Supplemental materials

1. Post-hoc power analyses

We determined the power of our studies by bootstrapping our experimental data for each key prediction. All code for these post-hoc power analyses is available at https://osf.io/9hw68/. All our key predictions (see below) relied on differences between conditions. We analyzed these differences by computing bootstrapped 95% confidence intervals over the difference in effect sizes and seeing if the interval crossed 0 (suggesting there may be no difference in effect sizes) or not.

The bootstrap process followed two stages. First, we bootstrapped the data to obtain a replicate of the two target experimental conditions (depending on prediction; see below). Next, for each experimental replicate, we conducted the analysis from the manuscript: bootstrapping the results to obtain a confidence interval over the difference in effect sizes. We ran 5000 experimental replicates and within each replicate we bootstrapped the difference in effect sizes with 1000 replicates.

Experiment 1

Prediction 1. Percentage of redundant color words should be different in the four-item displays across languages.

In 94.9% of the bootstrapped replicates, the corresponding confidence interval did not cross 0, suggesting that the two effects were reliably different and that our experiment's power is 0.949.

Prediction 2. Percentage of redundant color words in English should be higher in the 16-item displays relative to the four-item displays.

In 92.0% of the bootstrapped replicates, the corresponding confidence interval did not cross 0, suggesting that the two effects were reliably different and that our experiment's power is 0.92.

Prediction 3. Percentage of redundant color words in Spanish should be higher in the 16-item displays relative to the four-item displays.

In 100% of the bootstrapped replicates, the corresponding confidence interval did not cross 0, suggesting that the two effects were reliably different. Naturally, this does not imply that our power is 1, as it is likely that at least one replicate would eventually not produce a difference. Note, however, that the difference across conditions was striking in Spanish (see Figure 1S below), with almost no Spanish speaker producing color words in the four-item display, and the majority of them producing color words more than half of the time in the 16-item display. Given that we found no replicates that went against our predictions in the 5000 bootstrap samples, this suggests that the probability of a replicate not showing our predicted effect is lower than 1/500 = 1e-04.

Experiment 2a

Prediction 1. English listeners should show increased target fixations during the adjusted NP in the shape competitor trials relative to the color competitor trials.

In 100% of the bootstrapped replicates, the corresponding confidence interval did not cross 0, suggesting that the two effects (fixations on target in shape competitor trials, and fixation on target in color competitor trials) were different. Given that we ran 5000 bootstrapped samples, this suggests that the probability of a replicate not showing our predicted effect is lower than 1/5000=1e-04.

Prediction 2. Spanish listeners tested in Spanish should show increased target fixations during the adjusted NP in the color competitor trials relative to the shape competitor trials.

Similar to Prediction 1, in 100% of the bootstrapped replicates, the corresponding confidence interval did not cross 0, suggesting that the two effects were different. Given that we ran 5000 bootstrapped samples, this suggests that the probability of a replicate not showing our predicted effect is lower than 1/5000=1e-04.

Experiment 2b

Prediction 1. Spanish listeners tested in English should show increased target fixations during the adjusted NP in the shape competitor trials relative to the color competitor trials.

In 100% of the bootstrapped replicates, the corresponding confidence interval once again did not cross 0, suggesting that the two effects (fixations on target in shape competitor trials, and fixation on target in color competitor trials) were different. Given that we ran 5000 bootstrapped samples, this suggests that the probability of a replicate not showing our predicted effect is lower than 1/5000=1e-04.

2. Supplemental figures for main analyses

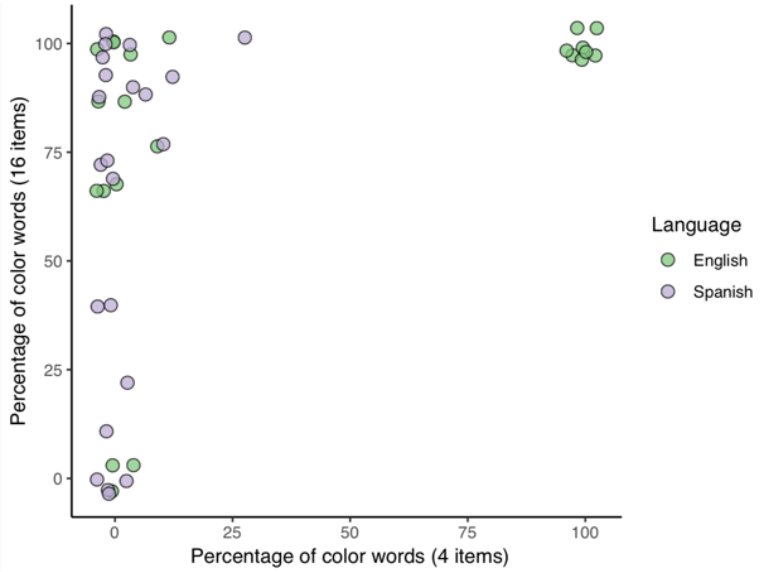

Figure 1: Percentage of redundant color adjectives used by each subject (represented by a dot). The x axis shows the percentage of color words that subjects produced in 4-shape displays and the y axis shows the percentage of color words that they produced in 16-shape displays. The color of the dot indicates the speaker's language. Data have been minimally jittered to avoid overplotting.

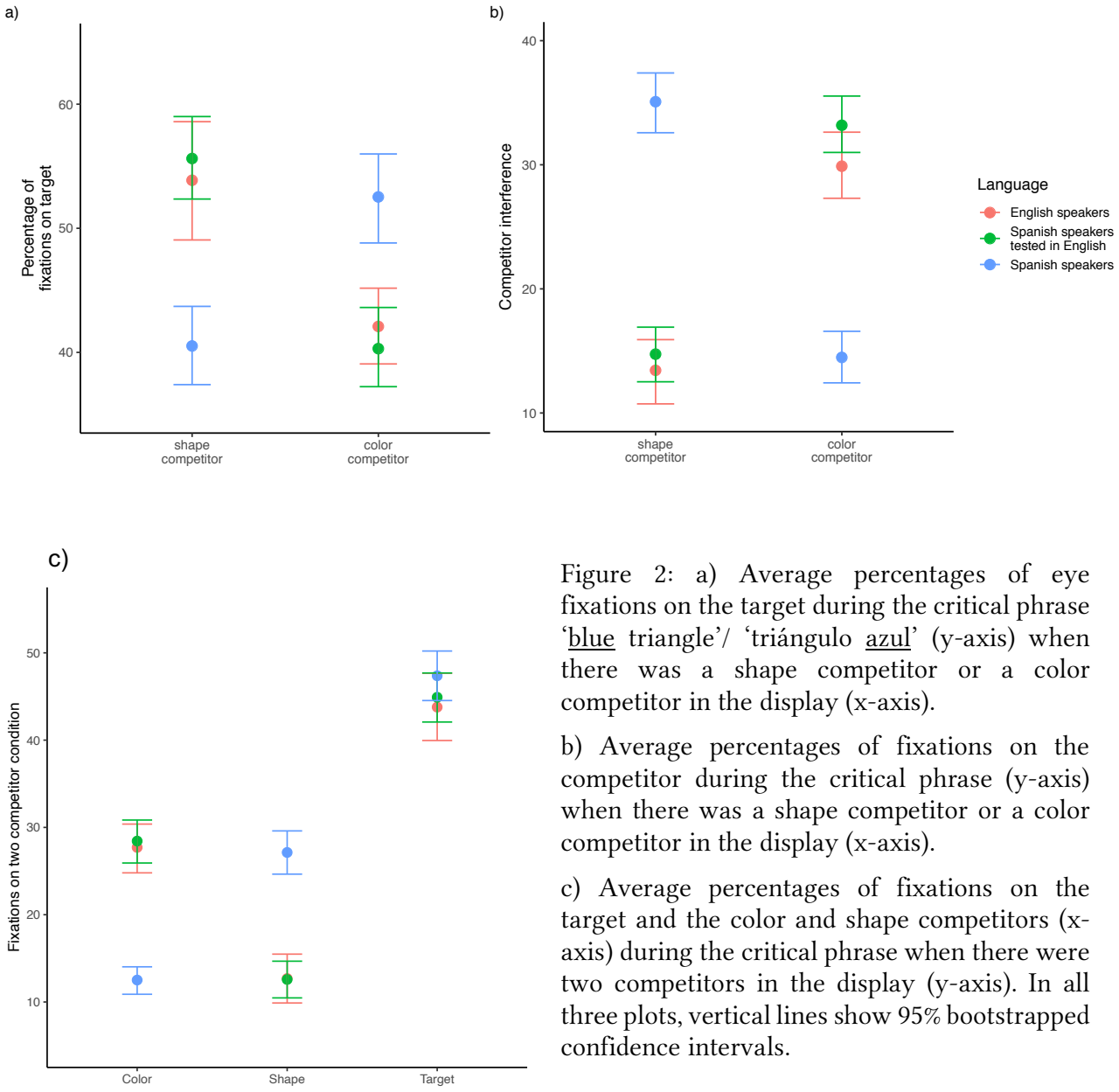

Figure 2: a) Average percentages of eye fixations on the target during the critical phrase 'blue triangle'/ 'triángulo azul' (y-axis) when there was a shape competitor or a color competitor in the display (x-axis).

b) Average percentages of fixations on the competitor during the critical phrase (y-axis) when there was a shape competitor or a color competitor in the display (x-axis).

c) Average percentages of fixations on the target and the color and shape competitors (x-axis) during the critical phrase when there were two competitors in the display (y-axis). In all three plots, vertical lines show 95% bootstrapped confidence intervals.

3 Regression analyses

3.1 Experiment 1

The tables below present the result of a logistic mixed-effects model [1, 2] predicting participant's use of redundant color words as a function of number of items in the display (coded numerically, either four or sixteen) and language (dummy-coded, with English set to 0 and Spanish to 1). We also included the maximal random effects structure:

$$Target \sim Items * Language + (1 + Items|Participant)$$

Note that we do not include language-based slopes per participant because language varies across participants.

Scaled residuals:				
Min	1Q	Median	3Q	Max
-3.5538	-0.0524	-0.0001	0.1037	3.5441

Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
Subject	(Intercept)	118.7392	10.8968	
	Items	0.2717	0.5213	-0.95

Fixed effects:				
	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-5.2552	3.1689	-1.658	0.09724 .
Items	0.5742	0.1768	3.247	0.00116 **
Language(Sp)	-10.6457	5.4160	-1.966	0.04934 *
Items:Language(Sp)	0.4653	0.3210	1.450	0.14713

3.2 Experiment 2

The following tables show the results of two logistic mixed-effects models [1, 2] comparing English vs. Spanish listeners (Experiment 2a), and comparing Spanish listeners tested in English vs Spanish listeners tested in Spanish (Experiment 2b). In each case we identified the largest maximal model that converged through the buildmer package [3].

In each trial we considered only the NP window (adjusted by 200 milliseconds), and we excluded trials where participants failed to select the correct target. Time was standardized by subtracting the mean and dividing by two times the standard deviations in each trial. Language and condition were both sum-coded (see individual regressions for details).

3.2.1 Experiment 2a

We first considered the maximal logical model:

$$\begin{aligned} Target \sim & Time * Language * Cond + \\ & (1 + Time + Cond | Participant) + \\ & (1 + Time | Item) \end{aligned}$$

In the language variable, English was coded as 0.5 and Spanish as -0.5. In the condition variable, the Shape Competitor condition was coded as 0.5 and the Color Competitor condition as -0.5 [1]. The largest model that converged was

$$\begin{aligned} Target \sim & Time * Language * Cond + \\ & (1 | Participant) + \\ & (1 + Time | Item) \end{aligned}$$

Our predictions focused on the speed of identifying the target and we thus focused on how language type and condition interacted with time.

Scaled residuals:				
Min	1Q	Median	3Q	Max
-4.9940	-0.7506	-0.3079	0.7679	6.1587

Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
Participant	(Intercept)	0.1578	0.3973	
Item	(Intercept)	0.2040	0.4516	
	Time	0.3344	0.5783	-0.29

Fixed effects:					
	Estimate	Std. Error	z value	Pr($>\|z\|$)	Pr($>\|t\|$)
(Intercept)	-0.132697	0.107819	-1.230735	0.218	0.21842
Time	1.850445	0.119071	15.540725	0.000	$< 2e-16$ ***
Cond	-0.009818	0.184884	-0.053103	0.958	0.95765
Language	0.091396	0.112124	0.815133	0.415	0.41500
Time:Cond	0.720336	0.237137	3.037636	0.002	0.00238 **
Cond:Language	1.178885	0.026640	44.253191	0.000	$< 2e-16$ ***
Time:Language	-0.010657	0.028990	-0.367623	0.713	0.71315
Time:Cond:Language	1.783963	0.057598	30.972843	0.000	$< 2e-16$ ***

The significant positive effect of the interaction between condition and language ($\hat{\beta}_{C:L} = 1.18$), and time, condition and language ($\hat{\beta}_{T:C:L} = 1.78$), show

evidence for our hypothesis. This means that English listeners in the Shape Competitor condition (both coded as 0.5, creating interactions of $0.25 * \hat{\beta}_{C:L}$ and $0.25 * \hat{\beta}_{T:C:L}$) and Spanish listeners in the Color Competitor condition (both coded as -0.5, creating the same interaction effects) identified the target significantly faster than English listeners in the Color Competitor condition and Spanish listeners in the Shape Competitor condition (interaction effects = $-0.25 * \hat{\beta}_{C:L}$ and $-0.25 * \hat{\beta}_{T:C:L}$ in both cases).

3.2.2 Experiment 2b

In the regression for Experiment 2b we compared Spanish listeners tested in Spanish and English.

We first considered the maximal logical model:

$$\begin{aligned} Target \sim & Time * Language * Cond + \\ & (1 + Time + Cond + Language | Participant) + \\ & (1 + Time | Item) \end{aligned}$$

Notice that this maximal model now included random language slopes per participant. This was now possible because the same participants completed the task in both languages. In the language variable, English was coded as 0.5 and Spanish as -0.5. In the condition variable, the Shape Competitor condition was coded as 0.5 and the Color Competitor condition as -0.5 [1].

The largest model that converged was

$$\begin{aligned} Target \sim & Time * Language * Cond + \\ & (1 + Time + Cond | Participant) + \\ & (1 + Time | Item) \end{aligned}$$

Our predictions focused on the speed of identifying the target and we thus focused on how language type and condition interacted with time.

Scaled residuals:				
Min	1Q	Median	3Q	Max
-5.1131	-0.7543	-0.3568	0.7809	5.7258

The significant positive effect of the interaction between condition and language ($\hat{\beta}_{C:L} = 1.35$), and time, condition and language ($\hat{\beta}_{T:C:L} = 1.55$), show evidence for our hypothesis. This means that when tested in English in the Shape Competitor condition (both coded as 0.5, creating interactions of $0.25 * \hat{\beta}_{C:L}$ and $0.25 * \hat{\beta}_{T:C:L}$) and in Spanish in the Color Competitor condition (both coded as -0.5, creating the same interaction effects), participants identified the target significantly faster than when tested in English in the Color Competitor condition and in Spanish in the Shape Competitor condition (interaction effects = $-0.25 * \hat{\beta}_{C:L}$ and $-0.25 * \hat{\beta}_{T:C:L}$ in both cases).

Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
Participant	(Intercept)	0.04886	0.2210	
	Cond	0.12844	0.3584	-0.29
	Time	0.04669	0.2161	0.48 -0.02
Item	(Intercept)	0.14438	0.3800	
	Time	0.28916	0.5377	-0.11

Fixed effects:					
	Estimate	Std. Error	z value	Pr($>\|z\|$)	Pr($>\|t\|$)
(Intercept)	-0.12984	0.08901	-1.45875	0.145	0.1446
Time	1.78146	0.11839	15.04720	0.000	$< 2e-16$ ***
Language	0.08839	0.01320	6.69488	0.000	$2.16e-11$ ***
Cond	0.09969	0.17007	0.58616	0.558	0.5578
Time:Language	-0.05508	0.02848	-1.93438	0.053	0.0531 .
Language:Cond	1.35165	0.02641	51.17822	0.000	$< 2e-16$ ***
Time:Cond	0.54240	0.22056	2.45924	0.014	0.0139 *
Time:Language:Cond	1.54676	0.05659	27.33337	0.000	$< 2e-16$ ***

3.3 Analysis of the Two Competitors vs Color Competitor conditions

The Two Competitors condition was used as filler trials intended to add variability to the types of displays used in the study. However, previous studies have compared the results of this condition with those of the Color Competitor condition in order to investigate the derivation of contrastive inferences in English (Sedivy, 2003, 2004; Aparicio et al., 2016; Rubio-Fernandez et al., under review). We report additional analyses of these two conditions for completeness, although they do not directly test our key hypotheses.

Previous studies comparing the Two Competitors vs Color Competitor conditions used long preview windows that allowed English listeners to anticipate the noun in the Two Competitors condition through the derivation of a contrastive inference (see Rubio-Fernandez et al. (under review) for discussion). Since we used short preview windows of 400ms in all our conditions, we expected weaker results than previous studies. That is, we suspected that English listeners in Experiment 2a and Spanish listeners tested in English in Experiment 2b may not have sufficient preview time to derive an anticipatory inference in the Two Competitors condition.

We analyzed English listeners and Spanish listeners tested in English on the Two Competitors and the Color Competitor conditions. We considered the maximal logical model:

$$Target \sim Time * Language * Cond + (1 + Time|Participant) + (1 + Time|Item)$$

In the language variable, English listeners were coded as -0.5 and Spanish listeners tested in English were coded as 0.5. In the condition variable, the filler trials were coded as -0.5 and the Color Competitor condition as 0.5 [1]. The largest model that converged was

$$Target \sim Time * Language * Cond + (1|Participant) + (1 + Time|Item)$$

Scaled residuals:				
Min	1Q	Median	3Q	Max
-2.9615	-0.8065	-0.4493	0.9054	6.5398

Random effects:				
Groups	Name	Variance	Std.Dev.	Corr
Participant	(Intercept)	0.1184	0.3441	
Item	(Intercept)	0.1685	0.4105	
	Time	0.4117	0.6416	-0.28

Fixed effects:					
	Estimate	Std. Error	z value	Pr($>\|z\|$)	Pr($>\|t\|$)
(Intercept)	-0.36207	0.09712	-3.72817	0.000	0.000193 ***
Time	1.21279	0.13194	9.19220	0.000	$< 2e-16$ ***
Cond	0.14696	0.16828	0.87334	0.382	0.382476
Language	0.01158	0.09826	0.11788	0.906	0.906164
Time:Cond	0.25449	0.26487	0.96078	0.337	0.336661
Cond:Language	-0.13876	0.02718	-5.10531	0.000	$3.30e-07$ ***
Time:Language	0.08668	0.02829	3.06369	0.002	0.002186 **
Time:Cond:Language	0.37164	0.05644	6.58502	0.000	$4.55e-11$ ***

The additional analyses of the Two Competitors vs Color Competitor conditions showed that English listeners did not reveal sensitivity to pragmatic contrast in the Two Competitors condition. We explain this lack of an effect as a result of the short preview window used in our study.

Contrary to what we observed with English listeners, Spanish listeners tested in English revealed a target advantage in the Two Competitors condition relative to the Color Contrast condition. This pattern of results confirms that

pragmatic contrast can affect real-time language processing in some visual contexts (see Sedivy, 2003, 2004; Aparicio et al., 2016; Rubio-Fernandez et al., under review). However, we interpret the target advantage observed in Experiment 2b as a familiarity effect since Spanish participants had interpreted color adjectives contrastively when first tested in Spanish in Experiment 2a (see Fig. 4 and eye-tracking analyses in the main text). We suppose that this first experience with the task may have allowed Spanish listeners to derive contrastive inferences in Experiment 2b, despite the short preview window.

In summary, the results of the Two Competitors vs Color Competitor conditions seem to suggest that, without a sufficiently long preview window or some experience with the task, pragmatic contrast may not affect real-time processing in this paradigm. Future studies should further investigate these open questions.

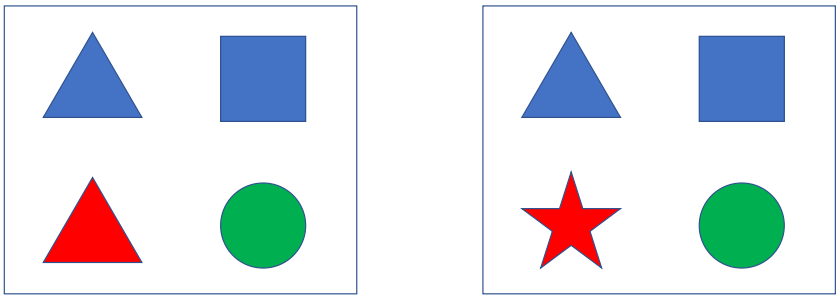

Figure 3: Sample displays.

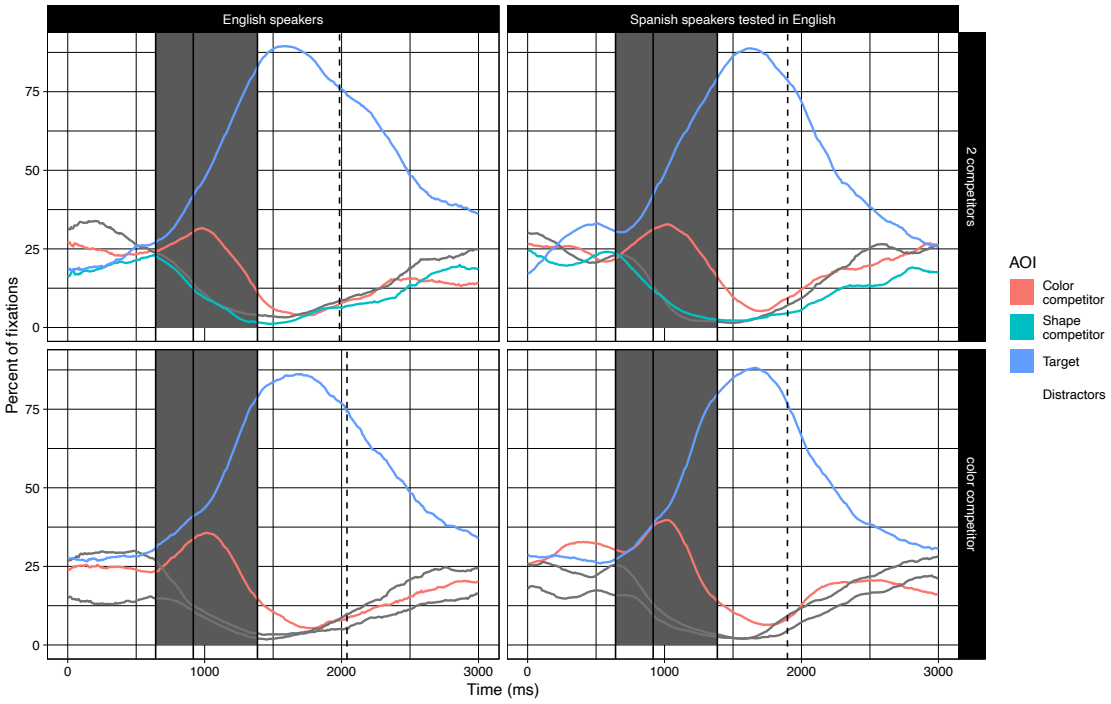

Figure 4: Supplemental looking plots.

References

[1] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[2] T Florian Jaeger. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, 59(4):434–446, 2008.

[3] Cesko C. Voeten. *buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression*, 2020. R package version 1.5.