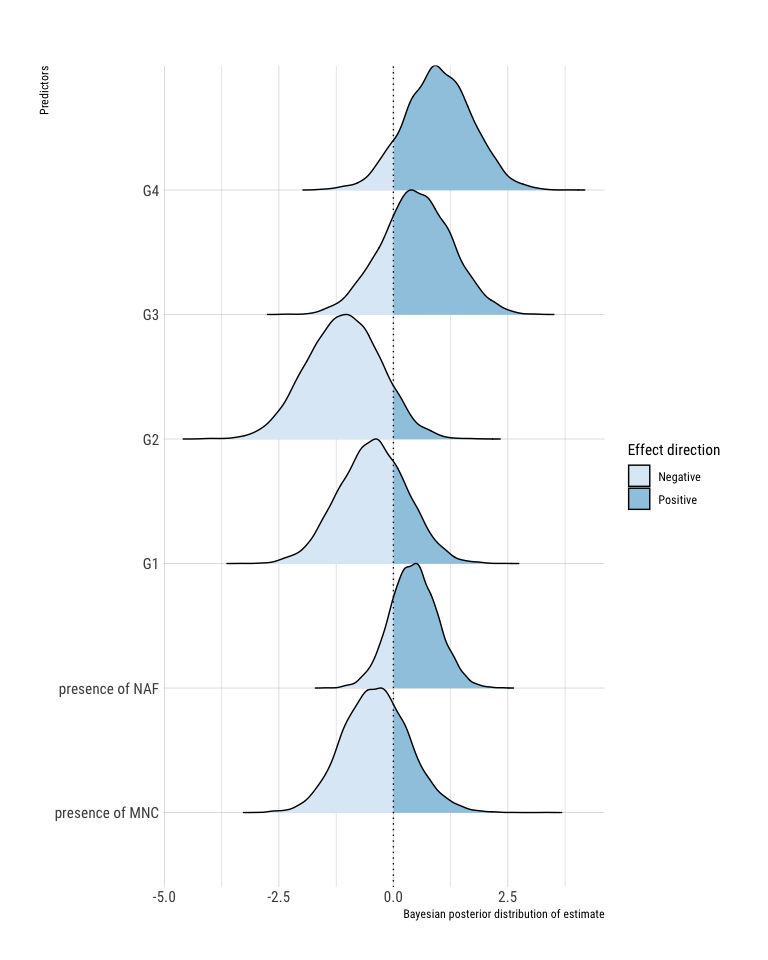
RESULTS

**MICRO**

The model results: median of posterior estimated coefficients, 95 intervals of credibility, Probability Direction, and Rope are shown in the table 1.

| Parameter | Median | CI\_low | CI\_high | pd | ROPE\_Percentage |
| --- | --- | --- | --- | --- | --- |
| Intercept | -1.41 | -3.33 | 0.42 | 0.93 | 0.05 |
| G1 | 0.98 | -0.45 | 2.42 | 0.90 | 0.08 |
| G2 | 0.50 | -1.00 | 2.02 | 0.75 | 0.15 |
| G3 | -1.08 | -2.66 | 0.43 | 0.91 | 0.07 |
| G4 | -0.44 | -1.92 | 0.98 | 0.72 | 0.16 |
| presence of NAF | 0.44 | -0.56 | 1.45 | 0.81 | 0.20 |
| presence of MNC | -0.38 | -1.79 | 1.05 | 0.70 | 0.17 |

Fig.1 report the posterior diastributions of estimated coefficents For each predictor, the distribution of the estimate of the regression coefficient is reported and the area corresponding to the values with negative effect and those with positive effect is indicated with different colors. Considering that the total area of the distribution of the values of the coefficients estimates is equal to 1 (100%), it is possible to calculate the area corresponding to the positive and negative values. In this way we obtain a measure of the importance of the effect estimated by the model called P\_direction (PD).

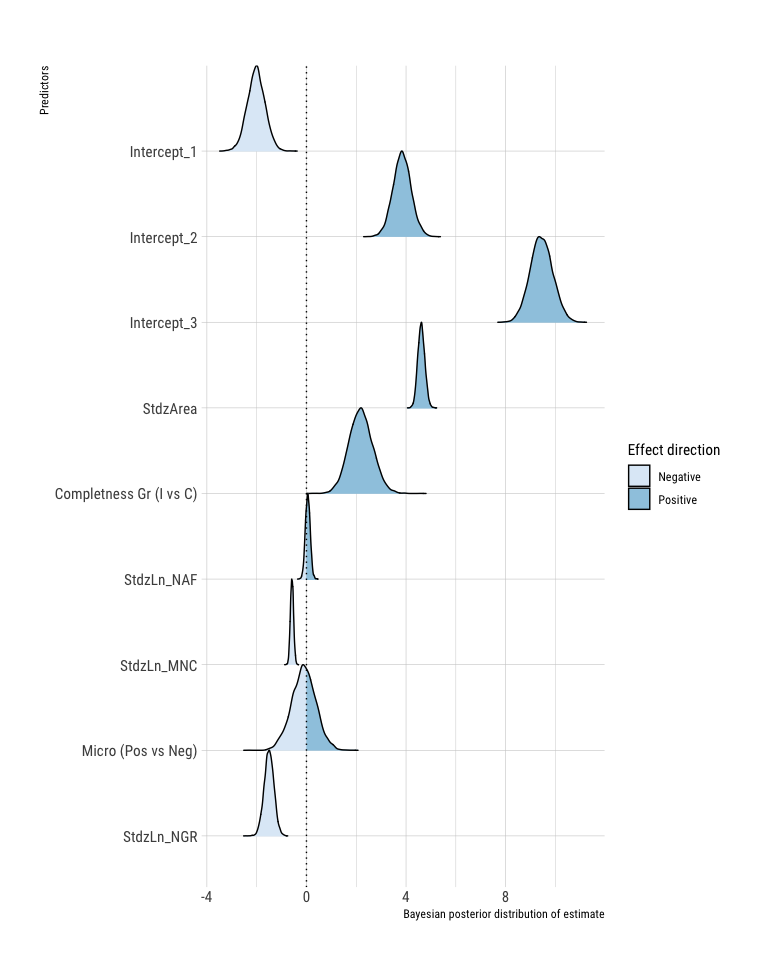


Basically as the proportion of grade 1 granulomas per lymph node increases there is a 90% probability that the probability that the lymph is microbiologically positive increases (keeping the value of all other variables constant). There is a similar effect for the variable G2 even if with a probability of 74%. As the proportion of grade 3 and 4 granulomas increases, there is a tendency to decrease the probability that the lymph node is microbiologically positive with a probability of 91% and 72% respectively. In lymph nodes in which granulomas with acid fast bacteria are present there is a greater probability that they are microbiologically positive in 81% of cases. On the contrary, in 70% of cases the presence of MNC in granulomas reduces the likelihood of the lymph nodes to be microbiologically positive. However, the relevance of the predictors in terms of distance from 0 is not particularly high, and the uncertainty of the estimates also appears to be rather large. Probably the aggregation of information used to analyze data at the lymph node level is not particularly effective in capturing the underlying effects.

**Histo** The model results: median of posterior estimated coefficients, 95 intervals of credibility, Probability Direction, and % in Rope are shown in the table 2.

| Parameter | Median | CI\_low | CI\_high | pd | ROPE\_Percentage |
| --- | --- | --- | --- | --- | --- |
| Intercept\_1 | -2.02 | -2.69 | -1.29 | 1.00 | 0.00 |
| Intercept\_2 | 3.84 | 3.16 | 4.60 | 1.00 | 0.00 |
| Intercept\_3 | 9.43 | 8.54 | 10.31 | 1.00 | 0.00 |
| StdzArea | 4.61 | 4.35 | 4.90 | 1.00 | 0.00 |
| Completness Gr (I vs C) | 2.18 | 1.26 | 3.13 | 1.00 | 0.00 |
| StdzLn\_NAF | 0.05 | -0.13 | 0.24 | 0.71 | 0.63 |
| StdzLn\_MNC | -0.58 | -0.70 | -0.45 | 1.00 | 0.00 |
| Micro (Pos vs Neg) | -0.11 | -1.05 | 0.88 | 0.59 | 0.16 |
| StdzLn\_NGR | -1.50 | -1.92 | -1.15 | 1.00 | 0.00 |

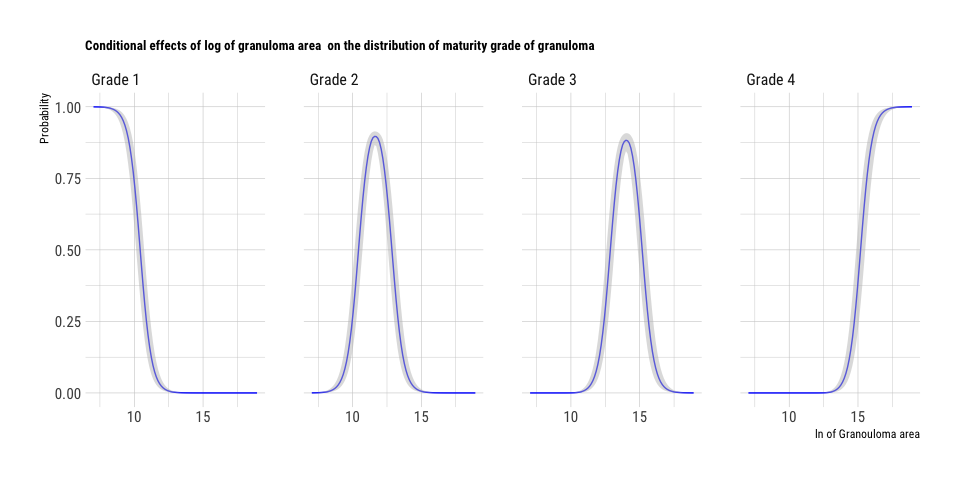
Fig.2 report the posterior distributions of estimated coefficents For each predictor, the distribution of the estimate of the regression coefficient is reported and the area corresponding to the values with negative effect and those with positive effect is indicated with different colors. Considering that the total area of the distribution of the values of the coefficients estimates is equal to 1 (100%), it is possible to calculate the area corresponding to the positive and negative values. In this way we obtain a measure of the importance of the effect estimated by the model called P\_direction (PD).

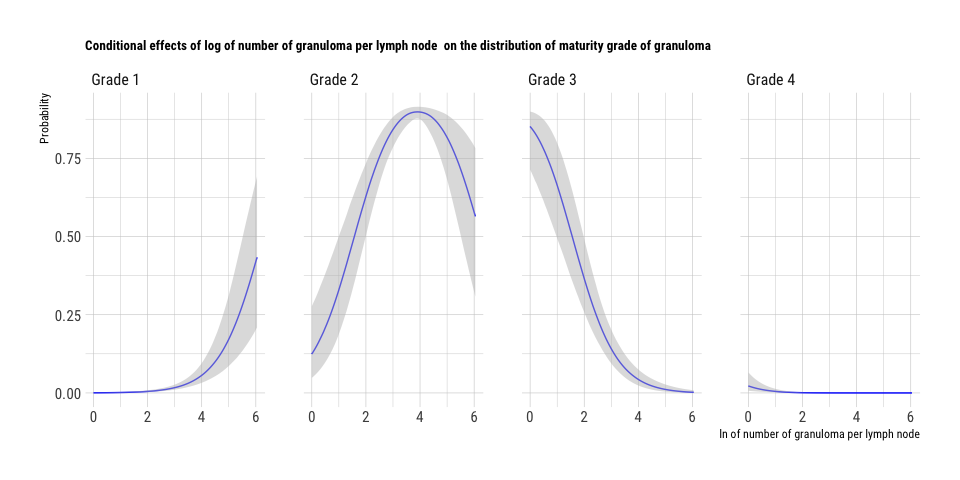


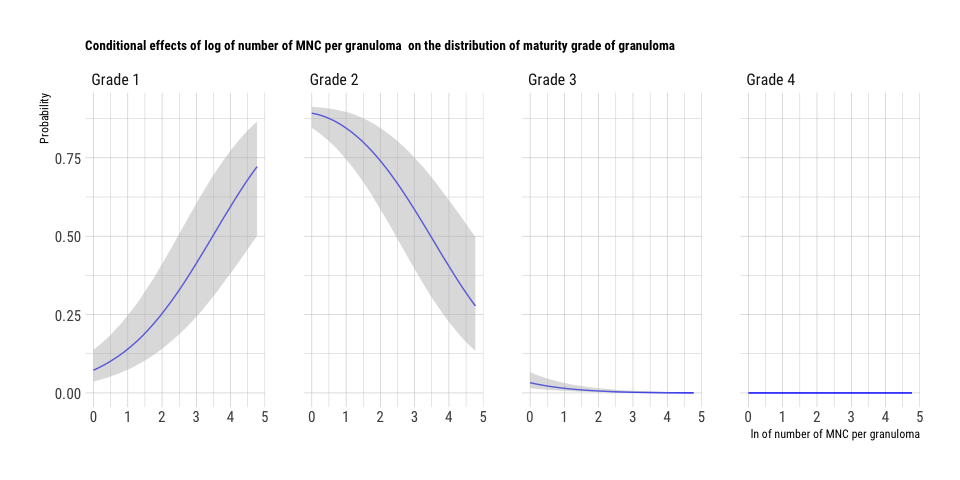
The model suggests that as granuloma size increases significantly (the regression coefficient is very far from 0!) the probability that the granuloma is classified in the higher grades of the histological scale. On the contrary, the increase in the number of granulomas in the lymph nodes reduces the probability that the granulomas of that lymph node are of high histological grade. The microbiological state does not seem to be relevant, while according to the model, the probability of granulomas being classified in high degrees decreases as the number of MNCs increases. As for the number of acid-fast bacteria, the model suggests that there is a positive effect, i.e. as the number of bacteria increases, the probability that the granuloma is of high histological grade increases but this result is supported only with a probability of 70%. Finally, incomplete granulomas compared to complete ones are more likely to be classified in high histological grades.

In this type of models it is more effective to visualize, at least for the most relevant predictors, the marginal effects, i.e. by keeping the other predictor values constant, observing how the probability of a granuloma of being classified in one of the four histological degrees varies as the predictor varies.

The following graphs show the effects of the predictors: size, number of granulomas, completeness, number of mnc.







## References

Berk, Richard. 2010. “What You Can and Can’t Properly Do with Regression.” *Journal of Quantitative Criminology* 26 (4): 481–87.

Bürkner, Paul-Christian. 2017. “Brms: An R Package for Bayesian Multilevel Models Using Stan.” *Journal of Statistical Software* 80 (1): 1–28.

Bürkner, Paul-Christian, and Matti Vuorre. 2019. “Ordinal Regression Models in Psychology: A Tutorial.” *Advances in Methods and Practices in Psychological Science* 2 (1): 77–101.

Cohen, Jacob. 1992. “Statistical Power Analysis.” *Current Directions in Psychological Science* 1 (3): 98–101.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge University Press.

Gelman, Andrew, Daniel Lee, and Jiqiang Guo. 2015. “Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization.” *Journal of Educational and Behavioral Statistics* 40 (5): 530–43.

Kruschke, John. 2014. *Doing Bayesian Data Analysis: A Tutorial with R, Jags, and Stan*. Academic Press.

Makowski, Dominique, Mattan S Ben-Shachar, SH Chen, and Daniel Lüdecke. 2019. “Indices of Effect Existence and Significance in the Bayesian Framework.” *Frontiers in Psychology* 10: 2767.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC press.

Team, Stan Developent, and others. 2016. “RStan: The R Interface to Stan.” *R Package Version* 2 (1).

Tutz, Gerhard. 1997. “Sequential Models for Ordered Responses.” In *Handbook of Modern Item Response Theory*, 139–52. Springer.