SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments: Online Appendices

SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments: Online Appendices

Kyle Mahowald, Peter Graff, Jeremy Hartman, Edward Gibson

Language, Volume 92, Number 3, September 2016, pp. s1-s14 (Article)

Published by Linguistic Society of America
DOI: *https://doi.org/10.1353/lan.2016.0051*

SNAP JUDGMENTS: A SMALL N ACCEPTABILITY PARADIGM (SNAP) FOR LINGUISTIC ACCEPTABILITY JUDGMENTS: ONLINE APPENDICES

KYLE MAHOWALD
Massachusetts Institute of Technology

PETER GRAFF
Intel Corporation

JEREMY HARTMAN
University of Massachusetts Amherst

EDWARD GIBSON
Massachusetts Institute of Technology

APPENDIX A: RATING STUDY RESULTS

'*z*-bad' is the average *z*-score for the hypothesized 'bad' option. '*z*-good' is the average *z*-score for the hypothesized good option. 'Z.diff' is the difference between *z*-good and *z*-bad and is the effect size. Beta is the estimate from the linear mixed-effects model, which has a standard error '*SE*' and a *t*-value '*t*'. 'χ^2' is the chi-squared value comparing the full model to an intercept-only model, and 'χ^2 *p*' is the *p*-value obtained by that comparison. Simple '*p*' is just the *p*-value calculated using the *t*-value. Pred is TRUE if the effect goes in the significant direction. Sig is TRUE if there is a significant effect.

Rows in yellow are rows in which the effect goes in the predicted direction but is not significant.

EXPERIMENT	*z*-BAD	*z*-GOOD	Z.DIFF	BETA	*SE*	*t*	χ^2	χ^2 *p*	*p*	PRED	SIG
35.3.Hazout:36–36	−0.05	−0.04	0.01	0	0.06	0.08	0.01	0.935	0.936	TRUE	FALSE
34.4.Lasnik:24a–24b	0.2	0.21	0.01	0.01	0.08	0.12	0.02	0.901	0.904	TRUE	FALSE
34.1.Basilico:11a–12a	−0.46	−0.44	0.03	0.02	0.1	0.24	0.06	0.813	0.81	TRUE	FALSE
34.4.Lasnik:22a–22b	0.03	0.06	0.03	0.03	0.06	0.48	0.23	0.629	0.631	TRUE	FALSE
35.3.Hazout:73b–73b	−0.29	−0.17	0.11	0.11	0.07	1.72	2.89	0.089	0.085	TRUE	FALSE
33.1.Fox:47c–48b	−0.39	−0.26	0.12	0.12	0.07	1.69	2.68	0.102	0.091	TRUE	FALSE
32.2.Nunes:fn 35iia–fn35iib	−0.89	−0.78	0.12	0.12	0.06	2.02	4.06	0.044	0.043	TRUE	TRUE
35.2.Hazout:1b–1b	−0.35	−0.18	0.17	0.17	0.07	2.59	6.57	0.01	0.01	TRUE	TRUE
32.4.Lopez:9c–10c	−0.56	−0.36	0.2	0.2	0.05	3.59	11.01	0.001	< 0.0001	TRUE	TRUE
32.3.Culicover:37a–37a	−0.37	−0.15	0.22	0.21	0.07	3.21	9.95	0.002	0.001	TRUE	TRUE
33.4.Neeleman:97a–98	−0.33	−0.09	0.24	0.24	0.13	1.8	2.93	0.087	0.072	TRUE	FALSE
40.1.Heck:51–52	−0.63	−0.39	0.24	0.24	0.09	2.69	5.87	0.015	0.007	TRUE	TRUE
34.3.Landau:7c–7c	0.88	1.13	0.25	0.25	0.12	2.05	3.74	0.053	0.04	TRUE	FALSE
41.3.Landau:11a–11a	0.28	0.54	0.26	0.26	0.06	4.08	15.21	< 0.0001	< 0.0001	TRUE	TRUE
35.2.Larson:44b–44b	0.54	0.8	0.27	0.27	0.08	3.5	10.76	0.001	< 0.0001	TRUE	TRUE
34.1.Phillips:59c–60c	−0.74	−0.45	0.29	0.29	0.11	2.58	5.52	0.019	0.01	TRUE	TRUE
33.2.Bowers:49c–49c	−0.74	−0.46	0.29	0.29	0.09	3.35	8.65	0.003	0.001	TRUE	TRUE
34.2.Caponigro:11b–11c	–0.5	0.82	0.32	0.32	0.06	5.5	20.8	< 0.0001	< 0.0001	TRUE	TRUE
32.3.Fanselow:61a–61b	−0.96	−0.63	0.33	0.33	0.06	5.24	22	< 0.0001	< 0.0001	TRUE	TRUE
34.1.Phillips:23a–25a	−0.02	0.39	0.4	0.4	0.09	4.5	12.78	< 0.0001	< 0.0001	TRUE	TRUE
35.1.Bhatt:93a–b	−0.69	−0.29	0.41	0.4	0.08	5.02	14.82	< 0.0001	< 0.0001	TRUE	TRUE
32.3.Culicover:46a–48a	−0.2	0.21	0.41	0.41	0.07	6.3	22.52	< 0.0001	< 0.0001	TRUE	TRUE
34.3.Landau:38a–38c	−0.28	0.14	0.42	0.42	0.07	5.68	18.67	< 0.0001	< 0.0001	TRUE	TRUE
39.1.Sobin:8b–8f	−0.36	0.06	0.42	0.42	0.07	6.32	25	< 0.0001	< 0.0001	TRUE	TRUE
34.4.Boskovic:fn6iie–fn6iid	−0.34	0.12	0.46	0.46	0.1	4.54	13.78	< 0.0001	< 0.0001	TRUE	TRUE
35.3.Embick:62b–62b.Cf	0.4	0.86	0.46	0.46	0.08	5.54	17.72	< 0.0001	< 0.0001	TRUE	TRUE
34.4.Haegeman:2a–2b	−0.15	0.3	0.46	0.46	0.06	7.27	25.89	< 0.0001	< 0.0001	TRUE	TRUE
34.3.Landau:fn12i–fn12ii	−0.88	−0.42	0.46	0.46	0.06	7.73	27.48	< 0.0001	< 0.0001	TRUE	TRUE

34.1.Basilico:37a–37b	−0.37	0.1	0.47	0.47	0.09	5.31	16.56	< 0.0001	< 0.0001	TRUE	TRUE
39.1.Sobin:8c–8f	−0.32	0.15	0.47	0.47	0.06	8.02	35.62	< 0.0001	< 0.0001	TRUE	TRUE
35.2.Hazout:1a–1a	−0.86	−0.34	0.52	0.52	0.06	8.77	34.21	< 0.0001	< 0.0001	TRUE	TRUE
33.2.Bowers:7d–7d	0.56	1.12	0.56	0.56	0.1	5.48	17.11	< 0.0001	< 0.0001	TRUE	TRUE
34.1.Phillips:61a–61b	−1	−0.41	0.59	0.59	0.07	7.86	25.39	< 0.0001	< 0.0001	TRUE	TRUE
39.1.Sobin:20a–21a	−0.18	0.41	0.6	0.6	0.08	7.41	23.31	< 0.0001	< 0.0001	TRUE	TRUE
35.3.Embick:7a–7b	−0.47	0.14	0.62	0.62	0.09	6.98	22.55	< 0.0001	< 0.0001	TRUE	TRUE
34.1.Fox:37a–37b	−0.17	0.45	0.62	0.62	0.08	8.26	27.1	< 0.0001	< 0.0001	TRUE	TRUE
35.1.Bhatt:fn25ia–fn25ib	−1.08	−0.43	0.65	0.65	0.08	7.78	26.37	< 0.0001	< 0.0001	TRUE	TRUE
34.3.Takano:2b–d	−0.52	0.14	0.66	0.66	0.12	5.71	16.37	< 0.0001	< 0.0001	TRUE	TRUE
41.3.Landau:32a–32b	−0.6	0.06	0.66	0.66	0.07	8.98	29.15	< 0.0001	< 0.0001	TRUE	TRUE
34.1.Phillips:23a–24a	−0.41	0.3	0.71	0.71	0.13	5.28	14.52	< 0.0001	< 0.0001	TRUE	TRUE
33.2.Bowers:20a–20b	−0.53	0.18	0.71	0.71	0.12	5.84	16.21	< 0.0001	< 0.0001	TRUE	TRUE
34.4.Haegeman:2c–2b	−0.73	0	0.73	0.73	0.08	8.8	28.07	< 0.0001	< 0.0001	TRUE	TRUE
32.3.Culicover:25c–25d. WithOneself	−0.25	0.52	0.77	0.77	0.1	7.76	22.87	< 0.0001	< 0.0001	TRUE	TRUE
41.4.Bruening:61b–62b. StarredVariantIn61	−0.24	0.54	0.78	0.78	0.16	4.89	13.65	< 0.0001	< 0.0001	TRUE	TRUE
35.1.Bhatt:fn5ia–fn5ia	0.03	0.8	0.78	0.78	0.09	8.89	25.75	< 0.0001	< 0.0001	TRUE	TRUE
32.1.Martin:50b–51b	−0.49	0.29	0.78	0.78	0.07	11.65	42.71	< 0.0001	< 0.0001	TRUE	TRUE
32.3.Culicover:46b–46b	−0.32	0.49	0.81	0.81	0.08	9.89	30.41	< 0.0001	< 0.0001	TRUE	TRUE
40.4.Hicks:2a–2b	0.21	1.04	0.82	0.83	0.12	6.65	19.3	< 0.0001	< 0.0001	TRUE	TRUE
34.2.Panagiotidis:12a–b	0.07	0.92	0.84	0.84	0.13	6.5	19.02	< 0.0001	< 0.0001	TRUE	TRUE
38.2.Hornstein:fn2.iii–iii	−0.24	0.6	0.84	0.84	0.1	8.68	28.84	< 0.0001	< 0.0001	TRUE	TRUE
32.3.Culicover:34c–34e	−0.28	0.56	0.84	0.84	0.08	10.76	34.15	< 0.0001	< 0.0001	TRUE	TRUE
32.1.Martin:50a–51a	−0.21	0.65	0.87	0.87	0.09	9.2	27	< 0.0001	< 0.0001	TRUE	TRUE
35.2.Hazout:5a–5c	−0.67	0.21	0.89	0.89	0.06	15.54	42.33	< 0.0001	< 0.0001	TRUE	TRUE
40.4.Hicks:10a–10b	−0.8	0.11	0.91	0.91	0.08	11.63	36.63	< 0.0001	< 0.0001	TRUE	TRUE
32.3.Culicover:23c–23d. SentenceDP	0.2	1.12	0.92	0.93	0.15	6.11	16.76	< 0.0001	< 0.0001	TRUE	TRUE
34.3.Takano:2a–c	−0.36	0.58	0.94	0.93	0.09	10.1	27.12	< 0.0001	< 0.0001	TRUE	TRUE
34.1.Fox:4–4	−1.01	−0.09	0.93	0.93	0.08	11.24	36.45	< 0.0001	< 0.0001	TRUE	TRUE
41.4.Bruening:62a–87a. StarredVariantIn87	−0.31	0.65	0.95	0.95	0.12	8.01	23.2	< 0.0001	< 0.0001	TRUE	TRUE
34.3.Heycock:93a–93b	0.04	1.01	0.98	0.97	0.07	14	38.46	< 0.0001	< 0.0001	TRUE	TRUE
38.3.Landau:62a–62b	−0.12	0.87	0.98	0.98	0.1	10.32	29.29	< 0.0001	< 0.0001	TRUE	TRUE
32.3.Culicover:44a–45a	−0.61	0.4	1.01	1.01	0.06	16.01	44.47	< 0.0001	< 0.0001	TRUE	TRUE
40.2.Johnson:78–79	−0.57	0.48	1.05	1.04	0.08	12.3	31.13	< 0.0001	< 0.0001	TRUE	TRUE
41.3.Constantini:1b–1b.BothVsBothBoth	−0.14	0.91	1.04	1.04	0.08	13.69	37.01	< 0.0001	< 0.0001	TRUE	TRUE
34.2.Caponigro:fn6ia–fn6ib.EagerlyIn2ndPos	−0.06	0.98	1.03	1.04	0.07	15.02	37.98	< 0.0001	< 0.0001	TRUE	TRUE
34.1.Basilico:29b–30b	−0.97	0.07	1.05	1.05	0.1	10.71	30.49	< 0.0001	< 0.0001	TRUE	TRUE
34.3.Takano:11a–11b	−0.92	0.12	1.05	1.05	0.09	11.4	31.77	< 0.0001	< 0.0001	TRUE	TRUE
34.1.Fox:1–1	−1.01	0.08	1.09	1.09	0.07	14.94	46.32	< 0.0001	< 0.0001	TRUE	TRUE
37.4.Nakajima:fn1ia–fn1iiia	−0.91	0.21	1.12	1.11	0.13	8.69	23.29	< 0.0001	< 0.0001	TRUE	TRUE
35.1.Bhatt:5a–5c	−0.4	0.72	1.12	1.12	0.07	15.57	43.81	< 0.0001	< 0.0001	TRUE	TRUE
33.2.Bowers:56c–56d	−0.37	0.77	1.14	1.14	0.16	7.09	20.34	< 0.0001	< 0.0001	TRUE	TRUE
32.2.Alexiadou:fn11iib–fn11iic	−0.56	0.58	1.14	1.14	0.1	11.71	32.22	< 0.0001	< 0.0001	TRUE	TRUE
33.1.denDikken:56a–58a	−0.51	0.66	1.17	1.17	0.09	13.22	35.62	< 0.0001	< 0.0001	TRUE	TRUE
32.3.Culicover:fn6ia–fn6ib	−0.77	0.41	1.18	1.18	0.07	17.1	48.58	< 0.0001	< 0.0001	TRUE	TRUE
36.4.denDikken:35a–35b	−0.26	0.95	1.21	1.21	0.09	13.25	37.76	< 0.0001	< 0.0001	TRUE	TRUE
35.1.Bhatt:1b–1b	−0.52	0.69	1.21	1.21	0.07	17.76	50.67	< 0.0001	< 0.0001	TRUE	TRUE

34.3.Landau:fn13ii–fn13ii	−0.49	0.73	1.22	1.23	0.1	12.21	34.15	< 0.0001	< 0.0001	TRUE	TRUE
41.3.Vicente:6b–8b	−0.98	0.26	1.24	1.24	0.08	15.94	42.74	< 0.0001	< 0.0001	TRUE	TRUE
33.2.Bowers:7a–7a. PerfectlyIn2ndPos3rdPos	−0.42	0.82	1.25	1.25	0.07	17.18	40.23	< 0.0001	< 0.0001	TRUE	TRUE
41.1.Muller:28a–28b	−0.86	0.42	1.28	1.28	0.11	11.3	31.48	< 0.0001	< 0.0001	TRUE	TRUE
35.1.McGinnis:63a–63b	−0.35	0.94	1.28	1.28	0.09	14.49	36.9	< 0.0001	< 0.0001	TRUE	TRUE
38.2.Hornstein:2b–2c	−0.12	1.24	1.35	1.35	0.09	14.74	45.23	< 0.0001	< 0.0001	TRUE	TRUE
33.2.Bowers:19a–19b	−0.35	1.02	1.37	1.37	0.1	13.18	39.73	< 0.0001	< 0.0001	TRUE	TRUE
35.3.Embick:72a–72b	−0.37	1.05	1.41	1.41	0.13	11.17	30.27	< 0.0001	< 0.0001	TRUE	TRUE
32.1.Martin:15a–15b	−0.33	1.12	1.45	1.45	0.13	11.35	29.79	< 0.0001	< 0.0001	TRUE	TRUE
32.4.Lopez:16a–16b	−0.44	1.03	1.48	1.48	0.07	20.04	55.9	< 0.0001	< 0.0001	TRUE	TRUE
34.1.Basilico:50–51	−0.81	0.68	1.49	1.49	0.14	10.89	29.51	< 0.0001	< 0.0001	TRUE	TRUE
33.1.denDikken:57a–57b	−0.65	0.87	1.51	1.52	0.1	15.16	39.21	< 0.0001	< 0.0001	TRUE	TRUE
34.1.Basilico:7a–7b	−0.46	1.06	1.52	1.52	0.09	16.29	40.28	< 0.0001	< 0.0001	TRUE	TRUE
32.1.Martin:48a–48b	−0.93	0.67	1.6	1.6	0.12	13.14	33.95	< 0.0001	< 0.0001	TRUE	TRUE
32.3.Fanselow:59a–59b	−0.49	1.12	1.61	1.61	0.08	20.91	48.82	< 0.0001	< 0.0001	TRUE	TRUE
35.3.Hazout:30a–30a	−0.67	0.98	1.64	1.64	0.15	10.76	27.84	< 0.0001	< 0.0001	TRUE	TRUE
37.2.deVries:70a–70b	−0.68	0.97	1.65	1.65	0.07	22.04	50.1	< 0.0001	< 0.0001	TRUE	TRUE
35.2.Larson:61a–61b	−0.81	0.85	1.66	1.66	0.09	17.71	43.62	< 0.0001	< 0.0001	TRUE	TRUE
38.4.Boskovic:74–75	−0.8	0.87	1.67	1.67	0.08	20.91	44.93	< 0.0001	< 0.0001	TRUE	TRUE
35.3.Hazout:65a–65b	−0.99	0.73	1.72	1.72	0.12	13.94	35	< 0.0001	< 0.0001	TRUE	TRUE
33.2.Bowers:13b–13b	−0.81	0.98	1.79	1.79	0.11	16.49	40.15	< 0.0001	< 0.0001	TRUE	TRUE
35.1.Bhatt:13a–13a	−1.03	0.79	1.82	1.82	0.07	27.59	61.45	< 0.0001	< 0.0001	TRUE	TRUE
34.1.Basilico:4b–4c	−0.81	1.03	1.84	1.84	0.06	28.8	64.27	< 0.0001	< 0.0001	TRUE	TRUE
36.4.denDikken:38b–38b	−1.02	0.89	1.91	1.91	0.08	24.38	58.24	< 0.0001	< 0.0001	TRUE	TRUE
37.2.Sigurdsson:3c–3e	−0.92	1.08	2	2	0.07	29.93	58.39	< 0.0001	< 0.0001	TRUE	TRUE

APPENDIX B: FORCED-CHOICE RESULTS

'Gramm' is the proportion of people who choose the hypothesized acceptable sentence. 'Beta' is the model estimate of the effect size, which has a standard error of *SE* and a *z*-value (distance from 0 in units of standard error) of z. The *p*-value is calculated directly from the *z*-value. Pred is TRUE if the effect goes in the significant direction, FALSE otherwise. Sig is TRUE if there is a significant effect.

Rows in red represent contrasts where the effect is significant in the opposite direction of that predicted. Rows in pink show effects in the opposite direction of what was predicted but are not significant. Rows in yellow are rows in which the effect goes in the predicted direction but is not significant.

EXPERIMENT	GRAMM	BETA	*z*	*SE*	*p*	PRED	SIG
35.3.Hazout:36–36	0.39	−0.79	−4.03	0.2	< 0.0001	FALSE	TRUE
34.4.Lasnik:24a–24b	0.35	−0.73	−3.43	0.21	0.001	FALSE	TRUE
32.2.Nunes:fn35iia–fn35iib	0.44	−0.25	−1.4	0.18	0.162	FALSE	FALSE
32.4.Lopez:9c–10c	0.46	−0.19	−1.2	0.15	0.23	FALSE	FALSE
39.1.Sobin:8b–8f	0.46	−0.19	−1.09	0.17	0.276	FALSE	FALSE
34.4.Lasnik:22a–22b	0.47	−0.17	−0.7	0.25	0.484	FALSE	FALSE
34.1.Basilico:11a–12a	0.51	0.04	0.17	0.26	0.865	TRUE	FALSE
34.4.Haegeman:2a–2b	0.58	0.51	2.2	0.23	0.028	TRUE	TRUE
34.1.Phillips:23a–25a	0.62	0.65	2.45	0.26	0.014	TRUE	TRUE
33.4.Neeleman:97a–98	0.62	0.7	1.93	0.37	0.054	TRUE	FALSE
40.1.Heck:51–52	0.69	0.94	3.88	0.24	< 0.0001	TRUE	TRUE
39.1.Sobin:8c–8f	0.7	1.02	4.6	0.22	< 0.0001	TRUE	TRUE
34.3.Landau:fn12i–fn12ii	0.71	1.03	5.47	0.19	< 0.0001	TRUE	TRUE
34.1.Basilico:37a–37b	0.72	1.04	4.59	0.23	< 0.0001	TRUE	TRUE

33.1.Fox:47c–48b	0.71	1.05	4.06	0.26	< 0.0001	TRUE	TRUE
35.2.Larson:44b–44b	0.69	1.11	3.85	0.29	< 0.0001	TRUE	TRUE
34.3.Landau:7c–7c	0.71	1.18	6.07	0.19	< 0.0001	TRUE	TRUE
34.1.Phillips:61a–61b	0.77	1.3	10.07	0.13	< 0.0001	TRUE	TRUE
34.2.Panagiotidis:12a–b	0.75	1.45	4.1	0.35	< 0.0001	TRUE	TRUE
34.4.Boskovic:fn6iie–fn6iid	0.73	1.54	4.56	0.34	< 0.0001	TRUE	TRUE
32.3.Fanselow:61a–61b	0.78	1.55	8.99	0.17	< 0.0001	TRUE	TRUE
32.3.Culicover:37a–37a	0.79	1.57	9.36	0.17	< 0.0001	TRUE	TRUE
41.3.Constantini:1b–1b.BothVsBothBoth	0.79	1.58	7.19	0.22	< 0.0001	TRUE	TRUE
41.3.Landau:11a–11a	0.8	1.64	7.28	0.22	< 0.0001	TRUE	TRUE
32.3.Culicover:25c–25d.WithOneself	0.8	1.77	5.75	0.31	< 0.0001	TRUE	TRUE
41.4.Bruening:61b–62b.StarredVariantIn61	0.74	1.84	3.77	0.49	< 0.0001	TRUE	TRUE
34.2.Caponigro:11b–11c	0.83	2.01	8.24	0.24	< 0.0001	TRUE	TRUE
35.3.Embick:7a–7b	0.83	2.03	6.12	0.33	< 0.0001	TRUE	TRUE
34.1.Phillips:23a–24a	0.83	2.06	5.95	0.35	< 0.0001	TRUE	TRUE
35.1.Bhatt:93a–b	0.85	2.14	9.59	0.22	< 0.0001	TRUE	TRUE
39.1.Sobin:20a–21a	0.83	2.18	6.41	0.34	< 0.0001	TRUE	TRUE
40.4.Hicks:2a–2b	0.86	2.24	6.22	0.36	< 0.0001	TRUE	TRUE
33.1.denDikken:57a–57b	0.87	2.24	8.85	0.25	< 0.0001	TRUE	TRUE
33.2.Bowers:49c–49c	0.85	2.26	8.48	0.27	< 0.0001	TRUE	TRUE
35.1.Bhatt:fn25ia–fn25ib	0.89	2.31	8.81	0.26	< 0.0001	TRUE	TRUE
34.3.Landau:38a–38c	0.88	2.41	7.65	0.32	< 0.0001	TRUE	TRUE
35.1.Bhatt:fn5ia–fn5ia	0.88	2.49	6.74	0.37	< 0.0001	TRUE	TRUE
32.3.Culicover:46b–46b	0.89	2.51	7.03	0.36	< 0.0001	TRUE	TRUE
34.1.Phillips:59c–60c	0.86	2.59	7.84	0.33	< 0.0001	TRUE	TRUE
41.4.Bruening:62a–87a.StarredVariantIn87	0.83	2.6	4.57	0.57	< 0.0001	TRUE	TRUE
34.3.Takano:2b–d	0.86	2.65	5.06	0.52	< 0.0001	TRUE	TRUE
32.1.Martin:48a–48b	0.89	2.75	7.9	0.35	< 0.0001	TRUE	TRUE
32.1.Martin:50a–51a	0.91	2.79	7.2	0.39	< 0.0001	TRUE	TRUE
34.3.Takano:2a–c	0.9	2.8	7.52	0.37	< 0.0001	TRUE	TRUE
33.2.Bowers:20a–20b	0.88	2.82	8.01	0.35	< 0.0001	TRUE	TRUE
35.2.Hazout:1b–1b	0.86	2.91	3811.11	0	< 0.0001	TRUE	TRUE
35.3.Embick:62b–62b.Cf	0.87	2.93	7.25	0.4	< 0.0001	TRUE	TRUE
33.2.Bowers:7d–7d	0.9	2.95	7	0.42	< 0.0001	TRUE	TRUE
35.3.Hazout:73b–73b	0.9	2.98	8.21	0.36	< 0.0001	TRUE	TRUE
38.3.Landau:62a–62b	0.92	3	7.66	0.39	< 0.0001	TRUE	TRUE
33.2.Bowers:56c–56d	0.89	3.04	5.6	0.54	< 0.0001	TRUE	TRUE
38.2.Hornstein:fn2.iii–iii	0.95	3.06	13.01	0.23	< 0.0001	TRUE	TRUE
32.3.Culicover:34c–34e	0.91	3.32	6.97	0.48	< 0.0001	TRUE	TRUE
35.3.Hazout:65a–65b	0.97	3.38	11.94	0.28	< 0.0001	TRUE	TRUE
32.3.Fanselow:59a–59b	0.97	3.38	12.16	0.28	< 0.0001	TRUE	TRUE
35.2.Hazout:1a–1a	0.9	3.41	5.78	0.59	< 0.0001	TRUE	TRUE
32.1.Martin:50b–51b	0.91	3.42	6.34	0.54	< 0.0001	TRUE	TRUE
33.2.Bowers:7a–7a.PerfectlyIn2ndPos3rdPos	0.97	3.54	14.39	0.25	< 0.0001	TRUE	TRUE
32.3.Culicover:46a–48a	0.9	3.57	5.13	0.7	< 0.0001	TRUE	TRUE
32.3.Culicover:23c–23d.SentenceDP	0.94	3.6	5.91	0.61	< 0.0001	TRUE	TRUE
34.1.Fox:4–4	0.91	3.63	5.55	0.65	< 0.0001	TRUE	TRUE
34.3.Takano:11a–11b	0.92	3.65	5.1	0.71	< 0.0001	TRUE	TRUE
34.1.Fox:1–1	0.93	3.77	5.43	0.69	< 0.0001	TRUE	TRUE
35.3.Hazout:30a–30a	0.97	3.79	9.45	0.4	< 0.0001	TRUE	TRUE
37.4.Nakajima:fn1ia–fn1iiia	0.93	3.84	5.63	0.68	< 0.0001	TRUE	TRUE

34.1.Basilico:29b–30b	0.94	3.93	5.71	0.69	< 0.0001	TRUE	TRUE
34.1.Basilico:4b–4c	0.98	3.99	13.1	0.3	< 0.0001	TRUE	TRUE
35.2.Hazout:5a–5c	0.93	4.03	4.92	0.82	< 0.0001	TRUE	TRUE
41.3.Landau:32a–32b	0.92	4.17	3.88	1.07	< 0.0001	TRUE	TRUE
33.2.Bowers:13b–13b	0.99	4.44	11.69	0.38	< 0.0001	TRUE	TRUE
37.2.Sigurdsson:3c–3e	0.99	4.44	11.69	0.38	< 0.0001	TRUE	TRUE
40.4.Hicks:10a–10b	0.92	4.79	3.11	1.54	0.002	TRUE	TRUE
34.3.Landau:fn13ii–fn13ii	0.92	5.52	4.18	1.32	< 0.0001	TRUE	TRUE
34.1.Fox:37a–37b	0.92	6.06	5.86	1.04	< 0.0001	TRUE	TRUE
40.2.Johnson:78–79	0.94	6.7	5.87	1.14	< 0.0001	TRUE	TRUE
35.2.Larson:61a–61b	0.95	6.76	4.62	1.46	< 0.0001	TRUE	TRUE
32.2.Alexiadou:fn11iib–fn11iic	0.95	7.45	7.53	0.99	< 0.0001	TRUE	TRUE
34.4.Haegeman:2c–2b	0.93	7.5	7.21	1.04	< 0.0001	TRUE	TRUE
34.1.Basilico:7a–7b	0.97	7.59	6.8	1.12	< 0.0001	TRUE	TRUE
32.1.Martin:15a–15b	0.97	7.88	7.75	1.02	< 0.0001	TRUE	TRUE
38.4.Boskovic:74–75	0.97	7.95	1318.99	0.01	< 0.0001	TRUE	TRUE
32.3.Culicover:fn6ia–fn6ib	0.96	7.96	7.7	1.03	< 0.0001	TRUE	TRUE
33.2.Bowers:19a–19b	0.95	8.04	6.54	1.23	< 0.0001	TRUE	TRUE
34.2.Caponigro:fn6ia–fn6ib.EagerlyIn2ndPos	0.96	8.17	7.6	1.08	< 0.0001	TRUE	TRUE
37.2.deVries:70a–70b	0.97	8.26	7.25	1.14	< 0.0001	TRUE	TRUE
41.1.Muller:28a–28b	0.96	8.4	8.05	1.04	< 0.0001	TRUE	TRUE
38.2.Hornstein:2b–2c	0.97	8.49	7.31	1.16	< 0.0001	TRUE	TRUE
41.3.Vicente:6b–8b	0.97	8.49	7.31	1.16	< 0.0001	TRUE	TRUE
35.1.Bhatt:1b–1b	0.98	8.51	1004.54	0.01	< 0.0001	TRUE	TRUE
34.3.Heycock:93a–93b	0.96	8.61	6.13	1.4	< 0.0001	TRUE	TRUE
35.1.Bhatt:13a–13a	0.98	8.69	1834.83	0	< 0.0001	TRUE	TRUE
36.4.denDikken:38b–38b	0.96	8.76	8.92	0.98	< 0.0001	TRUE	TRUE
35.1.McGinnis:63a–63b	0.93	8.78	6.95	1.26	< 0.0001	TRUE	TRUE
34.1.Basilico:50–51	0.96	8.82	7.08	1.25	< 0.0001	TRUE	TRUE
32.4.Lopez:16a–16b	0.98	9.1	6.59	1.38	< 0.0001	TRUE	TRUE
35.3.Embick:72a–72b	0.98	9.1	7.07	1.29	< 0.0001	TRUE	TRUE
32.3.Culicover:44a–45a	0.96	9.33	5.31	1.76	< 0.0001	TRUE	TRUE
35.1.Bhatt:5a–5c	0.96	9.74	7.12	1.37	< 0.0001	TRUE	TRUE
33.1.denDikken:56a–58a	0.95	10.35	6.8	1.52	< 0.0001	TRUE	TRUE
36.4.denDikken:35a–35b	0.96	10.61	6.56	1.62	< 0.0001	TRUE	TRUE

APPENDIX C: REFERENCES FOR *LINGUISTIC INQUIRY* PAPERS

See full set of materials in the Materials folder at the Open Science Foundation, http://osf.io/5wm2a.

ALEXIADOU, ARTEMIS, and ELENA ANAGNOSTOPOULOU. 2001. The subject-in-situ generalization and the role of case in driving computations. *Linguistic Inquiry* 32.193–231. DOI: 10.1162/00243890152001753.

BASILICO, DAVID. 2003. The topic of small clauses. *Linguistic Inquiry* 34.1–35. DOI: 10.1162/00243890376 3255913.

BECKER, MISHA. 2006. There began to be a learnability puzzle. *Linguistic Inquiry* 37.441–56. DOI: 10.1162/ ling.2006.37.3.441.

BECK, SIGRID, and KYLE JOHNSON. 2004. Double objects again. *Linguistic Inquiry* 35.97–123. DOI: 10.1162/ 002438904322793356.

BHATT, RAJESH, and ROUMYANA PANCHEVA. 2004. Late merger of degree clauses. *Linguistic Inquiry* 35.1–45. DOI: 10.1162/002438904322793338.

BOECKX, CEDRIC, and SANDRA STJEPANOVIĆ. 2001. Head-ing toward PF. *Linguistic Inquiry* 32.345–55. DOI: 10.1162/00243890152001799.

BOŠKOVIĆ, ŽELJKO. 2002. On multiple *wh*-fronting. *Linguistic Inquiry* 33.351–83. DOI: 10.1162/002438902 760168536.

BOŠKOVIĆ, ŽELJKO. 2007. On the locality and motivation of Move and Agree: An even more minimal theory. *Linguistic Inquiry* 38.589–644. DOI: 10.1162/ling.2007.38.4.589.

BOŠKOVIĆ, ŽELJKO, and HOWARD LASNIK. 2003. On the distribution of null complementizers. *Linguistic Inquiry* 34.527–46. DOI: 10.1162/002438903322520142.

BOWERS, JOHN. 2002. Transitivity. *Linguistic Inquiry* 33.183–224. DOI: 10.1162/002438902317406696.

BRUENING, BENJAMIN. 2010a. Double object constructions disguised as prepositional datives. *Linguistic Inquiry* 41.287–305. DOI: 10.1162/ling.2010.41.2.287.

BRUENING, BENJAMIN. 2010b. Ditransitive asymmetries and a theory of idiom formation. *Linguistic Inquiry* 41.519–62. DOI: 10.1162/LING_a_00012.

CAPONIGRO, IVANO, and LISA PEARL. 2009. The nominal nature of *where, when,* and *how*: Evidence from free relatives. *Linguistic Inquiry* 40.155–64. DOI: 10.1162/ling.2009.40.1.155.

CAPONIGRO, IVANO, and CARSON T. SCHÜTZE. 2003. Parameterizing passive participle movement. *Linguistic Inquiry* 34.293–307. DOI: 10.1162/002438903321663415.

COSTANTINI, FRANCESCO. 2010. On infinitives and floating quantification. *Linguistic Inquiry* 41.487–96. DOI: 10.1162/LING_a_00006.

CULICOVER, PETER W., and RAY JACKENDOFF. 2001. Control is not movement. *Linguistic Inquiry* 32.493–512. DOI: 10.1162/002438901750372531.

DEN DIKKEN, MARCEL. 2005. Comparative correlatives comparatively. *Linguistic Inquiry* 36.497–532. DOI: 10.1162/002438905774464377.

DEN DIKKEN, MARCEL, and ANASTASIA GIANNAKIDOU. 2002. From *hell* to polarity: 'Aggressively non-D-linked' *wh*-phrases as polarity items. *Linguistic Inquiry* 33.31–61. DOI: 10.1162/002438902317382170.

DE VRIES, MARK. 2006. The syntax of appositive relativization: On specifying coordination, false free relatives, and promotion. *Linguistic Inquiry* 37.229–70. DOI: 10.1162/ling.2006.37.2.229.

EMBICK, DAVID. 2004. On the structure of resultative participles in English. *Linguistic Inquiry* 35.355–92. DOI: 10.1162/0024389041402634.

FANSELOW, GISBERT. 2001. Features, θ-roles, and free constituent order. *Linguistic Inquiry* 32.405–37. DOI: 10.1162/002438901750372513.

FOX, DANNY. 2002. Antecedent-contained deletion and the copy theory of movement. *Linguistic Inquiry* 33.63–96. DOI: 10.1162/002438902317382189.

FOX, DANNY, and HOWARD LASNIK. 2003. Successive-cyclic movement and island repair: The difference between sluicing and VP-ellipsis. *Linguistic Inquiry* 34.143–54. DOI: 10.1162/002438903763255959.

HADDICAN, BILL. 2007. The structural deficiency of verbal pro-forms. *Linguistic Inquiry* 38.539–47. DOI: 10.1162/ling.2007.38.3.539.

HAEGEMAN, LILIANE. 2003. Notes on long adverbial fronting in English and the left periphery. *Linguistic Inquiry* 34.640–49. DOI: 10.1162/ling.2003.34.4.640.

HAEGEMAN, LILIANE. 2010. The movement derivation of conditional clauses. *Linguistic Inquiry* 41.595–621. DOI: 10.1162/LING_a_00014.

HAZOUT, ILAN. 2004a. Long-distance agreement and the syntax of *for-to* infinitives. *Linguistic Inquiry* 35. 338–43. DOI: 10.1162/ling.2004.35.2.338.

HAZOUT, ILAN. 2004b. The syntax of existential constructions. *Linguistic Inquiry* 35.393–430. DOI: 10.1162/0024389041402616.

HECK, FABIAN. 2009. On certain properties of pied-piping. *Linguistic Inquiry* 40.75–111. DOI: 10.1162/ling.2009.40.1.75.

HEYCOCK, CAROLINE, and ROBERTO ZAMPARELLI. 2003. Coordinated bare definites. *Linguistic Inquiry* 34. 443–69. DOI: 10.1162/002438903322247551.

HICKS, GLYN. 2009. *Tough*-constructions and their derivation. *Linguistic Inquiry* 40.535–66. DOI: 10.1162/ling.2009.40.4.535.

HIROSE, TOMIO. 2007. Nominal paths and the head parameter. *Linguistic Inquiry* 38.548–53. DOI: 10.1162/ling.2007.38.3.548.

HORNSTEIN, NORBERT. 2007. A very short note on existential constructions. *Linguistic Inquiry* 38.410–11. DOI: 10.1162/ling.2007.38.2.410.

JOHNSON, KYLE. 2009. Gapping is not (VP-) ellipsis. *Linguistic Inquiry* 40.289–328. DOI: 10.1162/ling.2009.40.2.289.

KALLULLI, DALINA. 2007. Rethinking the passive/anticausative distinction. *Linguistic Inquiry* 38.770–80. DOI: 10.1162/ling.2007.38.4.770.

LANDAU, IDAN. 2003. Movement out of control. *Linguistic Inquiry* 34.471–98. DOI: 10.1162/0024389033 22247560.

LANDAU, IDAN. 2007. EPP extensions. *Linguistic Inquiry* 38.485–523. DOI: 10.1162/ling.2007.38.3.485.

LANDAU, IDAN. 2010. The explicit syntax of implicit arguments. *Linguistic Inquiry* 41.357–88. DOI: 10.1162/LING_a_00001.

LARSON, RICHARD K., and FRANC MARUŠIČ. 2004. On indefinite pronoun structures with APs: Reply to Kishimoto. *Linguistic Inquiry* 35.268–87. DOI: 10.1162/002438904323019075.

LASNIK, HOWARD, and MYUNG-KWAN PARK. 2003. The EPP and the subject condition under sluicing. *Linguistic Inquiry* 34.649–60. DOI: 10.1162/ling.2003.34.4.649.

LÓPEZ, LUIS. 2001. On the (non)complementarity of θ-theory and checking theory. *Linguistic Inquiry* 32.694–716. DOI: 10.1162/002438901753373050.

MARTIN, ROGER. 2001. Null case and the distribution of PRO. *Linguistic Inquiry* 32.141–66. DOI: 10.1162/002438901554612.

MCGINNIS, MARTHA. 2004. Lethal ambiguity. *Linguistic Inquiry* 35.47–95. DOI: 10.1162/00243890432 2793347.

MÜLLER, GEREON. 2010. On deriving CED effects from the PIC. *Linguistic Inquiry* 41.35–82. DOI: 10.1162/ling.2010.41.1.35.

NAKAJIMA, HEIZO. 2006. Adverbial cognate objects. *Linguistic Inquiry* 37.674–84. DOI: 10.1162/ling.2006.37.4.674.

NEELEMAN, AD, and HANS VAN DE KOOT. 2002. The configurational matrix. *Linguistic Inquiry* 33.529–74. DOI: 10.1162/002438902762731763.

NUNES, JAIRO. 2001. Sideward movement. *Linguistic Inquiry* 32.303–44. DOI: 10.1162/00243890152001780.

PANAGIOTIDIS, PHOEVOS. 2003. *One*, empty nouns, and θ-assignment. *Linguistic Inquiry* 34.281–92. DOI: 10.1162/ling.2003.34.2.281.

PHILLIPS, COLIN. 2003. Linear order and constituency. *Linguistic Inquiry* 34.37–90. DOI: 10.1162/0024389 03763255922.

REZAC, MILAN. 2010. ϕ-Agree versus ϕ-feature movement: Evidence from floating quantifiers. *Linguistic Inquiry* 41.496–508. DOI: 10.1162/LING_a_00007.

RICHARDS, NORVIN. 2004. Against bans on lowering. *Linguistic Inquiry* 35.453–63. DOI: 10.1162/00243890 41402643.

SIGURÐSSON, HALLDÓR ÁRMANN. 2006. The nominative puzzle and the low nominative hypothesis. *Linguistic Inquiry* 37.289–308. DOI: 10.1162/ling.2006.37.2.289.

SOBIN, NICHOLAS. 2004. Expletive constructions are not 'lower right corner' movement constructions. *Linguistic Inquiry* 35.503–8. DOI: 10.1162/ling.2004.35.3.503.

SOBIN, NICHOLAS. 2008. *Do so* and VP. *Linguistic Inquiry* 39.147–60. DOI: 10.1162/ling.2008.39.1.147.

STEPANOV, ARTHUR; and PENKA STATEVA. 2009. When QR disobeys superiority. *Linguistic Inquiry* 40.176–85. DOI: 10.1162/ling.2009.40.1.176.

STROIK, THOMAS. 2001. On the light verb hypothesis. *Linguistic Inquiry* 32.362–69. DOI: 10.1162/ling.2001 .32.2.362.

TAKANO, YUJI. 2003. How antisymmetric is syntax? *Linguistic Inquiry* 34.516–26. DOI: 10.1162/ling.2003 .34.3.516.

VICENTE, LUIS. 2010. A note on the movement analysis of gapping. *Linguistic Inquiry* 41.509–17. DOI: 10.1162/LING_a_00008.

APPENDIX D: DISCUSSION OF ITEMS THAT DO NOT SHOW CLEAR RESULTS IN THE PREDICTED DIRECTION

35.3.Hazout:36

(#) There seem/*seems to have appeared [some new candidates] in the course of the presidential campaign.

The rating study revealed no significant difference between the two variants ($\beta = 0$), and the starred variant was significantly preferred in the forced-choice experiment. This judgment seems to reflect a trend in colloquial English to use the singular *There seems* in these 'verbal existential sentences', even when the agreeing phrase is plural. At the very least, there may be individual variation in sentences like this.

34.4.Lasnik:24a–24b

a. ?The detective asserted two students to have been at the demonstration during each other's hearings.
b. ?*The detective asserted that two students were at the demonstration during each other's hearings.

Example (b) is proposed to be unacceptable only when the final PP modifies the matrix clause and not the embedded clause. Our items were written to ensure that this is the only plausible interpretation, but participants still preferred (b) by a significant margin in the forced-choice experiment.

34.4.Lasnik:22a–22b

a. John proved three chapters to have been plagiarized with one convincing example each.
b. ?*John proved that three chapters were plagiarized with one convincing example each.

This example showed a nonsignificant trend in favor of (a) in the rating study and a nonsignificant trend toward (b) in the forced-choice study. Again, we took care to ensure that the final PP modifies the matrix verb across all our items.

32.4.Lopez:9c–10c

a. We proved Smith to the authorities to be the thief.
b. *We proved to the authorities Smith to be the thief.

People significantly preferred (a) in the rating study, but the opposite trend emerged in the forced-choice study, which suggests that this is not a clear contrast. In fact, Hartman (2011) has argued that sentences like (a) are degraded on independent grounds, which might explain why most subjects did not prefer them over (b).

39.1.Sobin:8b–8f

a. Bill devoured a ham, and Mary did a similar thing with a chicken.
b. *Bill devoured a ham, and Mary did so with a chicken.

In this contrast, we found a significant predicted effect in the rating study but a trend in the opposite direction in the forced-choice experiment. It is possible, in this case, that the *did so* construction in (b) is semantically unclear out of context, but clearer (and more natural sounding) when presented with the more semantically transparent (a). This would explain the difference between the rating study and the forced-choice study.[1]

APPENDIX E: MATH BEHIND SNAP JUDGMENTS

Formally, we can think of our experiment as a draw from a binomial distribution, where p is the underlying population parameter for how likely someone is to choose sentence A over sentence B, n is the total number of trials, and k is the number of trials on which someone chose sentence A over sentence B.

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

To obtain a confidence interval from a binomial distribution where the sample is unanimous while also taking advantage of our prior knowledge about how MOST experiments turn out, we will use a Bayesian credible interval—which is the Bayesian version of a confidence interval and can be thought of as the probability that a given parameter falls within some interval—on the posterior distribution. We get the posterior distribution by combining our binomial likelihood with a beta prior distribution (Gelman et al. 2004) on the parameter p, which gives a distribution of possible values for our parameter p. This prior distribution is the distribution over the value of p BEFORE we have collected any data. In other words, before we flip the coin, we do not know its weight p. We might think that it is very likely that the coin is fair and that p is near 0.50. Or maybe we think that p is close to 1. The shape of the distribution is controlled by the shape parameters α and β. Formally, the beta distribution is:

$$P(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)},$$

[1] These results also demonstrate that different experimental tasks can sometimes give different results. Specifically, it seems that (b)'s unacceptability is largely context-dependent.

where *B* is the beta function. We could, in principle, use any distribution with support on [0,1], but we use the beta distribution because it is the conjugate prior for the binomial and thus lets us obtain a closed-form solution.

Informally, we can think of the job of the prior as being to add in our prior belief about the underlying distribution. We can literally think of this as adding the results of imaginary trials that we have not actually conducted. For instance, if we suspect that the coin is fair, we might use a beta prior of Beta(5, 5)—meaning α and β are both 5. Then, we present five people with sentence A and sentence B and ask which is better. In this case, *p* is the underlying probability of choosing A. We get the following result:

A A A A A

Without the prior, our best guess for the underlying parameter *p* is 1 since 5/5 is 1. If we use the Beta(5, 5) prior, however, we can think of this as adding five a priori As and five a priori Bs to our five experimentally obtained As such that we imagine we have ten As and five Bs, as in the following (where the italicized values come from the prior):

A A A A A B B B B B **A A A A A**

In this case, our best estimate of the underlying parameter *p* is (5 As + 5 As) / (15 trials) = 0.66. If we were very confident that the sentences are equally acceptable (i.e. the coin is fair; $p \sim 0.5$), we could use a Beta(100, 100) prior. With a prior like that, we would have to conduct many more trials in order to move our estimate substantially away from 0.50. After getting five As, we would still have an estimate of 51%.

If we thought it was very likely that one of the sentences was better, but we did not know which, we might instead use a beta prior of Beta(.1, .1). This would mean that, after asking five people who all choose A, our new estimate for how likely a random person is to choose A would be: 5.1/(5.1 + .1) = 98%. Figure 3 shows the shape of the beta distribution for two possible settings of the shape parameters. If the shape parameters are unequal, then the distribution is skewed. When the two shape parameters are equal, the distribution is symmetric.

Formally, we can multiply the beta prior and the binomial likelihood together to get the posterior probability.

$$P(k|n,p) * P(p|\alpha,\beta) = P(k|n,\alpha,\beta) = \binom{n}{k} \frac{B(k+\alpha, n-k+\beta)}{B(\alpha,\beta)}$$

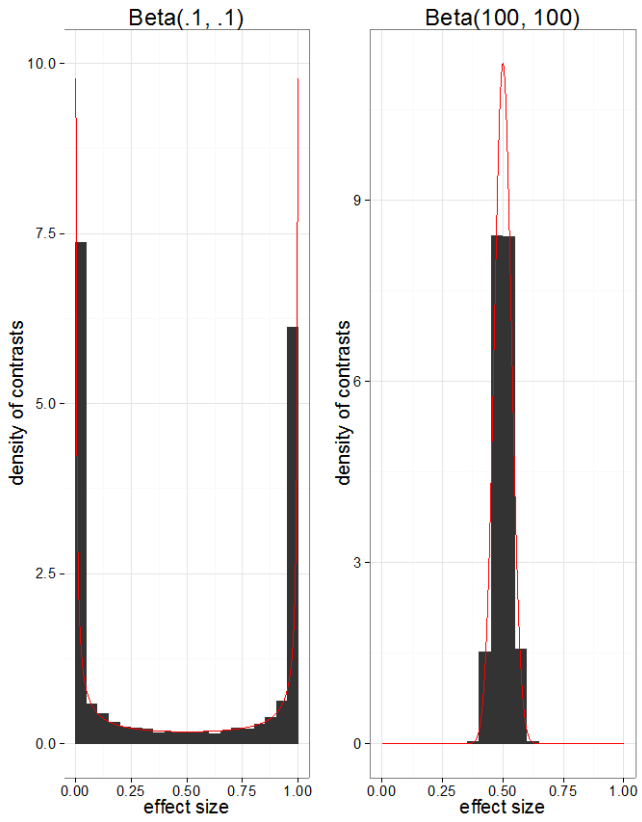

FIGURE 3. The histograms represent a density map of a draw from a beta distribution with the shape parameters indicated. The red line is the probability density of the beta distribution at each value for *p* between 0 and 1. The plot on the left conforms to an instance in which, most of the time, the probability *p* is extreme (toward 0 or 1), as in the experiments we tested here. The plot on the right corresponds to a situation in which we have a strong prior belief that the probability *p* is near 0.5.

In our case, we want to know what our prior expectations about *p* should be. Should our prior look more like Figure 3a or Figure 3b? Because we have formal results for 100 contrasts, we can use these empirical results to set our prior.[2] In other words, when we have a new contrast for which we do not have much data but which we believe likely to produce a unanimous result, we can imagine that the contrast has an underlying parameter *p* (where *p* is once again the probability of choosing sentence A) and that *p* is drawn from the same distribution of judgments that gave rise to the 100 contrasts we observed. If we do not believe that the contrast is likely to produce a unanimous result, the assumption that the parameter *p* is drawn from the same distribution as the 100 contrasts we tested experimentally is potentially invalid since, in general, the effects that we tested were hypothesized to be very strong.

[2] The prior that is obtained by our experimental results ends up very similar to what is obtained from the results from Sprouse, Schütze, and Almeida's (2013) data (available on Jon Sprouse's webpage).

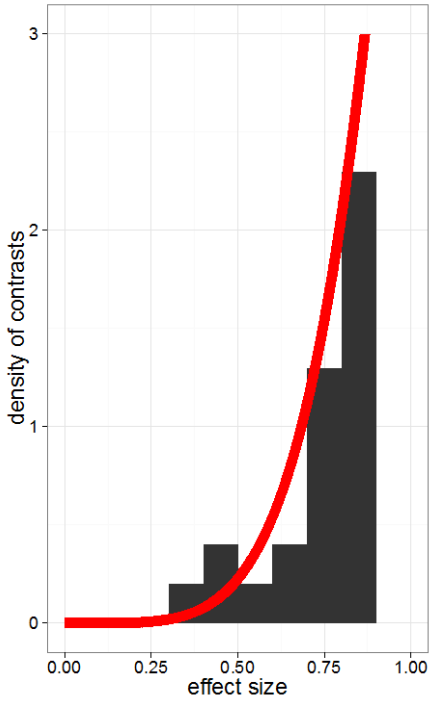

FIGURE 4. This plot corresponds to a smoothed histogram (averaged over many trials) of the data from our forced-choice experiment where, for each contrast, one variant is randomly assigned to be sentence A and one to be sentence B. Most of the time, there is a strong preference for one sentence or the other. The best fit for the beta distribution is Beta(5.9, 1.1)—which is shown by the red line.

In order to determine the prior empirically, for each contrast in our experiment, we assume that the hypothesized 'good' sentence is sentence A. We then draw a histogram of the effect sizes and fit the beta distribution to the histogram (as seen in Fig. 4). Averaging over 100 samples, the best fit is Beta(5.9, 1.1), with standard error .12 and .01, respectively. Rounding to the nearest whole number, we can think of this as having seen six As and one B BEFORE we run our experiment. Thus, if we run an experiment and get three As and zero Bs, we can act as if we have nine As and one B. We can use this prior to construct 95% Bayesian credible intervals for the underlying probability in the population of someone preferring sentence A over sentence B. Specifically, the Bayesian credible interval gives us a continuous interval, for which there is a 95% probability that the true underlying probability falls in that region.

We also checked to see if the recommendations here were robust to other reasonable choices of prior. There is some theoretical question as to whether it makes sense to use the full available information in order to set the prior or if we should instead 'forget' which sentence is hypothesized to be good and assume that it is equally likely that the good sentence is A or B. The logic here is that including information as to which sentence is supposed to be good would be equivalent to doing an experiment where a researcher wants to test the efficacy of a medicine and then includes her prior belief that the medicine will probably work as evidence in the experiment. While she might be very confident in the medicine's efficacy, she cannot include that prior belief as part of her analysis or else she could end up concluding that data that are consistent with pure noise are actually a result in favor of the hypothesis. But, because the

whole point of the SNAP Judgment paradigm is to use the existing information, we do not believe those concerns are particularly relevant here.[3]

To check how robust the paradigm is to choice of prior, we tried this approach where A and B are equally likely to be the 'good' sentence. To do that, we randomly assign one sentence in each contrast as A and one as B. Using this approach, we find a Beta(.6, .6) prior. For five unanimous participants, this gives us a mean of .90 with a 95% CI of [.67, 1]. So the CI's lower bound is only slightly lower than when we include all the information. To get the lower bound to .75 when we use this prior, we would need to include seven participants in the experiment (as compared to five in our main analysis). We would also arrive at similar conclusions if we used the Jeffreys uninformative prior Beta(.5, .5)—a prior that is standardly used in many applications since it is locally uniform. Hence, the outcome is similar under other plausible alternative priors. We use the asymmetric, full-information prior in our main analysis, but we recognize that there may be good theoretical reasons to instead use the symmetric prior.

APPENDIX F: STATISTICAL POWER

The idea of computing statistical power is to ask, if there is an underlying 'true effect' size *D* that is being looked for in the experiment, what is the likelihood that the experiment correctly detects a significant effect? (Note that, in reality, we can never know the 'true effect size' because that would require infinite data. We can only sample.) If $D = .8$ for a sentence in the forced-choice experiment, that would mean the true underlying effect was .80. If the statistical power of our experiment is .95 (based on the sample size and design), that would mean that 95% of the time we would find a significant effect given the underlying effect size of .80. (Power would be lower if the effect size were smaller.) To compute statistical power and possible error rates using linear mixed-effects models, we repeated the following procedure 100 times for each contrast, took the mean of those 100 iterations, and then averaged across contrasts.

a) Fit a linear mixed-effects model to the real data as described in the main text.
b) Use the random-effects structure and residual variance from the model fit to the actual data in a). For the fixed-effect estimate, use *D*, which we systematically vary and report for several values in the table below. In effect, this lets us use the actual variance in the world (by subject, by item, and residual variance) to estimate the noise we should expect in an experiment.
c) Use the parameters from b) to simulate a new set of data equivalent in sample size to the original experiment and with the same subject and item breakdown as the original experiment.
d) Fit a new linear mixed-effects model to the simulated data in c) and test for effect size and significance.
e) Use the effect sizes and significance levels found in d) to calculate power, type S, and type M error.

We used the simulated effect size and significance measures to calculate statistical power given varying underlying effect sizes as well as two measures recommended: type S (sign) error and type M (magnitude) error (Gelman & Carlin 2014). Power here refers to the proportion of the time a 'true effect' would be detected in the experiment given true effect size *D*. Type S error refers to the proportion of the time a significant effect is found in the OPPOSITE direction of the true effect. That is, if the type S error rate is .05, that means that 5% of the time, we should expect to find a significant effect in the opposite direction of the true effect. Type M error refers to the expected absolute overestimation rate given that a significant

[3] See Cox & Mayo 2011 and Gelman 2012 for more discussion of how to use prior information responsibly in scientific inference.

effect is found (that is, when significant, the absolute value of the estimated effect size divided by the true effect size). This means that, conditioned on finding a significant effect, we should expect it to be *M* times more extreme than the underlying true effect.

The tables below report power and estimated error rates for various true effect sizes. Note that, in the rating study, a true effect size less than .4 is quite small (only 19% of our estimated effect sizes are this small) and possibly not large enough for robust acceptability generalizations. For the forced-choice study, an effect size less than .70 is quite small, and only 11% of our data fits that description.

D (TRUE EFFECT SIZE)	STATISTICAL POWER	TYPE S ERROR RATE	TYPE M ERROR RATE
.2	0.63	0.0	1.29
.4	0.96	0.0	1.01
.6	1.00	0.0	1.00

TABLE F1. Ratings study (all values where significance is defined by $p < 0.05$).

D (TRUE EFFECT SIZE)	STATISTICAL POWER	TYPE S ERROR RATE	TYPE M ERROR RATE
.6	.48	.04	1.71
.7	.80	0.0	1.17
.8	.93	0.0	1.06

TABLE F2. Forced-choice study (all values where significance is defined by $p < 0.05$).

* Note that for the forced-choice study, the type M error rate refers to the overestimation rate of the difference between the effect size *D* (defined as the proportion choosing the good sentence) and .5 (50% baseline in which neither sentence is better than another). So a 1.17 type M error rate for $D = .7$ means that, on average, if the contrast is significant at $p < 0.05$, the difference between the estimated *d* and .5 is 1.17 higher than it should be (where what it 'should be' is $.7 - .5 = .2$).

REFERENCES

COX, SIR DAVID, and DEBORAH MAYO. 2011. Statistical scientist meets a philosopher of science: A conversation. *Rationality, Markets and Morals* 2.103–14. Online: http://www.rmm-journal.de/downloads/Article_Cox_Mayo.pdf.

GELMAN, ANDREW. 2012. Ethics and statistics: Ethics and the statistical use of prior information. *CHANCE* 25.52–54. DOI: 10.1080/09332480.2012.752294.

GELMAN, ANDREW, and JOHN CARLIN. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9.641–51. DOI: 10.1177/1745691614551642.

GELMAN, ANDREW; JOHN B. CARLIN; HAL S. STERN; and DONALD B. RUBIN. 2004. *Bayesian data analysis*. Boca Raton, FL: CRC Press.

SPROUSE, JON; CARSON T. SCHÜTZE; and DIOGO ALMEIDA. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua* 134.219–48. DOI: 10.1016/j.lingua.2013.07.002.

[kmahowald@gmail.com]
[graffmail@gmail.com]
[hartman@linguist.umass.edu]
[egibson@mit.edu]