Supplementary materials for “Are there factive predicates? An empirical investigation”

Judith Degen

Department of Linguistics
Stanford University
450 Serra Mall
Stanford, CA 94305, USA
jdegen@stanford.edu

Judith Tonhauser

Department of Linguistics
University of Stuttgart
Keplerstr. 17
70174 Stuttgart, Germany
judith.tonhauser@ling.uni-stuttgart.de

A. 20 COMPLEMENT CLAUSES. The following clauses realized the complements of the predicates in Exps. 1, 2, and 3:

1. Mary is pregnant.
2. Josie went on vacation to France.
3. Emma studied on Saturday morning.
4. Olivia sleeps until noon.
5. Sophia got a tattoo.
6. Mia drank 2 cocktails last night.
7. Isabella ate a steak on Sunday.
8. Emily bought a car yesterday.
9. Grace visited her sister.
10. Zoe calculated the tip.
11. Danny ate the last cupcake.
12. Frank got a cat.
13. Jackson ran 10 miles.
14. Jayden rented a car.
15. Tony had a drink last night.
16. Josh learned to ride a bike yesterday.
17. Owen shoveled snow last winter.
18. Julian dances salsa.
19. Jon walks to work.
20. Charley speaks Spanish.

B. CONTROL STIMULI IN EXPS. 1, 2, AND 3.

(1) Control stimuli in Exps. 1

  a. Is Zack coming to the meeting tomorrow?

  b. Is Mary's aunt sick?

  c. Did Todd play football in high school?

  d. Is Vanessa good at math?

  e. Did Madison have a baby?

  f. Was Hendrick's car expensive?

(2) Control stimuli in Exps. 2

  a. Entailing control stimuli

    i. **What is true:** Frederick managed to solve the problem. (Tested inference: Frederick solved the problem.)

    ii. **What is true:** Zack bought himself a car this morning. (Tested inference: Zack owns a car.)

    iii. **What is true:** Tara broke the window with a bat. (Tested inference: The window broke.)

    iv. **What is true:** Vanessa happened to look into the mirror. (Tested inference: Vanessa looked into the mirror.)

  b. nonentailing control stimuli

i. **What is true:** Dana watched a movie last night. (Tested inference: Dana wears a wig.)

ii. **What is true:** Hendrick is renting an apartment. (Tested inference: The apartment has a balcony.)

iii. **What is true:** Madison was unsuccessful in closing the window. (Tested inference: Madison closed the window.)

iv. **What is true:** Sebastian failed the exam. (Tested inference: Sebastian did really well on the exam.)

(3) Control stimuli in Exps. 3

a. Contradictory control stimuli

i. Madison laughed loudly and she didn't laugh.

ii. Dana has never smoked in her life and she stopped smoking recently.

iii. Hendrick's car is completely red and his car is not red.

iv. Sebastian lives in the USA and has never been to the USA.

b. noncontradictory control stimuli

i. Vanessa is really good at math, but I'm not.

ii. Zack believes that I'm married, but I'm actually single.

iii. Tara wants me to cook for her and I'm a terrific cook.

iv. Frederick is both smarter and taller than I am.

C. DATA EXCLUSION. Table A1 presents how many participants' data were excluded from the analysis based on the exclusion criteria. The first column records the experiment, the second ('recruited') how many participants were recruited, and the final column ('remaining') how many participants' data entered the analysis. The 'Exclusion criteria' columns show how many participants' data were excluded based on the four exclusion criteria:

- 'multiple': Due to an experimental glitch, some participants participated in Exps. 1b, 2b or 3b more than once. Of these participants, we only analyzed the data from the first time they participated.
- 'language': Participants' data were excluded if they did not self-identify as native speakers of American English.
- 'controls': Participants' data were excluded if their response mean on the 6 control items was more than 2 sd above the group mean (Exp. 1a), if they gave a wrong rating ('yes') to more than one of the six controls (Exp. 1b), if their response mean on the entailing or the nonentailing controls was more than 2 sd below or above, respectively, the group means

(Exp. 2a), if they gave more than one wrong rating to one of the eight controls, where a wrong rating is a 'yes' to a nonentailing control and a 'no' to an entailing one (Exp. 2b), if their response means on the contradictory or noncontradictory controls were more than 2 sd below or above, respectively, the group means (Exp. 3a), and if they gave more than one wrong response to one of the eight control sentences, where a wrong response was a 'yes' to a noncontradictory control or a 'no' to a contradictory one (Exp. 3b).

- 'variance': Participants' data were excluded if they always selected roughly the same point on the response scale, that is, if the variance of their response distribution was more than 2 sd below the group mean variance.

		Exclusion criteria				
	recruited	multiple	language	controls	variance	remaining
Exp. 1a	300	n.a.	13	16	5	266
Exp. 1b	600	75	43	46	n.a.	436
Exp. 2a	300	n.a.	14	27	0	259
Exp. 2b	600	169	35	21	n.a.	375
Exp. 3a	300	n.a.	19	18	0	263
Exp. 3b	600	170	30	47	n.a.	353

TABLE A1. Data exclusion in Exps. 1, 2, and 3

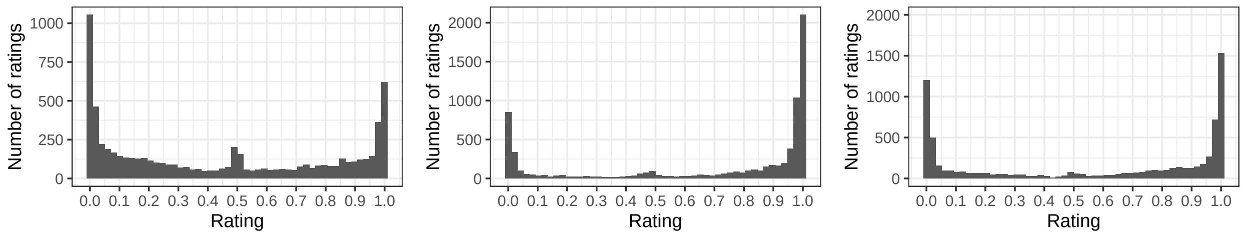
D. MODEL DETAILS FOR EXPERIMENTS 1, 2, AND 3. This supplement provides details on the data analysis conducted for Exps. 1, 2, and 3. We first motivate the use of Beta regression rather than linear regression in Exps. 1a, 2a, and 3a (section D.1) and then provide a brief primer on how to interpret Bayesian mixed effects Beta regression models (section D.2). We then report the model outputs for Exps. 1, 2, and 3 (section D).

D.1. MOTIVATION FOR USING BAYESIAN MIXED EFFECTS BETA REGRESSION. There are three separate pieces to motivate: the use of mixed effects, the use of Bayesian rather than frequentist models, and the use of Beta regression rather than linear regression.

Using mixed effects refers to the practice of modeling the outcome variable, here slider ratings or proportions of 'yes' ratings, as a function of not just fixed effects of interest (i.e. predicate) but also as the result of possible random variability that is not of theoretical interest (e.g. random by-participant or by-item variability, Gelman & Hill 2006). This is standard practice in psycholinguistic studies and allows the researcher to trust that any observed effects of theoretical interest are true average effects rather than the result of idiosyncratic behavior (e.g. of participants or items).

Using Bayesian models rather than frequentist models is increasingly becoming the norm in psycholinguistic studies as computational power has increased and running Bayesian models has become more accessible with the introduction of R packages such as `brms` (Bürkner 2017). The presence of an effect in frequentist models is evaluated by checking whether the p-value is smaller than $.05$, where the p-value is defined as the probability of obtaining data that is as skewed or more skewed than the observed data if the null-hypothesis was true, that is, if the hypothesized effect was absent. Parameter estimates in frequentist models are obtained via maximum-likelihood techniques, that is, by estimating the parameter values that maximize the probability of observing the data. Bayesian models, by contrast, return a full posterior distribution over parameter values that take into account not just the probability of the data under the parameter values, but also the prior probability of parameter values. In order to evaluate the evidence for an effect of a predictor of interest, one common practice is to report 95% credible intervals and the posterior probability that the predictor coefficient β is either lower or greater than zero, $P(\beta < 0)$ or $P(\beta > 0)$, depending on the direction of the expected effect. A 95% credible interval (CI) demarcates the range of values that comprise 95% of probability mass of the posterior beliefs such that no value inside the CI has a lower probability than any point outside it (Jaynes & Kempthorne 1976, Morey et al. 2016). There is substantial evidence for an effect if zero is (by a reasonably clear margin) not included in the 95% CI, and $P(\beta > 0)$ or $P(\beta < 0)$ is close to zero or one. Posterior probabilities indicate the probability that the parameter has a certain value, given the data and model—these probabilities are thus *not* frequentist p-values. In order to present statistics as close to widely used frequentist practices, and following Nicenboim & Vasishth 2016, we defined an inferential criterion that seems familiar (95%), but the strength of evidence should not be taken as having clear cut-off points (such as in a null-hypothesis significance testing framework).

Using Beta regression rather than linear regression was motivated by the violation of two of the assumptions of linear regression: first, that residuals be normally distributed (where 'residuals' refers to the residual error for each data point after fitting the model), and second, that the error term exhibit homoscedasticity (that it be roughly the same across different conditions). Slider ratings data has the property of being bounded by its endpoints (which we code as 0 and 1, respectively). This often leads to 'bunching' behavior at the endpoints (see Figure A1 for the distribution of raw ratings in Exps. 1a, 2a, and 3a).



(A) Exp. 1a ratings. (B) Exp. 2a ratings. (C) Exp. 3a ratings.

FIGURE A1. Histograms of raw slider ratings in Exps. 1a, 2a, and 3a.

This 'bunching' behavior, in turn, can lead to the violation of both of the above assumptions of linear regression. Intuitively, these assumptions are violated because conditions that elicit ratings closer to endpoints necessarily have a compressed variance; consequently, a condition's mean and its variance are not independent. Beta regression is useful here because it allows for modeling an arbitrarily distributed outcome variable in the [0,1] interval. The Beta distribution is characterized by two parameters, one capturing the mean μ of the distribution and one capturing its precision ϕ, a measure of dispersion. The greater the precision, the more concentrated the values are around the mean, that is, the lower the variance of the distribution. We follow Smithson & Verkuilen 2006 in modeling μ and ϕ separately for each predictor. That is, we allow each predictor to affect both the mean and the precision of the outcome variable's distribution.

D.2. Coding choices and interpreting model output. The outcome variable in Exps. 1a, 2a, and 3a (slider ratings) contained the values 0 and 1, which Beta regression is undefined for. We therefore applied a common transformation to ratings before the main analysis that rescales values y to fall in the open unit interval (0,1) (Smithson & Verkuilen 2006). First, we apply $y' = (y - a)/(b - a)$, where b is the highest possible slider rating and a is the smallest possible slider rating. The range is then compressed to not include 0 and 1 by applying $y' = [y'(N-1)+1/2]/N$, where N is the total number of observations.

The mean parameter μ is modeled via a logit link function (default for Beta regression in `brms`), though other links that squeeze μ into the $[0,1]$ interval are possible. The dispersion parameter ϕ is modeled via a log link, which ensures that values of ϕ are strictly positive, which is necessary because a variance cannot be negative.

We allowed both μ and ϕ to vary as a function of predicate, with reference level set to main clause control in Exp. 1a, entailing control in Exp. 2a, and contradictory control in Exp. 3a. We also allowed random intercept adjustments to each parameter by participant and by item, where item was defined as a unique combination of a predicate and a complement clause. Four chains converged after 2000 iterations each (warmup = 1000, $\hat{R} = 1$ for all estimated parameters) with a target acceptance rate of .95 and a maximum treedepth of 15.

D.3. Model outputs for Experiments 1, 2, and 3. The three tables in this section show the model outputs for Exps. 1, 2, and 3, respectively: Table A2 for Exps. 1a and 1b, Table A3 for Exps. 2a and 2b, and Table A4 for Exps. 3a and 3b. Each table shows maximum a posteriori (MAP) model estimates for certainty ratings from the Beta regression model (left and middle column, mean μ and precision ϕ) and the logistic regression model (right column, β) with 95% credible intervals.

TABLE A2. Maximum a posteriori (MAP) model estimates for certainty ratings from Exp. 1a (left and middle column, mean μ and precision ϕ) and Exp. 1b (right column, β) with 95% credible intervals. Contrast of each predicate is with main clause control reference level. MAP estimates for which there is no evidence that they are different from 0 are italicized.

	Exp. 1a: Beta regression		Exp. 1b: logistic regression
Predictor	Estimated μ	Estimated ϕ	Estimated β
Intercept	−1.88 [−1.98; −1.78]	1.16 [1.02; 1.29]	−6.37 [−7.10; −5.72]
acknowledge	2.62 [2.45; 2.80]	−0.67 [−0.86; −0.49]	8.00 [7.31; 8.78]
admit	2.45 [2.29; 2.61]	−0.65 [−0.83; −0.48]	7.29 [6.60; 8.04]
announce	2.14 [1.98; 2.30]	−0.74 [−0.92; −0.57]	6.70 [6.04; 7.45]
be annoyed	3.56 [3.37; 3.75]	−0.30 [−0.51; −0.10]	9.41 [8.64; 10.22]
be right	0.52 [0.36; 0.68]	−0.21 [−0.40; −0.01]	2.33 [1.50; 3.21]
confess	2.29 [2.13; 2.45]	−0.71 [−0.88; −0.54]	6.75 [6.07; 7.50]
confirm	1.31 [1.15; 1.46]	−0.46 [−0.64; −0.28]	4.27 [3.60; 5.01]
demonstrate	1.78 [1.63; 1.93]	−0.59 [−0.76; −0.41]	5.31 [4.64; 6.06]
discover	2.90 [2.72; 3.07]	−0.67 [−0.85; −0.48]	8.46 [7.75; 9.25]
establish	1.42 [1.26; 1.58]	−0.55 [−0.73; −0.38]	4.51 [3.83; 5.26]
hear	2.71 [2.53; 2.88]	−0.79 [−0.98; −0.61]	8.22 [7.50; 9.01]
inform	3.00 [2.82; 3.18]	−0.60 [−0.79; −0.41]	9.13 [8.38; 9.94]
know	3.37 [3.18; 3.56]	−0.49 [−0.70; −0.29]	9.72 [8.94; 10.57]
pretend	0.32 [0.15; 0.49]	−0.19 [−0.38; −0.00]	3.22 [2.50; 4.03]
prove	1.16 [1.01; 1.31]	−0.21 [−0.40; −0.03]	3.92 [3.21; 4.67]
reveal	2.59 [2.42; 2.76]	−0.72 [−0.90; −0.55]	7.41 [6.72; 8.16]
say	0.78 [0.63; 0.93]	*-0.12* [−0.30; 0.06]	3.10 [2.35; 3.93]
see	3.01 [2.83; 3.19]	−0.70 [−0.90; −0.51]	8.74 [8.03; 9.55]
suggest	0.79 [0.63; 0.94]	*-0.17* [−0.36; 0.02]	3.22 [2.49; 3.99]
think	0.62 [0.47; 0.78]	*0.01* [−0.17; 0.20]	2.54 [1.75; 3.40]

TABLE A3. Maximum a posteriori (MAP) model estimates for inference ratings from Exp. 2a (left and middle column, mean μ and precision ϕ) and Exp. 2b (right column, β) with 95% credible intervals. Contrast of each predicate is with entailing control reference level. MAP estimates for which there is no evidence that they are different from 0 are italicized. Predicates are marked for which there is no evidence that they differ from entailing controls (bold: no difference on either categorical or continuous measure; italics: no difference on at least one measure).

	Exp. 2a: Beta regression		Exp. 2b: logistic regression
Predictor	Estimated μ	Estimated ϕ	Estimated β
Intercept	3.00 [2.87; 3.15]	2.24 [2.05; 2.42]	6.20 [5.52; 6.98]
acknowledge	−0.74 [−0.94; −0.54]	−0.64 [−0.89; −0.39]	−1.58 [−2.47; −0.70]
admit	−0.67 [−0.87; −0.48]	−0.52 [−0.77; −0.27]	−1.83 [−2.72; −0.97]
announce	−1.51 [−1.71; −1.31]	−1.35 [−1.59; −1.10]	−4.39 [−5.17; −3.68]
be annoyed	−0.55 [−0.76; −0.35]	−0.50 [−0.75; −0.23]	−1.36 [−2.31; −0.46]
be right	*-0.03* [−0.22; 0.16]	*0.09* [−0.16; 0.34]	*0.36* [−0.93; 1.89]
confess	−0.85 [−1.05; −0.66]	−0.64 [−0.90; −0.38]	−2.79 [−3.60; −2.04]
confirm	−0.22 [−0.41; −0.03]	*-0.00* [−0.26; 0.26]	*-0.85* [−1.84; 0.14]
demonstrate	−1.23 [−1.44; −1.02]	−1.18 [−1.44; −0.93]	−3.35 [−4.15; −2.63]
discover	−0.27 [−0.46; −0.08]	*-0.10* [−0.35; 0.14]	*-0.50* [−1.57; 0.64]
establish	−0.78 [−0.98; −0.58]	−0.68 [−0.92; −0.42]	−1.92 [−2.76; −1.12]
hear	−2.92 [−3.12; −2.72]	−2.05 [−2.27; −1.83]	−7.57 [−8.41; −6.84]
inform	−1.37 [−1.57; −1.16]	−1.29 [−1.54; −1.05]	−3.93 [−4.71; −3.22]
know	−0.40 [−0.60; −0.21]	−0.28 [−0.54; −0.03]	−0.27 [−1.39; 0.96]
nonMent.C	−6.22 [−6.42; −6.02]	−0.27 [−0.52; −0.04]	−11.92 [−12.94; −11.01]
pretend	−4.47 [−4.72; −4.24]	−2.13 [−2.38; −1.87]	−10.78 [−11.83; −9.85]
prove	*-0.08* [−0.29; 0.14]	*-0.13* [−0.40; 0.14]	*0.92* [−0.57; 2.91]
reveal	−0.80 [−1.00; −0.60]	−0.67 [−0.92; −0.41]	−2.37 [−3.21; −1.54]
say	−2.20 [−2.41; −2.00]	−1.92 [−2.15; −1.69]	−5.79 [−6.59; −5.09]
see	−0.23 [−0.43; −0.02]	*-0.16* [−0.41; 0.09]	*-0.49* [−1.51; 0.62]
suggest	−3.47 [−3.67; −3.27]	−1.90 [−2.12; −1.67]	−8.75 [−9.62; −7.95]
think	−3.64 [−3.84; −3.44]	−1.85 [−2.08; −1.62]	−9.71 [−10.65; −8.87]

TABLE A4. Maximum a posteriori (MAP) model estimates for contradictoriness ratings from Exp. 3a (left and middle column, mean μ and precision ϕ) and Exp. 3b (right column, β) with 95% credible intervals. Contrast of each predicate is with contradictory control reference level. MAP estimates for which there is no evidence that they are different from 0 are italicized.

	Exp. 3a: Beta regression		Exp. 3b: logistic regression
Predictor	Estimated μ	Estimated ϕ	Estimated β
Intercept	2.74 [2.54; 2.93]	1.46 [1.23; 1.66]	6.17 [5.55; 6.88]
acknowledge	−2.15 [−2.39; −1.90]	−1.04 [−1.29; −0.79]	−4.89 [−5.65; −4.21]
admit	−2.03 [−2.26; −1.77]	−1.05 [−1.30; −0.78]	−4.86 [−5.59; −4.19]
announce	−2.90 [−3.14; −2.64]	−1.56 [−1.80; −1.32]	−6.46 [−7.20; −5.80]
be annoyed	−2.24 [−2.47; −1.98]	−1.28 [−1.53; −1.03]	−5.18 [−5.91; −4.51]
be right	−0.40 [−0.67; −0.12]	*-0.18* [−0.48; 0.12]	−1.90 [−2.70; −1.12]
confess	−2.15 [−2.39; −1.89]	−1.12 [−1.37; −0.86]	−4.85 [−5.59; −4.18]
confirm	−1.74 [−1.99; −1.48]	−0.99 [−1.24; −0.73]	−4.25 [−4.99; −3.57]
demonstrate	−1.94 [−2.18; −1.69]	−1.05 [−1.30; −0.79]	−4.53 [−5.25; −3.84]
discover	−1.63 [−1.90; −1.38]	−1.02 [−1.27; −0.75]	−3.30 [−4.06; −2.61]
establish	−1.94 [−2.19; −1.70]	−1.00 [−1.27; −0.75]	−4.29 [−5.06; −3.62]
hear	−3.72 [−3.97; −3.45]	−1.29 [−1.54; −1.01]	−8.98 [−9.79; −8.27]
inform	−2.78 [−3.03; −2.52]	−1.51 [−1.75; −1.25]	−6.46 [−7.20; −5.81]
know	−1.37 [−1.63; −1.11]	−0.90 [−1.17; −0.64]	−3.64 [−4.39; −2.95]
non-contra. control	−5.26 [−5.54; −4.98]	*-0.24* [−0.51; 0.04]	−11.51 [−12.43; −10.71]
pretend	−3.72 [−3.98; −3.46]	−1.35 [−1.61; −1.10]	−8.97 [−9.78; −8.26]
prove	−1.18 [−1.44; −0.92]	−0.71 [−0.98; −0.45]	−3.36 [−4.10; −2.65]
reveal	−2.27 [−2.52; −2.02]	−1.24 [−1.50; −0.98]	−4.99 [−5.73; −4.32]
say	−3.17 [−3.42; −2.91]	−1.51 [−1.75; −1.25]	−7.11 [−7.87; −6.42]
see	−1.43 [−1.68; −1.18]	−0.82 [−1.08; −0.56]	−3.48 [−4.21; −2.78]
suggest	−3.58 [−3.83; −3.31]	−1.17 [−1.43; −0.92]	−8.92 [−9.71; −8.21]
think	−3.93 [−4.19; −3.66]	−1.20 [−1.47; −0.94]	−9.81 [−10.67; −9.06]

E. Comparisons of gradient and categorical ratings. The Spearman rank correlation coefficient, a value between -1 and 1, is a nonparametric measure of rank correlation: the higher the coefficient, the more the relation between the the two variables can be described using a monotonic function; if the coefficient is positive, the value of one variable tends to increase with an increase in the other. For instance, in the case of our Exps. 1, a coefficient of 1 would mean that there is a perfectly monotone increasing relation between the mean certainty ratings of the predicates in Exp. 1a and Exp. 1b: for any two predicates p_1 and p_2, if p_1 ranks below p_2 in Exp. 1a (that is, the mean certainty rating of p_1 is lower than that of p_2), then that ranking is preserved in Exp. 1b.

The three panels of Figure A2 visualize the strong correlation between by-predicate mean ratings in Exps. 1a, 2a, and 3a and by-predicate proportion of 'yes' responses in Exps. 1b, 2b, and 3b, respectively.

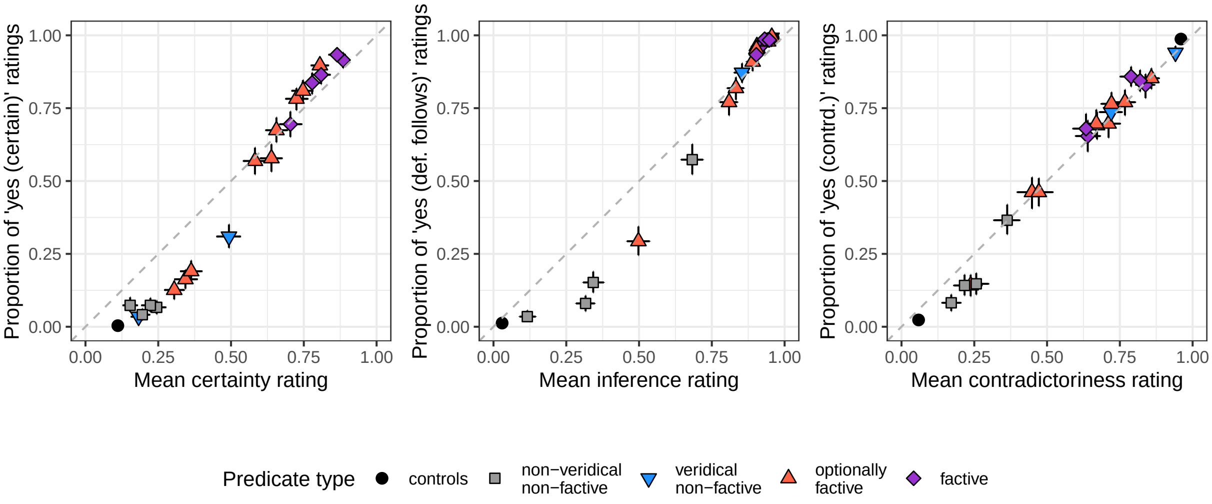
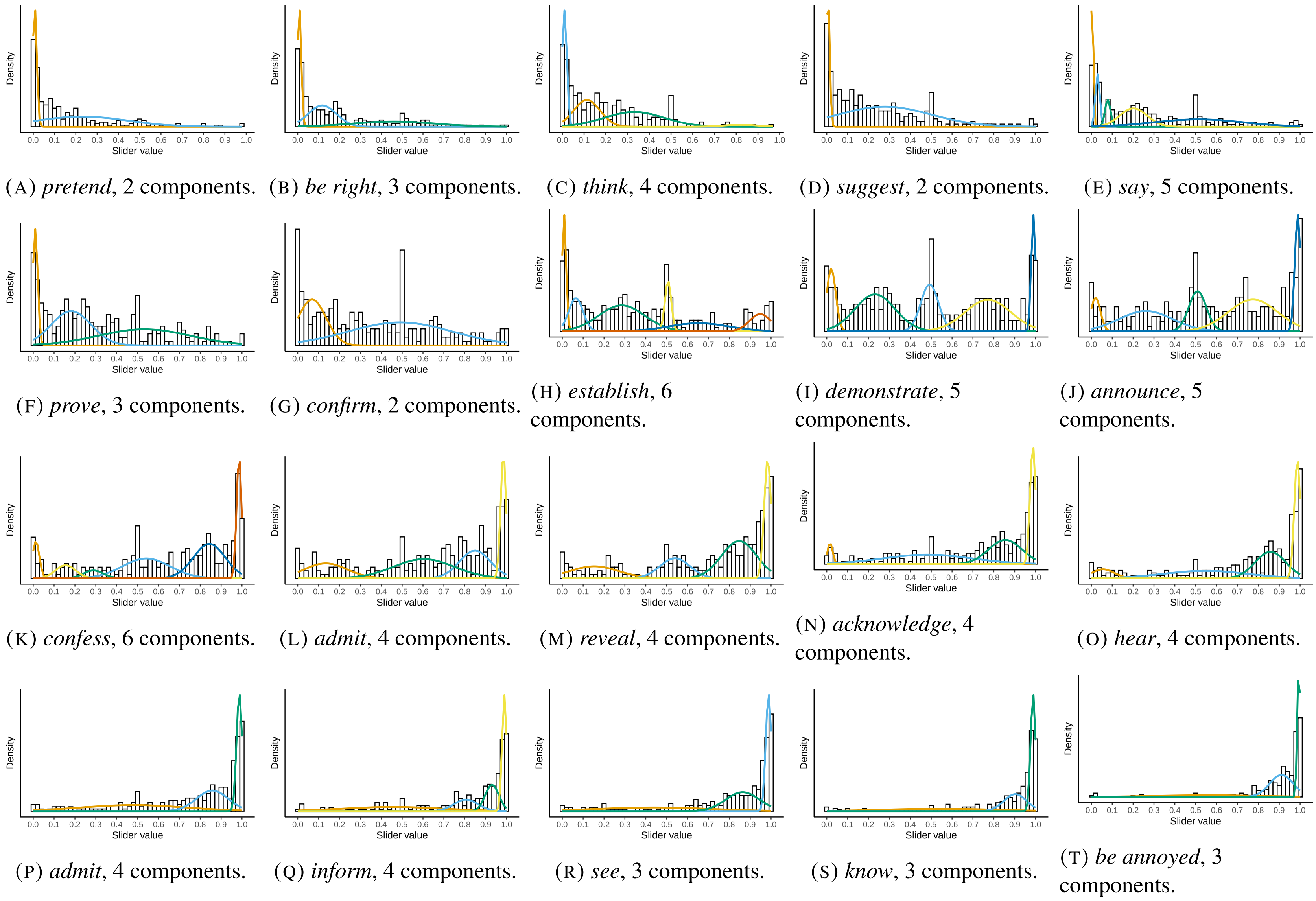


Figure A2. By-predicate proportion of 'yes' responses in two-alternative forced choice task against mean slider ratings in Exps. 1 (left, $r = .98$), Exps. 2 (middle, $r = .99$), and Exps. 3 (right, $r = .99$). Error bars indicate 95% bootstrapped confidence intervals.

F. Mixture models applied to data from Exp. 1a. Figure A3 plots the density of certainty ratings for the 20 predicates from Exp. 1a with the number of Gaussian components overlaid.



(A) *pretend*, 2 components. (B) *be right*, 3 components. (C) *think*, 4 components. (D) *suggest*, 2 components. (E) *say*, 5 components.

(F) *prove*, 3 components. (G) *confirm*, 2 components. (H) *establish*, 6 components. (I) *demonstrate*, 5 components. (J) *announce*, 5 components.

(K) *confess*, 6 components. (L) *admit*, 4 components. (M) *reveal*, 4 components. (N) *acknowledge*, 4 components. (O) *hear*, 4 components.

(P) *admit*, 4 components. (Q) *inform*, 4 components. (R) *see*, 3 components. (S) *know*, 3 components. (T) *be annoyed*, 3 components.

FIGURE A3. Histograms of certainty ratings from Exp. 1a with overlaid optimal number of Gaussian components.

REFERENCES

Bürkner, Paul-Christian. 2017. brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software* 80.1–28.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, England: Cambridge University Press.

Jaynes, Edwin, and Oscar Kempthorne. 1976. Confidence intervals vs. Bayesian intervals. *Foundations of probability theory, statistical inference, and statistical theories of science*, ed. by William Leonard Harper and Clifford Alan Hooker, 175–257. Dordrecht: Springer Netherlands.

Morey, Richard D.; Rink Hoekstra; Jeffrey N. Rouder; Michael D. Lee; and Eric J. Wagenmakers. 2016. The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin and Review* 23.103–123.

Nicenboim, Bruno, and Shravan Vasishth. 2016. Statistical methods for linguistic research: Foundational Ideas, Part II. *Linguistics and Language Compass* 10.591–613.

Smithson, Michael, and Jay Verkuilen. 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* 11.54.