Analysis of rating scales: A pervasive problem in bilingualism research and a solution with Bayesian ordinal models

João Veríssimo[1,2]

[1] Potsdam Research Institute for Multilingualism
[2] Center of Linguistics, School of Arts and Humanities, University of Lisbon

Preprint created on 24 Jun, 2021
Accepted for publication in *Bilingualism: Language and Cognition*

Abstract

Research in bilingualism often involves quantifying constructs of interest by the use of rating scales, for example, to measure language proficiency, dominance, or sentence acceptability. However, ratings are a type of ordinal data, which violates the assumptions of the statistical methods that are commonly used to analyse them. As a result, the validity of ratings is compromised and the ensuing statistical inferences can be seriously distorted. In this article, we describe the problem in detail and demonstrate its pervasiveness in bilingualism research. We then provide examples of how bilingualism researchers can employ an appropriate solution using Bayesian ordinal models. These models respect the inherent discreteness of ratings, easily accommodate non-normality, and allow modelling unequal psychological distances between response categories. As a result, they can provide more valid, accurate, and informative inferences about graded constructs such as language proficiency. Data and code are publicly available in an OSF repository at https://osf.io/grs8x.

Keywords: rating scales, ordinal models, Bayesian analysis, language proficiency, acceptability judgements

Research in bilingualism, as in psycholinguistics more generally, often involves quantifying constructs of interest by the use of *rating scales*. In these items, participants are asked to express a belief by selecting a response from an ordered set (e.g., "How well do you speak English?" '0=not well at all' to '6=very well'). The use of rating scales is ubiquitous in bilingualism research, in particular as part of the standard bilingualism questionnaires, for example, the Language History Questionnaire (Li, Sepanski, & Zhao, 2006), the Language Experience and Proficiency Questionnaire (LEAP-Q) (Marian, Blumenfeld, & Kaushanskaya, 2007), and the Bilingual Language Profile (Birdsong, Gertken, & Amengual, 2012).

One reason for the extensive use of rating scales is that they have wide applicability: they can be customised to measure many different constructs, including language proficiency (Hakuta, Bialystok, & Wiley, 2003), nativelikeness of speech (Flege, Yeni-Komshian, & Liu, 1999; Hopp, 2009), frequency of language mixing (Li, Sepanski, & Zhao, 2006), language attitudes (Birdsong, Gertken, & Amengual, 2012), sentence grammaticality (Cho & Slabakova, 2014), and semantic relatedness (Farhy & Veríssimo, 2019). Despite this heterogeneity, ratings are thought to yield quite valid and reliable measurements of such constructs (Marian, Blumenfeld, & Kaushanskaya, 2007). Ratings are also particularly useful to assess graded quantities in bilingualism research (e.g., language proficiency or dominance). Finally, they are broadly considered to be simple to design, administer, and analyse.

The simplicity and usefulness of rating scales may, however, be deceptive. Concerns have been raised that ratings may be too subjective, variable, and unidimensional to capture complex constructs like language proficiency (Hulstijn, 2012; Lemhöfer & Broersma, 2012; Tomoschuk, Ferreira, & Gollan, 2019; Zell & Krizan, 2014). Additionally, as discussed below, ratings constitute a type of data that violates the assumptions of the statistical methods that are commonly used to analyse them. As a result, their validity is compromised and the ensuing statistical inferences can be seriously distorted.

The problem

The approach that is near-universally used for the analysis of ratings is to assign a number to each response and then compute descriptive statistics like means and standard deviations, followed by inference from linear models, for example, ANOVAs, *t*-tests, and linear regressions. The general problem with this approach is that these methods are appropriate

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

I thank Laura Anna Ciaccio, Daniela Mertzen, Cecilia Puebla, Judith Schlenter, Nick Pandža, and two anonymous reviewers for helpful comments. I also thank Cecilia Puebla and Judith Schlenter for sharing their datasets and for granting me permission to make them public. This work has been partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 317633480 – SFB 1287, Project Q.

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

only for *metric variables*, which are continuous and inherently quantitative, whereas rating scales are *ordinal variables*, in which data consists of ordered, but discrete categories.

The inclination to analyse ratings with metric methods arises because responses can be easily assigned numerical values. However, ordinal data lacks the important property of *equidistance* between points, which is a requirement for the application of metric methods; that is, the interval between values cannot be assumed to be constant throughout the scale. For example, in the proficiency question of the LEAP-Q, the psychological distance between ‘7=good’ and ‘8=very good’ is likely to be larger than that between ‘5=adequate’ and ‘6=slightly more than adequate’, due to the distinct verbal labels. Even if exquisite care is taken in devising the labels, there are systematic psychological biases in how responses are perceived (DeCastellarnau, 2018; Krosnick & Presser, 2010). For example, extreme responses tend to be avoided (Douven, 2018), such that the psychological distance between an endpoint and its adjacent value may be particularly large (e.g., between ‘0=none’ and ‘1=very low’ in the LEAP-Q). Likewise, responses on each side of the midpoint may be qualitatively different, and thus perceived to be further apart.

Such tendencies and biases may also vary across tasks, populations, and individuals (Dawes, 2008; Kuncel, 1977; Schwarz, 1999). A striking example is that the relationship between self-ratings and objective measures of proficiency may vary as a function of language dominance and the particular languages used (Tomoschuk, Ferreira, & Gollan, 2019). For example, Spanish-English bilinguals who self-rate as a ‘4’ (on a 7-point scale) have greater objective proficiency than Chinese-English bilinguals, but the difference between groups disappears or reverses for bilinguals who self-rate as a ‘7’. Such results suggest that ratings vary according to an internal ‘frame of reference’, which may in turn differ across bilingual groups.

Consequences of analysing ordinal data with metric methods

If the distance between values of a rating scale cannot be assumed to be constant, then even the interpretation of a simple mean breaks down. Figure 1 shows how a 7-point proficiency scale may be mentally represented if participants avoid extreme responses and perceive crossing the midpoint as particularly impactful. In this case, the average proficiency expressed by the hypothetical responses {‘2’, ‘3’, ‘4’} is not identical to that expressed by {‘3’, ‘3’, ‘3’}, but actually reflects *greater* average proficiency. From this we see that when equidistance is violated, the very same metric mean (here, 3) may not express the same underlying quantity; rather, what the mean signifies depends on the particular distribution of responses. This lack of consistency between an underlying quantity and the metric mean naturally extends to effects, that is, to differences between means and regression slopes. Thus, at the very least, estimating ordinal effects with metric methods brings about a serious problem of interpretability.

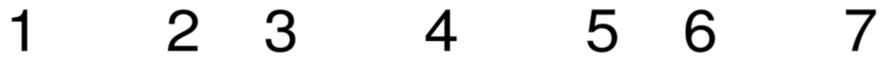

Figure 1. The hypothetical mental representation of a 7-point scale with unequally-spaced intervals.

Importantly, as demonstrated by Liddell and Kruschke (2018) with both simulated and real data, metric methods may produce serious inferential errors when applied to rating scales, namely: the detection of effects that do not exist; the failure to detect effects that do exist; and distortions of effect-size estimates. It is even possible for differences to be *inverted*, that is, for effects in one direction in the underlying construct (e.g., group A has larger proficiency than group B) to turn into effects in the opposite direction when mapped to metric means (e.g., group B > group A) (this can happen when the underlying distributions have different variances; see Liddell & Kruschke, 2018).

A (very) wrong model for ordinal data

The negative consequences described above arise because, in addition to unequal distances between responses, other properties of ordinal data are also fundamentally incompatible with the statistical models commonly used to analyse them. To see how this is the case, consider the following dataset.[1] Schlenter (2019) conducted a study in which bilingual speakers listened to 56 German sentences with canonical and non-canonical orders of thematic arguments. In order to determine the acceptability of the two variants, these sentences were first rated by a group of 42 native speakers, using a 7-point scale ('1=not acceptable'. . . '7=very acceptable').

A typical analysis would involve computing means in the two conditions (canonical: 6.35, non-canonical: 5.76), and then fitting a linear model, for example, a regression with condition as predictor. The results indicate that non-canonical sentences are less acceptable (b=-0.59, t=-9.64). As described above, it is hard to interpret such differences: it is not clear what it means for a construction to have 0.59 'units of acceptability' less, and moreover, if equidistance cannot be assumed, the interpretation of this difference depends on the distribution of responses.

In addition, it can be shown that the assumptions of linear regression are severely violated. This can be assessed with a *predictive check*, in which instances of 'predicted data' are generated from the fitted statistical model and then compared to the observed data that the models were fitted on; this check is shown in Figure 2, panel a. A comparison of the lighter lines (predicted) against the darker line (observed) reveals that the linear model: (a) grossly underestimates the proportions of '6's and '7's; (b) predicts impossible non-integer

[1] Data and code can be downloaded from https://osf.io/grs8x. This article was composed as a reproducible manuscript using the R-package *papaja* (Aust & Barth, 2020).

responses (e.g., 6.27); and (c) predicts a sizeable proportion of responses outside the 7-point scale.

The mismatches happen because linear models are inherently continuous and assume that errors are normally distributed. In other words, the best inference we can draw from this model is that the observed data in each condition comes from a normally-distributed population. However, because these assumptions yield impossible data, we know that this inference is fundamentally flawed. To be sure, the assumptions of statistical models are never fully satisfied ("all models are wrong"; Box, 1976). However, linear models with normally-distributed errors are particularly wrong when applied to ordinal data.

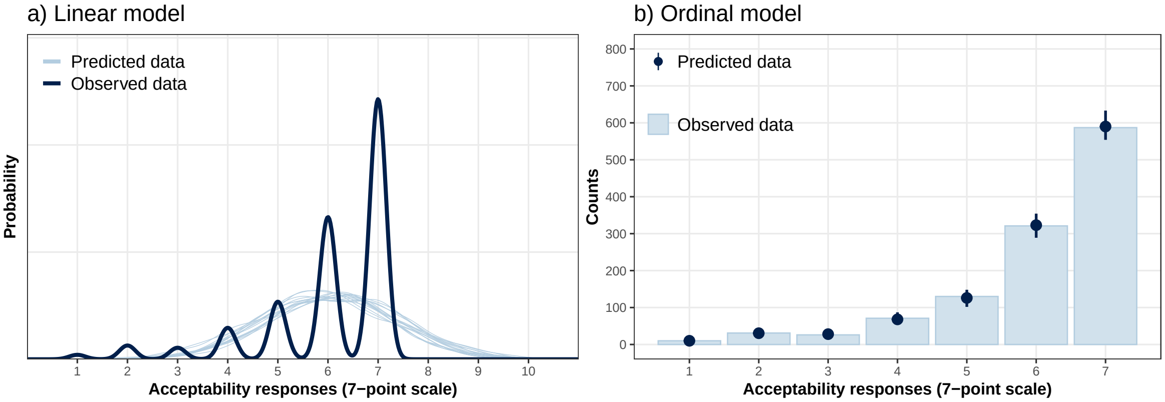

Figure 2. Predictive checks for a mixed-effects linear regression model (panel a), and for an ordinal model with flexible threshold locations (panel b). These assessments of model quality involve comparing the probability/counts of different responses (given in a 7-point acceptability scale) in data predicted by the fitted models (lighter lines in panel a, and dots in panel b) against the observed data that the models were fitted on (darker line in panel a, and bars in panel b). The two panels depict predicted and observed data differently because the linear model treats responses as continuous and the ordinal model appropriately treats them as discrete.

A pervasive problem in bilingualism

How widespread is the analysis of ordinal data with metric methods in bilingualism research? Although there are examples of the application of appropriate methods (e.g., Kissling, 2018; Tare et al., 2018), this is far from common. We quantified this tendency by searching for all articles published in *Bilingualism: Language and Cognition* in 2019–2020, containing the words 'rating' or related words. Forty-six articles analysed ordinal data, mostly proficiency ratings. Of these, only 3 appropriately summarised it with counts instead of means, but skipped inferential statistics or reported inappropriate ones; only 2 used a method that can be applied to ordinal data (Spearman's correlation), but still computed means. No article used both descriptive and inferential statistics that respect the properties of ordinal data. The remainder of this article describes and exemplifies a solution to this pervasive problem.

The solution

A better statistical model for ordinal data should fulfil the following requirements: first, predict discrete response categories, rather than continuous outcomes; second, accommodate non-normal response distributions; and third, allow for unequal distances between responses. Ordinal models satisfy all three requirements (McCullagh, 1980; McKelvey & Zavoina, 1975), and they are relatively simple for psycholinguists to fit, using readily-available software.

The particular class of ordinal model we describe here is the *thresholded-cumulative* model. Its basic idea is to assume an underlying *latent* continuous variable (e.g., proficiency), which is mapped to proportions of observed responses by being 'discretised', that is, chopped into intervals by thresholds placed at different points. Figure 3 shows an example (described in greater detail below). The normal distribution represents the variability of the latent variable (i.e., more or less proficiency) and the vertical thresholds divide the distribution into ordered responses ('1'...'7'). The proportion of '1's is given by the area under the curve up to the first threshold; the proportion of '2's by the area between the first two thresholds, etc. The threshold locations are estimated from the observed proportions, and in this way, we can model how the construct of interest is mapped to the responses.

In the next sections, we provide practical examples of fitting, interpreting, and assessing ordinal (thresholded-cumulative) models in the R programming language (R Core Team, 2020). Models will be constructed in the framework of Bayesian statistics, with the easy-to-use package *brms* (Bürkner, 2017). We opt for a Bayesian approach, because such models: (a) typically provide greater flexibility and can be easily extended in complexity (Bürkner, 2018; Bürkner & Vuorre, 2019); (b) converge more easily on accurate values (Liddell & Kruschke, 2018); (c) provide a more natural and informative quantification of uncertainty, because each quantity is accompanied by a full probability distribution (McElreath, 2020). However, note that ordinal models in a frequentist framework provide another valid solution for analysing ratings (see *ordinal* package; Christensen, 2019).

The current article is a brief introduction to the application of Bayesian ordinal models, with examples drawn from bilingual research. It does not constitute a full tutorial, as the examples below sidestep several issues that would be considered in a real analysis.[2] For more detailed tutorials and comprehensive treatments of Bayesian analyses, see Bürkner and Vuorre (2019), Vasishth, Nicenboim, Beckman, Li, and Kong (2018), and Schad, Betancourt, and Vasishth (2021).

Modelling unequal distances

The modelling of unequal distances can be better understood by comparing models with and without equidistance. We will make use of a simulated dataset of proficiency ratings,

[2] These include setting contrasts (Schad, Vasishth, Hohenstein, & Kliegl, 2020), selecting a random-effects structure (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017), specifying Bayesian priors, and various model checks (Schad, Betancourt, & Vasishth, 2021).

Table 1

Summary of a model with flexible thresholds, with estimates and 95% credible intervals for the position of each threshold. For comparison, estimates from the equidistant model are also displayed, as well as the 'true' population values which were used to generate the data.

	Estimate	L-95% CI	U-95% CI	Est. Equidistant	Population
Intercept[1]	-1.92	-2.08	-1.76	-2.16	-2.00
Intercept[2]	-1.21	-1.31	-1.11	-1.29	-1.25
Intercept[3]	-0.73	-0.82	-0.65	-0.42	-0.75
Intercept[4]	0.75	0.67	0.84	0.45	0.75
Intercept[5]	1.27	1.16	1.38	1.32	1.25
Intercept[6]	1.97	1.80	2.14	2.19	2.00

consisting of the (randomly generated) responses of 1,000 participants, given on a 7-point scale. The dataset aims to represent a heterogeneous group of L2 speakers with a mean level at the midpoint of the scale. Importantly, the data was generated assuming a representation similar to that in Figure 1, that is, with *unequal distances* between values. The counts of each response are shown in Figure 4 below (as bars), and suggest an exaggerated tendency for choosing the midpoint of the scale.

We use the `brm()` function to fit two (Bayesian) thresholded-cumulative models to this dataset. We use a *probit* link function, which means that we assume a normally-distributed latent variable. In the first model (`m.equidistant`), the distance between each threshold is assumed to be the same:

```
m.equidistant <- brm(Response ~ 1,
      data = ratings.unequal,
      family = cumulative(link="probit", threshold="equidistant"))
```

In the second model (`m.flexible`), we allow consecutive thresholds to be estimated at any distance from one another:

```
m.flexible <- brm(Response ~ 1,
      data = ratings.unequal,
      family = cumulative(link="probit", threshold="flexible"))
```

A summary of the flexible model is shown in Table 1 with estimates of each threshold's position and associated 95% credible intervals (this is the range within which a parameter falls with 95% probability). The estimates are expressed in standard deviation (SD) units; for example, the first threshold (labelled `Intercept[1]`) is 1.92 SDs below the mean of the latent distribution. The thresholds are more easily interpreted when visualised, as in Figure 3. In order to capture the observed proportions of each response ('1' to '7'), the thresholds are estimated to be closer together (e.g., second and third thresholds) or further apart.

The arrows at the bottom of Figure 3 depict the 'true' population values, which in this case (and unlike with real data) are already known. The models can be assessed by whether they can recover these underlying parameters, and it can be seen that the flexible model does a very good job. The equidistant model fares worse (see Table 1), especially for the third and fourth thresholds, which are estimated as much closer together than the population values.

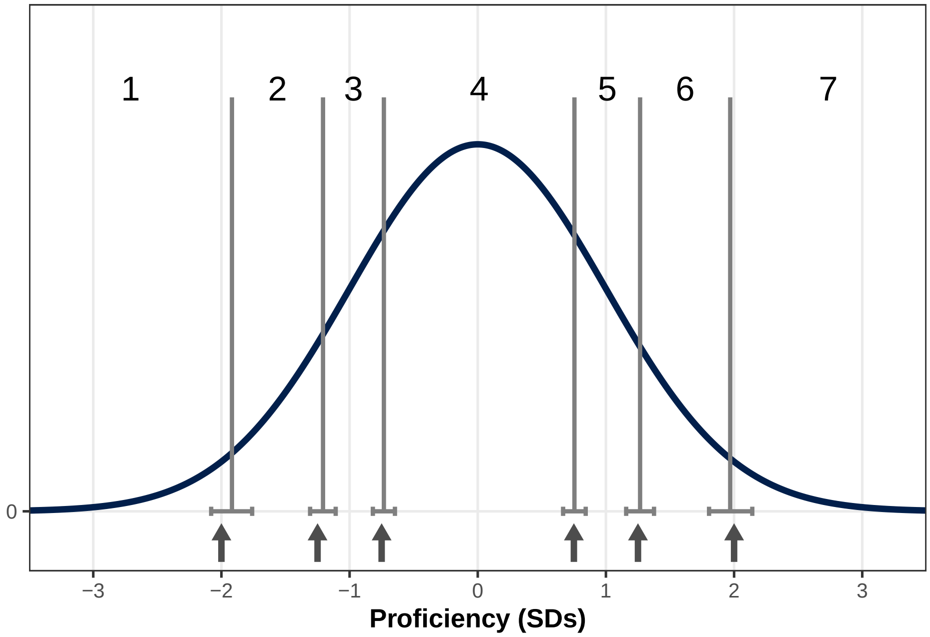

Figure 3. Ordinal model with flexible thresholds fitted to responses given on a 7-point proficiency scale (data was simulated in order to have unequal distances between responses). Estimates of each threshold are shown by vertical lines, with their 95% credible intervals at the bottom of each line. Predicted probabilities of each response ('1' to '7') correspond to the areas of the distribution bounded by the thresholds. Arrows indicate the 'true' threshold locations from which the data was generated. The latent distribution is assumed to be normal with mean=0, SD=1.

One way of assessing a model's goodness-of-fit is the previously discussed predictive check, in which data samples are predicted from the fitted model and compared to the observed data. For Bayesian models, these checks can be easily conducted with the `pp_check()` function; we use bar plots because the models are discrete rather than continuous:

```
pp_check(m.equidistant, type="bars", nsamples=100)
pp_check(m.flexible, type="bars", nsamples=100)
```

The results are displayed in Figure 4 (in a single plot), and show that the equidistant model severely underestimates the number of '4' responses and overestimates the number of '3' and '5' responses. By contrast, the predictions of the flexible model are very close to the observed data. Another way of assessing models is to formally compare them, for example with cross-validation (Bürkner, 2017; Vasishth, Nicenboim, Beckman, Li, & Kong, 2018);

this also shows an advantage for the flexible model (see Appendix S1 in the Supplementary Material).

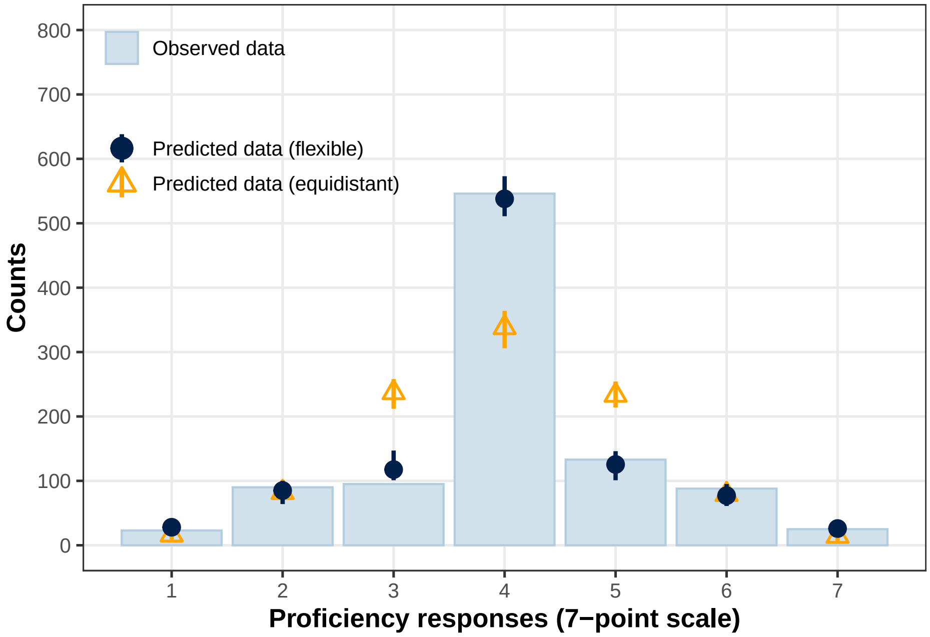

Figure 4. Predictive checks for ordinal models with flexible thresholds (full dots) and equidistant thresholds (empty triangles). These assessments of model quality involve comparing the probability/counts of different responses (given in a 7-point acceptability scale) in data predicted by the fitted models (dots and triangles) against the observed data that the models were fitted on (shown as bars).

From the various comparisons, we would conclude that flexible thresholds are necessary to appropriately model this dataset, and by extension, that the psychological distances between response values are not constant across the scale. In particular, the estimated threshold locations show that '5' responses actually express greater underlying proficiency than a metric 5 (and '3' responses express lower proficiency than a 3), so that the difference between a '3'-participant and a '5'-participant should be interpreted as particularly large. Such inferences about how the underlying construct relates to the observed responses are missed when metric models are used, and are an important advantage of ordinal models.

Effects of categorical predictors

To illustrate how the effects of predictors are estimated in ordinal models, we go back to Schlenter (2019)'s dataset, in which canonical and non-canonical sentences were rated on a 7-point acceptability scale. In thresholded-cumulative models, effects are estimated in the same way as in linear regression (i.e., by 'shifting' the mean of a normal distribution), but they take place at the latent variable, rather than on metric responses.

We fit a (mixed-effects) thresholded-cumulative model with condition (canonical, non-canonical) as predictor. Because each participant and item are associated with multiple responses, participant and item are included as random effects:

```
m.canonicity <- brm(Response ~ (1|Participant) + (1|Item) + Condition,
      data = acceptability.ratings,
      family = cumulative(link="probit", threshold="flexible"))
```

The effect of condition on acceptability is estimated as -0.68 [-0.83, -0.54] (see model summary in Table S1 in the Supplementary Material). The effect is expressed in SD units, which in this case can be interpreted as a standardized effect size, similar to Cohen's *d* (1988; Glass, McGaw, & Smith, 1981). Both the effect's magnitude and its narrow credible interval indicate a substantial difference between conditions, with lower acceptability for non-canonical sentences.

An important aid to interpretation is the examination of the model's *conditional effects*, that is, the predicted proportions of responses in each condition:

```
conditional_effects(m.canonicity, categorical=T)
```

The predictions indicate that canonical sentences have very high acceptability, with a large proportion of '7=very acceptable' responses (64% [52, 75]); also see Figure S1 in the Supplementary Material. The proportion of '6's is lower but still substantial (28% [20, 34]), and other responses are less frequent ('5' responses: 6% [3, 10]). In turn, the lower acceptability of non-canonical sentences is expressed by a much smaller proportion of '7's (37% [26, 49]), and by more lower-acceptability responses ('6' responses: 38% [34, 42]; '5's: 14% [10, 19]).

Such predictions are finer than those obtained from the linear model above (cf. '0.59 less acceptability'), because they are expressed as probabilities of discrete responses and not in the inappropriate metric scale. Note also that these proportions are not calculated directly from the data. Rather, they constitute better, more generalisable inferences, because they come from a model that: (a) estimates the effect of condition across all responses; (b) takes into account the whole structure of the data in terms of its participants and items; and (c) can potentially include other sources of information, like adjustment for covariates.

Finally, we assess the model's goodness-of-fit with a predictive check, displayed in Figure 2, panel b:

```
pp_check(m.canonicity, type="bars", nsamples=100)
```

Whereas the predictions of the linear model seriously mismatched the data (see above), the ordinal model fares remarkably well, and readily accommodates both the discreteness of responses and their non-normality.

Effects of continuous predictors

The usefulness of examining the model's conditional effects is particularly clear in the case of continuous predictors. To illustrate, we will make use of a dataset collected by Puebla (2016) in which 55 second language German speakers self-rated their speaking proficiency on a 7-point scale (responses covered only the four highest categories, 'functional', 'good', 'very good', and 'nativelike'). We will estimate the effect of age of acquisition (AoA) on self-ratings of proficiency (see, e.g., Hakuta, Bialystok, & Wiley, 2003).

Responses were coded as categories (e.g., 'good'), so we first establish their order, and then fit a thresholded-cumulative model with AoA as a continuous predictor:

```
proficiency.ratings$Response <- ordered(proficiency.ratings$Response,
      levels=c("functional", "good", "very good", "nativelike"))

m.aoa <- brm(Response ~ AoA, data = proficiency.ratings,
      family = cumulative(link="probit", threshold="flexible"))
```

The effect of AoA on the latent proficiency variable is estimated as -0.15 [-0.22, -0.09] per year (see model summary in Table S2 in the Supplementary Material). Thus, the effect across the whole AoA range (4–24 years) is 3 SDs on the latent scale, indicating a large difference between early and late bilinguals, with lower proficiency at later AoAs.

We plot the conditional effects of this model (Figure 5):[3]

```
  conditional_effects(m.aoa, categorical=T)
```

At the earlier AoAs there is a predominance of 'nativelike' responses. As AoA increases, 'nativelike' responses decline sharply and there is an increase in the proportions of the lower-proficiency categories (i.e., 'nativelike' responses are progressively replaced by 'very good' and 'good' responses). By an AoA of 13, all three top proficiency choices are likely responses, but there is a predominance of 'very good's. As AoA increases further (after puberty), the most common response becomes 'good' and some 'functional' responses start emerging.

These inferences are much more informative than what could be afforded by a linear model, because predictions (and their uncertainty) are appropriately expressed in terms of the different response categories, because predictions (and their uncertainty) are appropriately expressed in terms of the different response categories.

Conclusions

We have shown how Bayesian ordinal models are an appropriate solution for the analysis of rating scales, since they respect the discreteness of responses, conform to statistical assumptions, and allow modelling unequal psychological distances. This last aspect is

[3] In the case of many response categories, complexity may be reduced by plotting only a subset of interest or by averaging the predictions for several categories (Kissling, 2018).

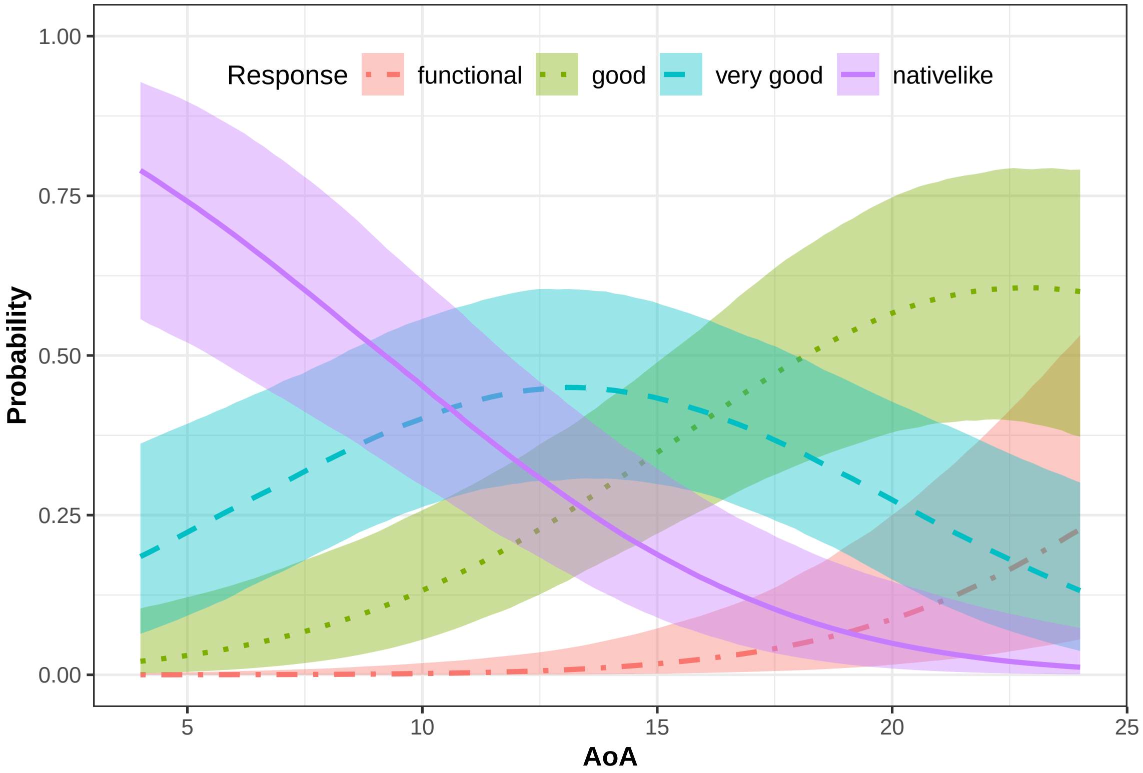

Figure 5. Effect of AoA on self-rated speaking proficiency. Separate predictions are plotted for each response category (i.e., different probabilities for ‘functional’, ‘good’, ‘very good’, ‘nativelike’). Given that responses are mutually exclusive, their predicted proportions add up to 100% at each AoA. Note that even though the model predicts an AoA effect for each response category, all predictions arise from a single (linear) effect of AoA on the latent proficiency variable. Shaded bands indicate 95% credible intervals. Data from Puebla (2016).

particularly relevant for estimating language proficiency, given that bilinguals of different groups can adopt different frames of reference in self-ratings (Tomoschuk, Ferreira, & Gollan, 2019). We note that ordinal models are not a panacea against all subjective biases in ratings; they should ideally be complemented by objective measures (Hulstijn, 2012; Lemhöfer & Broersma, 2012). Nevertheless, ordinal models can incorporate some of those biases by modelling how underlying constructs like proficiency map to ordered responses. As a result, they provide more valid and accurate inferences than the metric methods that are currently used in bilingualism research.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. R Package Version 0.1.0.9942.

Birdsong, D., Gertken, L. M., & Amengual, M. (2012). Bilingual Language Profile: An easy-to-use instrument to assess bilingualism. Measurement Instrument, https://sites.la.utexas.edu/bilingual/.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/gdm28w

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. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/gfxzpn

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

Cho, J., & Slabakova, R. (2014). Interpreting definiteness in a second language without articles: The case of L2 Russian. *Second Language Research*, *30*(2), 159–190. https://doi.org/ggktx6

Christensen, R. H. B. (2019). *ordinal - Regression models for ordinal data*. R Package Version 2019.12-10.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Second). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, *50*(1), 61–104. https://doi.org/ggktxk

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity*, *52*(4), 1523–1559. https://doi.org/gdqv89

Douven, I. (2018). A Bayesian perspective on Likert scales and central tendency. *Psychonomic Bulletin & Review*, *25*(3), 1203–1211. https://doi.org/gf5sjs

Farhy, Y., & Veríssimo, J. (2019). Semantic effects in morphological priming: The case of Hebrew stems. *Language and Speech*, *62*(4), 737–750. https://doi.org/ggkts3

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, *41*(1), 78–104. https://doi.org/dwfs84

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills: Sage Publications.

Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical evidence: A test of the critical-period hypothesis for second-language acquisition. *Psychological Science*, *14*(1), 31–38. https://doi.org/ffjrhs

Hopp, H. (2009). The syntax–discourse interface in near-native L2 acquisition: Off-line and on-line performance. *Bilingualism: Language and Cognition*, *12*(4), 463–483. https://doi.org/dk5xtw

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, *15*(2), 422–433. https://doi.org/ggktxv

Kissling, E. M. (2018). Pronunciation instruction can improve L2 learners' bottom-up processing for listening. *The Modern Language Journal*, *102*(4), 653–675. https://doi.org/gfkf3j

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & James D. (Eds.), *Handbook of survey research* (Second edition). Bingley, UK: Emerald.

Kuncel, R. B. (1977). The subject-item Interaction in itemmetric research. *Educational and Psychological Measurement*, *37*(3), 665–678. https://doi.org/bjrbmt

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/c9f897

Li, P., Sepanski, S., & Zhao, X. (2006). Language History Questionnaire: A Web-based interface for bilingual research. *Behavior Research Methods*, *38*(2), 202–210. https://doi.org/fph9q8

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

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/gcx746

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 109–127. https://doi.org/ggntw8

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (Second). CRC Press.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, *4*(1), 103–120. https://doi.org/dqfhpp

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

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods, 26*(1), 103–126. https://doi.org/ghbtt6

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language, 110*, 104038. https://doi.org/gf9tjp

Schlenter, J. (2019). *Predictive language processing in late bilinguals* (Doctoral Dissertation). Universität Potsdam. https://doi.org/ffhz

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93–105. https://doi.org/fqrx56

Tare, M., Golonka, E., Lancaster, A. K., Bonilla, C., Doughty, C. J., Belnap, R. K., & Jackson, S. R. (2018). The role of cognitive aptitudes in a study abroad language-learning environment. In C. Sanz & A. Morales-Front (Eds.), *The Routledge Handbook of Study Abroad Research and Practice* (First, pp. 406–420). Routledge. https://doi.org/gj9s

Tomoschuk, B., Ferreira, V. S., & Gollan, T. H. (2019). When a seven is not a seven: Self-ratings of bilingual language proficiency differ between and within language populations. *Bilingualism: Language and Cognition*, *22*(3), 516–536. https://doi.org/gfkm58

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

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111–125. https://doi.org/f5t93s

Supplementary Material for Veríssimo (2021, *BLC*)

João Veríssimo

24 Jun, 2021

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

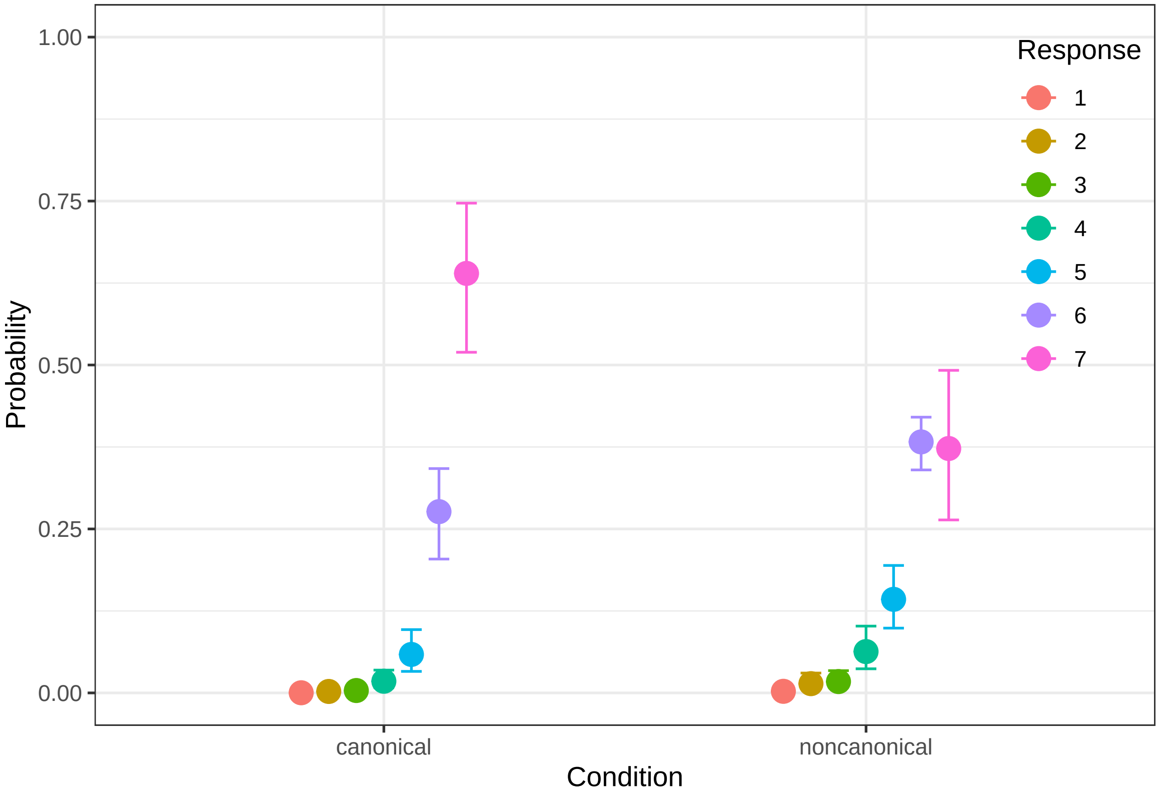

Figure S1. Conditional effects of condition (canonical, non-canonical) on sentence acceptability. Given that responses are mutually exclusive, their predicted proportions add up to 100% in each condition. Error bars indicate 95% credible intervals. Data from Schlenter (2019).

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, 2021; 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 for Supplementary Material

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. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103–126. https://doi.org/ghbtt6

Schlenter, J. (2019). *Predictive language processing in late bilinguals* (Doctoral Dissertation). Universität Potsdam. https://doi.org/ffhz

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