Supplementary Material for Veríssimo (2021, *BLC*)

João Veríssimo

Veríssimo, J. (2021). Analysis of rating scales: A pervasive problem in bilingualism research and a solution with Bayesian ordinal models. *Bilingualism: Language and Cognition*.

João Veríssimo, Potsdam Research Institute for Multilingualism, University of Potsdam. Current affiliation: Center of Linguistics, School of Arts and Humanities, University of Lisbon.

Correspondence concerning this article should be addressed to João Veríssimo, Faculdade de Letras da Universidade de Lisboa, Alameda da Universidade, 1600-214 Lisboa, Portugal. E-mail: jlverissimo@edu.ulisboa.pt

Supplementary Tables

Table S1

Summary of a Bayesian ordinal model (thresholded-cumulative, with flexible thresholds) fit to Schlenter's (2019) acceptability ratings of canonical and non-canonical sentences.

	Estimate	SE	L-95% CI	U-95% CI
Intercept[1]	-3.50	0.21	-3.93	-3.10
Intercept[2]	-2.81	0.18	-3.17	-2.45
Intercept[3]	-2.50	0.18	-2.86	-2.16
Intercept[4]	-1.98	0.17	-2.32	-1.66
Intercept[5]	-1.39	0.16	-1.71	-1.07
Intercept[6]	-0.36	0.16	-0.66	-0.05
Condition (non-canonical vs. canonical)	-0.68	0.07	-0.83	-0.54

Note. Condition is coded as 0='canonical', 1='non-canonical'; thus, the negative effect of Condition indicates lower acceptability for the non-canonical sentences. SE: Standard error; L-95% CI, U-95%: Lower and upper bounds of the 95% credible interval.

Table S2

Summary of a Bayesian ordinal model (thresholded-cumulative, with flexible thresholds) fit to Puebla's (2016) proficiency ratings.

	Estimate	SE	L-95% CI	U-95% CI
Intercept[1]	-4.45	0.68	-5.83	-3.15
Intercept[2]	-2.64	0.52	-3.67	-1.65
Intercept[3]	-1.42	0.45	-2.30	-0.54
AoA	-0.15	0.03	-0.22	-0.09

Note. SE: Standard error; L-95% CI, U-95%: Lower and upper bounds of the 95% credible interval

Supplementary Figures

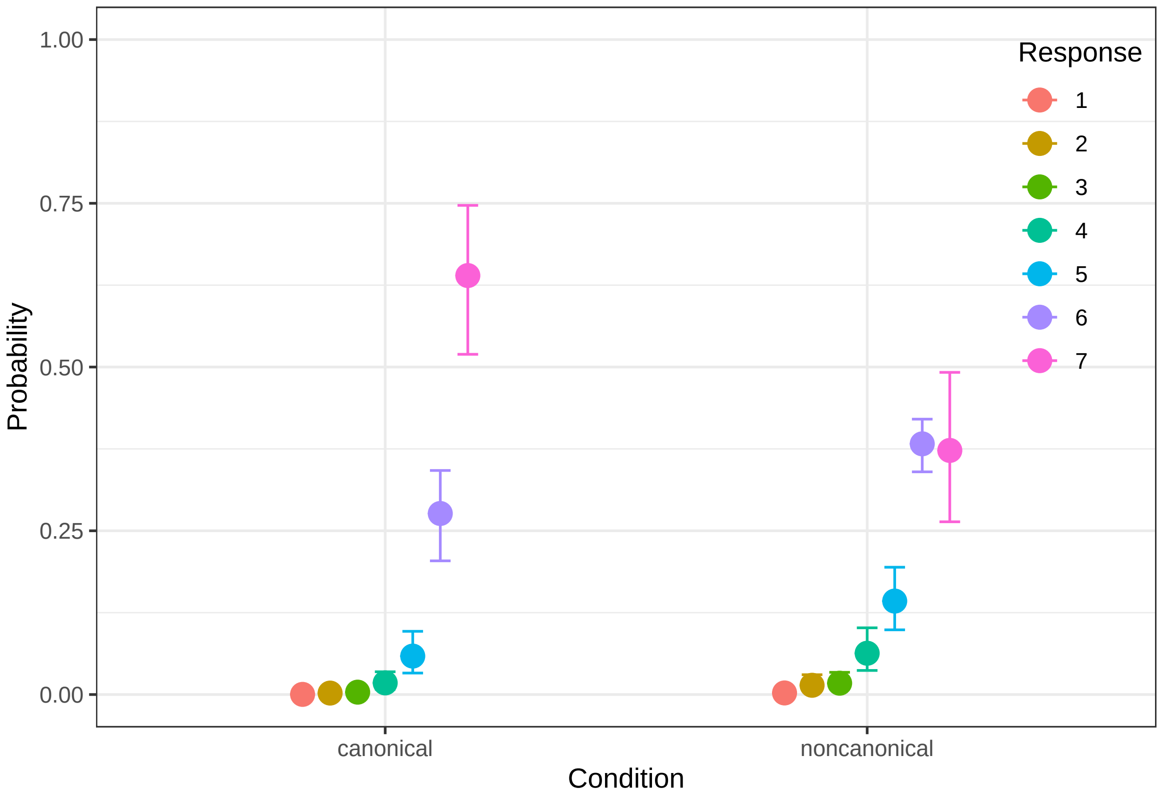

Figure S1. Conditional effects of condition (canonical, non-canonical) on sentence acceptability. Given that responses are mutually exclusive, their predicted proportions add up to 100% in each condition. Error bars indicate 95% credible intervals. Data from Schlenter (2019).

Appendix S1

Model comparisons and model complexity

Assessing the goodness-of-fit of different models is an important step in analyses with ordinal models, and in Bayesian analyses more generally (Schad, Betancourt, & Vasishth, 2020; Vasishth, Nicenboim, Beckman, Li, & Kong, 2018). In the main paper we show one way of visually assessing model quality, namely, by using predictive checks. Additionally, different models can also be formally compared in terms of their relative goodness-of-fit.

As an example, the two models reported in the paper, `m.equidistant` (with equidistant thresholds) and `m.flexible` (with flexible thresholds), can be compared. The flexible model is much more free, and thus can potentially fit the data better, but at the expense of requiring more parameters. We will use the function `loo_compare()` of the *brms* package, which returns the difference between the models' *expected log pointwise density* (ELPD). This is a measure of a model's predictive accuracy if applied to a new dataset. It can be computed by estimating how well each data point is predicted from all others (i.e., if the datapoint was taken out), a procedure referred to as leave-one-out cross-validation (LOO; Vehtari, Gelman, & Gabry, 2017):

```
m.equidistant <- add_criterion(m.equidistant, "loo")
m.flexible <- add_criterion(m.flexible, "loo")
loo_compare(m.equidistant, m.flexible)
```

```
##               elpd_diff se_diff
## m.flexible       0.0       0.0
## m.equidistant -126.7      15.3
```

The first row of the output shows that the flexible model is preferred, despite its greater complexity (the difference of 0 reflects the comparison of this model against itself). The equidistant model's ELPD is much smaller (by -126.70), and this amounts to a difference greater than 8 standard errors (SEs) relative to the flexible model (a difference greater than 2 SEs suggests that one model is better than the other; Bürkner, 2017; Vasishth, Nicenboim, Beckman, Li, & Kong, 2018).

Generally speaking, models with flexible thresholds are more appropriate, but there are cases in which equidistant models may suffice, and there may be advantages to their simplicity. In a flexible-threshold model, the number of parameters depends on the number of response categories. As an example, modelling responses to the 11-point proficiency scale of the LEAP-Q questionnaire (Marian, Blumenfeld, & Kaushanskaya, 2007) would require 10 parameters with a flexible-thresholds model (one for the threshold between each response), whereas an equidistant-thresholds model would require 2 parameters (one for the first threshold and one for the distance between each pair). Models with more parameters will typically fit the data better but they may also be more difficult to estimate. For example, if some response categories are chosen very rarely (as is often the case), estimating the more extreme thresholds comes with very large uncertainty.

In addition to the use of flexible thresholds, ordinal models of greater complexity can also be fitted (and compared) using the *brms* R-package. For example, *unequal-variances* models estimate latent distributions with different variances for each level of a predictor (e.g., in different conditions or groups). As demonstrated by Liddell and Kruschke (2018), ignoring that underlying distributions may have different variances can lead to serious distortions in the estimation of effects. More detailed examples of different types of ordinal models and of their comparison can be found in Bürkner and Vuorre (2019).

References

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/gddxwp

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/gfv26q

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/gfdbv8

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940–967. https://doi.org/bt2xwb

Puebla, C. (2016). *L2 proficiency survey*. Unpublished raw data, Potsdam Research Institute for Multilingualism, University of Potsdam.

Schad, D. J., Betancourt, M., & Vasishth, S. (2020). *Toward a principled Bayesian workflow in cognitive science*. Manuscript submitted for publication. Retrieved from https://arxiv.org/abs/1904.12765

Schlenter, J. (2019). *Predictive language processing in late bilinguals* (Doctoral dissertation). Universität Potsdam. https://doi.org/10.25932/publishup-43249

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. https://doi.org/gfzq3c

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/gdj2kz