Title:
SNAP Judgments: A Small N Acceptability Paradigm (SNAP) for Linguistic Acceptability Judgments

Authors:
Kyle Mahowald (corresponding author)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
77 Massachusetts Avenue, 46-3037
Cambridge, MA 02139
kylemaho@mit.edu

Peter Graff
Intel Corporation
2200 Mission College Blvd.
Santa Clara, CA 95054
peter.graff@intel.com

Jeremy Hartman
University of Massachusetts Amherst
N408 Integrative Learning Center
650 North Pleasant Street
Amherst, MA 01003
hartman@linguist.umass.edu

Edward Gibson
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
77 Massachusetts Avenue, 46-3037
Cambridge, MA 02139
egibson@mit.edu

Key words:
syntax, semantics, linguistic acceptability, quantitative linguistics

Abstract

While published linguistic judgments sometimes differ from the judgments found in large-scale formal experiments with naive participants, there is not a consensus as to how often these errors occur nor as to how often formal experiments should be used in syntax and semantics research. First, we present results of a large-scale replication of Sprouse, Schütze, and Almeida (2013) on 100 English contrasts randomly sampled from *Linguistic Inquiry* 2001-2010 and tested in both a forced-choice experiment and an acceptability rating experiment. Like Sprouse, Schütze, and Almeida, we find that the effect sizes of published linguistic acceptability judgments are not uniformly large or consistent but rather form a continuum from very large effects to small or non-existent effects. We then use this data as a prior in a Bayesian framework to propose a Small N Acceptability Paradigm for Linguistic Acceptability Judgments (SNAP Judgments). This proposal makes it easier and cheaper to obtain quantitative and statistically valid data in syntax and semantics research. Specifically, for a contrast of linguistic interest for which a researcher is confident that Sentence A is better than Sentence B, we recommend that the researcher should obtain judgments from 7 unique participants, using 7 unique sentences of each type. If all 7 participants agree that Sentence A is better than Sentence B, then the researcher can be confident that the result of a full forced choice experiment would likely be 75% or more agreement in favor of Sentence A (with a mean of 93%). We test this proposal by sampling from the existing data and find that it gives highly reliable performance.

1. Introduction

Historically, the method in syntax and semantics research was for the researcher to use his or her own intuitions about the acceptability of phrases and sentences. This informal method worked when the field was developing, and the contrasts were large as in (1), but as the field progressed, the contrasts needed for deciding among competing theories became more complex, and the judgments consequently became more subtle, as in (2):

(1) (Chomsky, 1957)

a. Colorless green ideas sleep furiously.

b. Furiously sleep ideas green colorless.

(2) (Chomsky, 1986)
a. What do you wonder who saw?
b. I wonder what who saw.

Many researchers have observed weaknesses of the informal method (e.g., Arppe & Järvikivi, 2007; Cowart, 1997; Featherston, 2005; Gibson & Fedorenko, 2010, 2013; Gibson, Piantadosi, & Fedorenko, 2013; Gross & Culbertson, 2011; Schutze, 1996; Sorace & Keller, 2005; Wasow & Arnold, 2005). Such weaknesses include potential cognitive biases on the part of the researcher and participants, difficulty in controlling for discourse context, the inability to find interactions among factors, and the inability to find probabilistic effects or relative effect sizes. Furthermore, for readers who do not natively speak the language that is being described, it is difficult to evaluate the size of informally reported contrasts. For this reason, a major advantage of formal methods is that they provide fellow researchers with quantitative information about the quality of data that is gathered: quantitative details enable an understanding of which comparisons support a theory, and which do not.

Below, we address several remaining arguments against the widespread adoption of quantitative methods in syntax and semantics research. We then present a formal replication of a large-scale experiment by Sprouse, Schütze, & Almeida (2013; henceforth SSA) on a set of sentences sampled from *Linguistic Inquiry* 2001-2010. Using these data from 100 pairwise comparisons randomly sampled from the same set of articles investigated by SSA, we present a novel proposal and provide empirical support for a Small N Acceptability Paradigm for linguistic judgments (SNAP judgments) which is robust to noise and which should dramatically decrease the burden on language researchers.

Arguments in favor of quantitative methods in syntax and semantics research

A. The current error rate in informal linguistic judgments is not acceptable.

SSA accept that there may be published judgments that would not be found in large-scale experiments, but they note that it is important to know the rate at which such examples occur. Consequently, SSA analyzed the judgments from 148 randomly sampled English acceptability judgments from Linguistic Inquiry (LI) 2001-2010. 127 out of these 148 experiments (86%) resulted in significant effects in the predicted direction using magnitude estimation; 130 (88%) resulted in significant predicted effects using Likert ratings; and 140 (95%) resulted in significant predicted effects in a forced choice experiment (where all values here are obtained using mixed models, which are most appropriate for this type of data (Barr, Levy, Scheepers, & Tily, 2013; Bates, Maechler, Bolker, & Walker, 2014)). 7 of the 148 experiments (5%) do not show predicted effects in any of the 3 experiments. Because SSA believe the forced choice judgments to be most reliable, SSA generalize from the 95% rate that informal acceptability intuitions reported in research articles on generative syntax have similar statistical properties as quantitative experiments comparing acceptability ratings for various experimental items by naive participants, because each allows an error rate of approximately 5%. That is, because a 5% error rate is the acceptable standard in cognitive psychology experiments, this error rate should also be acceptable in linguistic judgments. SSA write: "The field of experimental psychology has, by consensus, signaled a willingness to tolerate a divergence of 5% over the long run between the decision to classify differences as statistically significant and whether there is a real difference between the conditions" and that, while they do not unqualifiedly endorse 5% as an acceptable error rate, they find it to be a "reasonable starting point for the current discussion."

Following Gibson, Piantadosi & Fedorenko (2013), we do not believe that a 5% error rate is acceptable. Much of the recent debate on quantitative methods in syntax and semantics has focused on whether or not a significant p-value ($p < .05$) is obtained through a quantitative experiment (SSA and see Sprouse & Almeida (2012) for more discussion of effect size and statistical power in linguistic acceptability judgments). If the null hypothesis can be rejected, SSA state that the syntactic judgment "replicates." While SSA do not make a strong claim as to the appropriate role of Null Hypothesis

Significance Testing (NHST) in syntax and semantics judgments, we believe that there *is* a place for understanding the significance of judgments in formal experiments but we do not believe that the standards developed in the NHST paradigm are unproblematically applicable to linguistic judgments.

A 5% false-positive threshold in behavioral experiments is not a baseline error rate, since often a study will report a p-value much lower than the minimum threshold required for publication. A 5% error rate in linguistic acceptability judgments, however, suggests that 5% of all judgments actually do diverge from the results of a formal experiment. If the average linguistics paper has 33 examples (the average number of US-English examples found in the papers examined by SSA), then *every* paper is likely to contain an erroneous judgment: 1.64 on average. We think it is a mistake, then, to equate the 5% false-positive *threshold* in NHST with the 5% false-positive *rate* of informal acceptability intuitions. The NHST paradigm assumes that one has performed statistical significance testing for each particular effect under consideration; the $p < .05$ threshold is an easy way to classify the results of these tests, but it does not substitute for the important quantitative information gathered about each individual effect. On the other hand, with informal acceptability intuitions, the *method itself* has a 5% overall false-positive rate. The method provides no quantitative information about any individual effect, other than the fact that it has a 5% chance of being false.

B. Only formal experiments can give detailed information on the size of effects.
In the case of acceptability judgments, it is rarely the case that researchers actually care whether a sentence is some infinitesimally tiny amount better than another one. In fact, given a large enough sample size, one is likely to be able to find a statistically significant difference between *any* two sentence types that minimally differ. It is far more informative to investigate the size of the effect. In a rating study, the effect size can be measured as the difference in mean rating between Sentence A and Sentence B in terms of standard deviations. In a forced choice study, effect size can be estimated by the proportion of participants who choose Sentence A over Sentence B. Having a standardized system for obtaining and reporting native speaker judgments would allow

readers to know just how strong the generalization in question is. That is, when we see two sentences being compared, one with a * and one without, does that mean that 51/100 people would prefer the unstarred sentence? 90/100? 100/100? To be sure, we do not think it is necessary, or even useful, to apply a uniform quantitative threshold for acceptability contrasts. Our point is simply that *some* quantitative information about the size of the effect is a prerequisite for interpreting any individual acceptability contrast in a theoretically meaningful way.

C. Informal linguistic experiments make it difficult for researchers who either (a) are from other fields or (b) don't speak the target language in the materials.

Even if the informal linguistic judgments in journals agreed with results from formal experiments 100% of the time, there are still important reasons for performing formal experiments. Reporting statistically valid inferences about sentence judgments would make linguistics more accessible to researchers in other fields and to researchers who are unfamiliar with the language in question. Formal experiments using a consistent methodology make it easy to compare effects across languages.

D. It is *not* too costly and time consuming to do formal experiments, especially if one uses SNAP Judgments.

Another common concern is that formal experiments in syntax and semantic are too costly in time and money to justify the benefits (Culicover & Jackendoff, 2010). Although there is some cost to doing an experiment, the cost is now relatively low because of the existence of crowd-sourcing platforms like Amazon.com's Mechanical Turk that can provide robust results for cognitive behavioral experiments (Crump, McDonnell, & Gureckis, 2013; Mason & Suri, 2012; Sprouse, 2011). Such platforms provide cheap, reliable, fast labor, and there is free software available to perform such experiments (e.g., Turkolizer, Gibson, Piantadosi, & Fedorenko, 2011). In a syntax judgment experiment on Mechanical Turk, the researcher typically posts a survey consisting of a set of sentences (either visually or through audio recordings in order to capture prosody or other acoustic effects) and asking for a forced choice judgment, a

rating, or some other measure of acceptability. Participants fill out these surveys, and the researcher receives the data—usually within a few hours.

Still, it may seem like overkill to run a large-scale experiment to find out that "Rat cat ate the" is a less good English sentence than "The cat ate the rat." Moreover, not all researchers have easy access to Mechanical Turk or sufficient funding to run large-scale experiments. This lack of access to crowdsourcing platforms may especially affect those outside the United States (where Mechanical Turk is less readily available) and those who do fieldwork working on less widely spoken languages. As a result, many researchers eschew formal experiments altogether, leading to a gulf between theoretical syntax methods and experimental syntax methods. Like Myers (2009), we agree that there is space for a bridge between large-scale formal experiments and informal judgments. In order to address this issue, we propose the SNAP Judgments paradigm, which makes obtaining formal linguistic acceptability ratings easier and cheaper whether they are performed in the field, in the lab, or over the Internet.

2. Evaluation of Judgments from the Literature

We sampled a new set of sentences from these same 2001-2010 *Linguistic Inquiry* issues that SSA evaluated, and tested them in a rating experiment and a forced-choice experiment.[1]

2.1 Ratings Experiments

Participants

240 workers with US IP addresses were recruited through Amazon's Mechanical Turk crowdsourcing platform. 11 participants were excluded from the analysis because they did not self-identify as native speakers of English, leaving 229 participants. An additional 5 were excluded because they, on average, gave numerically higher ratings to the hypothesized unacceptable forms than to the hypothesized acceptable forms. These participants were thus probably not doing the task.

[1] SSA also did a magnitude estimation study, but, because those results were very similar to the rating study, we did not include magnitude estimation.

Stimuli

PG and an undergraduate assistant went through all the *Linguistic Inquiry* articles from 2001-2010 in which US English contrasts were presented and which were sampled from by SSA. We selected only contrasts that i) were English, ii) directly compared a grammatical and an ungrammatical sentence (irrespective of the particular judgment reported; assuming that OK>?>??>*?>*) and iii) were not primarily dependent on interpretation. This resulted in a total of 814 contrasts. Of these, 41 contrasts had already been tested by SSA. From the remaining 773 examples, we randomly sampled 101 contrasts.

For 96 of the contrasts sampled, JH constructed a template illustrating which properties of sentences other than the syntactic parse were allowed to vary across experimental items and which were not. Next, JH constructed an example item based on the original contrast reported in the paper. 6 templates/sample item pairs were assigned to 16 MIT undergraduate students in MIT's 9.59J Laboratory in Psycholinguistics class (taught by EG; MIT 24.900 Introduction to Linguistics prerequisite). Students were asked to create 10 experimental items (hypothesized grammatical/ungrammatical pairs; 20 sentences total) for the contrast they were assigned. Student items were hand-checked and corrected by PG and JH. For each contrast, we tested the ten student items, the original sentence pair reported in the research article and JH's sample item, resulting in 12 experimental items per contrast. For the 5 additional contrasts in our sample, 11 items were constructed by JH, which, together with the original sentence pair reported in the paper, also resulted in 12 items for those 5 contrasts. All 1212 contrast pairs were divided into 4 lists of 303 sentences each through a Latin square.

Procedure

Participants were asked to read each of the 606 sentences out loud to themselves and rate its naturalness on a Likert scale from 1-7. Order of presentation was randomized for every participant.

Results

After running the experiment, we noticed that 21 of our sentences had minor spelling mistakes or errors in punctuation. To be conservative, we excluded the 21 items these sentences from the analysis reported below. These errors were fixed in the subsequent forced choice experiment. We also noticed that one contrast was constructed erroneously, in that it did not represent the intended contrast in its source article. We therefore excluded this item from the analysis, both here and in the forced choice experiment.

To eliminate some of the effect of different participants using the rating scale differently, ratings for each participant were z-transformed (mean and standard deviation estimated within participants). For each item in each contrast, we then calculated a mean z-score and averaged these together to get an overall z-score for the "acceptable" sentence and the "unacceptable" sentence in each contrast. The effect size is the difference between these two z-scores.

All 100 contrasts showed a numerical trend in the predicted direction. Following Sprouse & Almeida (2012), we computed Cohen's d for each contrast (Cohen, 1994). Cohen's d is a measure of effect size that is equal to the difference in means between the two conditions (in this case, the acceptable condition vs. the unacceptable conditions) divided by the standard deviation of the data. Using Cohen's recommended effect size windows, we find 19/100 effects to be small effects ($d < .5$); 15/100 to be medium effects ($.5 < d < .8$); and 66/100 to be large effects ($d > .8$). Of the 19 small-effect contrasts, 7 actually have a Cohen's d of $< .2$, which is the minimum value that Cohen posits for a "small effect."

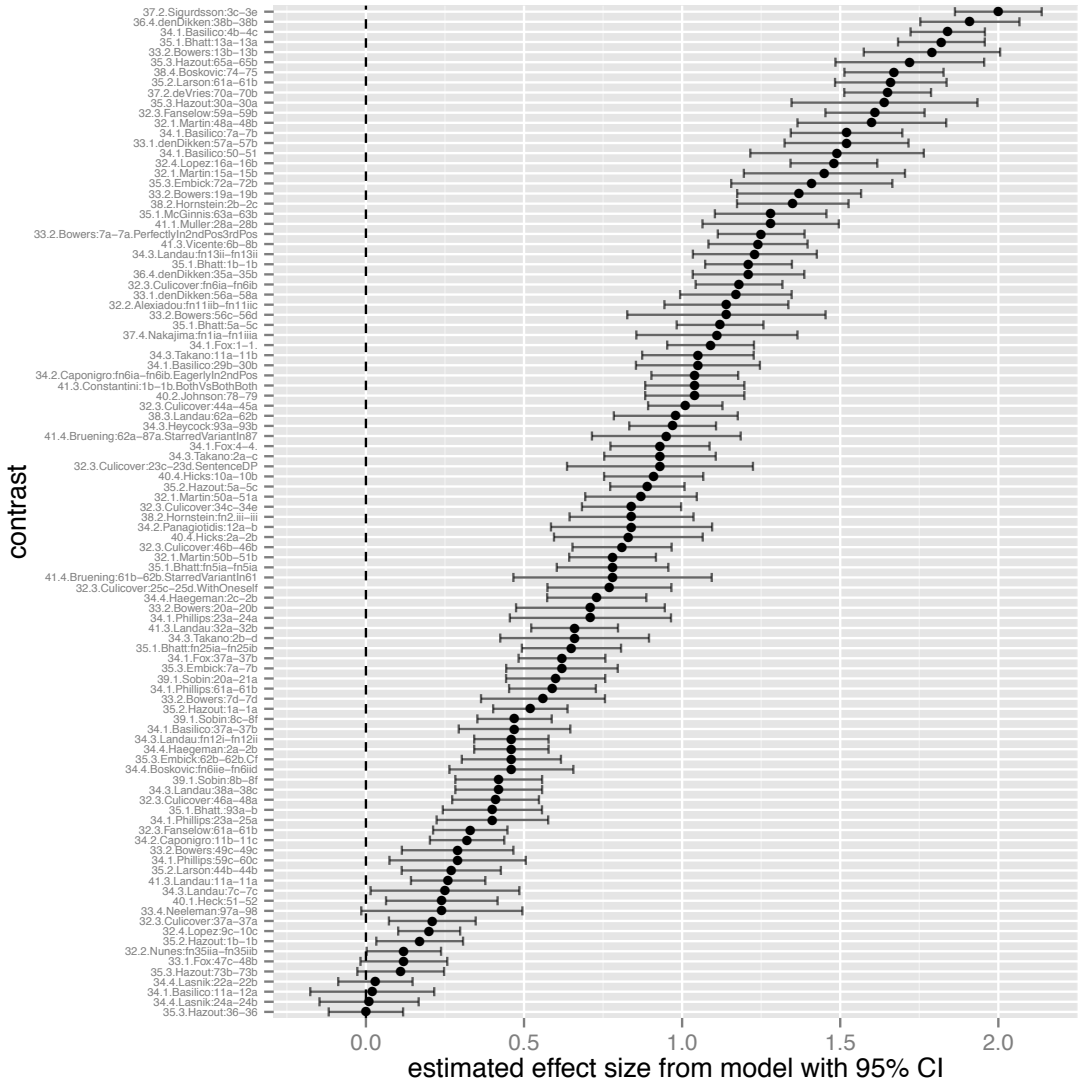

Figure 1: Each point is an effect size for the ratings experiment listed on the y-axis with 95% confidence intervals estimated from the linear mixed effect model. When the error bars extend through 0, the effect is not significant.

To control for individual variation by participant and item, we fit a linear mixed effects model with a sum-coded predictor for hypothesized acceptability (that is, whether or not the sentence is a reported as "acceptable" or "unacceptable" in the source *LI* paper) and random intercepts for both participant and item and random slopes for grammaticality grouped by both participant and item (random effects). The estimated coefficient for grammaticality is a best-guess estimate of the size of the effect, which is in this case the

difference in z-score rating between the "acceptable" variant and the "unacceptable" variant, after controlling for participant and item effects. The model also gives us standard error estimates on this output. Figure 1 plots the effect size estimates and 95% confidence-intervals estimated from the mixed effect model. Despite the fact that many of the papers from which these examples were drawn talk about these contrasts categorically as either grammatical or ungrammatical, Figure 1 reveals that the effect sizes of randomly chosen linguistic judgments do not show any discrete jumps (which one might expect given the frequent discussions of "grammatical," "marginal," or "ungrammatical" sentences) but rather form a continuum from no effect to huge effects.

To assess significance in an NHST framework, we used the linear mixed effect model described above fit using the R statistical programming language (R Team & others, 2012) and the `lme4` package (Bates et al., 2014). We performed a χ^2-model likelihood ratio test comparing a model with a fixed effect for grammaticality to the intercept model (a model assuming a single mean for both hypothesized grammatical and ungrammatical sentences) leaving the random effects structure intact in both models (Agresti, 2002; Baayen, Davidson, & Bates, 2008; Barr et al., 2013). This test asks whether the hypothesized grammaticality improves data likelihood significantly given the intercept, given normally distributed participant and item means, and normally distributed effect sizes of grammaticality for every participant and item. In other words, does hypothesized grammaticality explain a significant amount of variance in judgments?

To assess the statistical power of this experiment (the likelihood that we correctly detect a true effect), we simulated results using the same number of subjects and items that were used in the analysis.[2] The power analysis showed that, for a true effect size of 0.4, where effect size is the difference between the z-scores of the hypothesized grammatical and hypothesized ungrammatical sentences, we have a 96% chance of detecting a true effect at α=.05. 81% of our contrasts are estimated to have effect sizes at least this big. For a

[2] To simulate the random effect structure, we simulated subjects and items from the unconditional covariance matrix estimated from the mixed effect model.

true effect size of 0.2 (which would be small for experiments like these and would suggest very little difference between the two sentences), we have a 63% chance of detecting a true effect. The analysis shows that 92 of 100 contrasts in our random sample show significant effects in the predicted direction (92%). These results are summarized in Table 1.

Discussion

Of the 100 contrasts, the majority showed the predicted effect robustly. Of the eight that did not show a significant result, four showed clear trends in the predicted direction (35.3.Hazout:73b-73b, 33.1.Fox:47c-48b, 33.4.Neeleman:97a-98, 34.3.Landau:7c-7c) with estimated effect sizes over .1 and *ps* < .15. Four of the contrasts (35.3.Hazout:36-36, 34.4.Lasnik:24a-24b, 34.1.Basilico:11a-12a, 34.4.Lasnik:22a-22b) showed only numerical tendencies, with no clear trend in the predicted direction. We discuss the examples that did not show the predicted effects in more detail in Appendix D.

2.2 Forced Choice Experiments

Participants

240 workers with US IP addresses were recruited through Amazon's Mechanical Turk crowdsourcing platform. We excluded participants who took the test more than once and those who did not self-identify as native speakers of English. We also excluded participants who chose the hypothesized "acceptable" option less than 60% of the time. Because most participants chose the "acceptable" option the vast majority of the time, those who were choosing it less than 60% of the time were likely not doing the task. After these exclusions, 201 participants remained.

Stimuli

The stimuli were the same as the stimuli in the ratings experiment, except with minor spelling corrections.

Procedure

Participants were asked to read each pair of sentences out loud to themselves and choose which sounded more natural. The order of presentation was randomized across participants.

Results

As in the rating experiment, one contrast was removed due to an error in how it was constructed, such that it did not represent the intended contrast in its source article. We found a wide array of effect sizes in the remaining sample of 100 contrasts, where effect size is taken to be the proportion of trials in which the hypothesized acceptable sentence is preferred. First, 6 of 100 contrasts trended in the opposite direction from the predicted direction. The remaining 94/100 sentences showed an effect in the predicted direction. 81/100 had an effect size greater than .75, and roughly half (52/100) had an effect size greater than .9. Overall, these results demonstrated smaller effects than those reported by SSA, but were qualitatively similar.

To control for individual variation by participant and item, we fit a logistic linear mixed effects model predicting whether or not the participant preferred the hypothesized acceptable sentence over the hypothesized unacceptable one. We included a fixed effect intercept (that is, whether or not the sentence was reported as acceptable or unacceptable in the original article) and random intercepts for both participant and item. The estimated coefficient for the intercept was a logit-transformed best-guess estimate of how often a contrast would show a preference for the hypothesized acceptable form after controlling for participant and item effects. The model also provided us standard errors for the estimates, from which we can calculate 95% confidence intervals (CIs). Figure 2 plots the effect size estimates and 95% CI output from the mixed effect model.[3] As with the rating study, we see no discrete jumps but rather a continuum of effect sizes.

[3] A few of the forced choice experiments had mean proportions close to 1. Although logistic regressions are not accurate when proportions are near 1 or 0, this inaccuracy doesn't matter here, because these effects were very large.

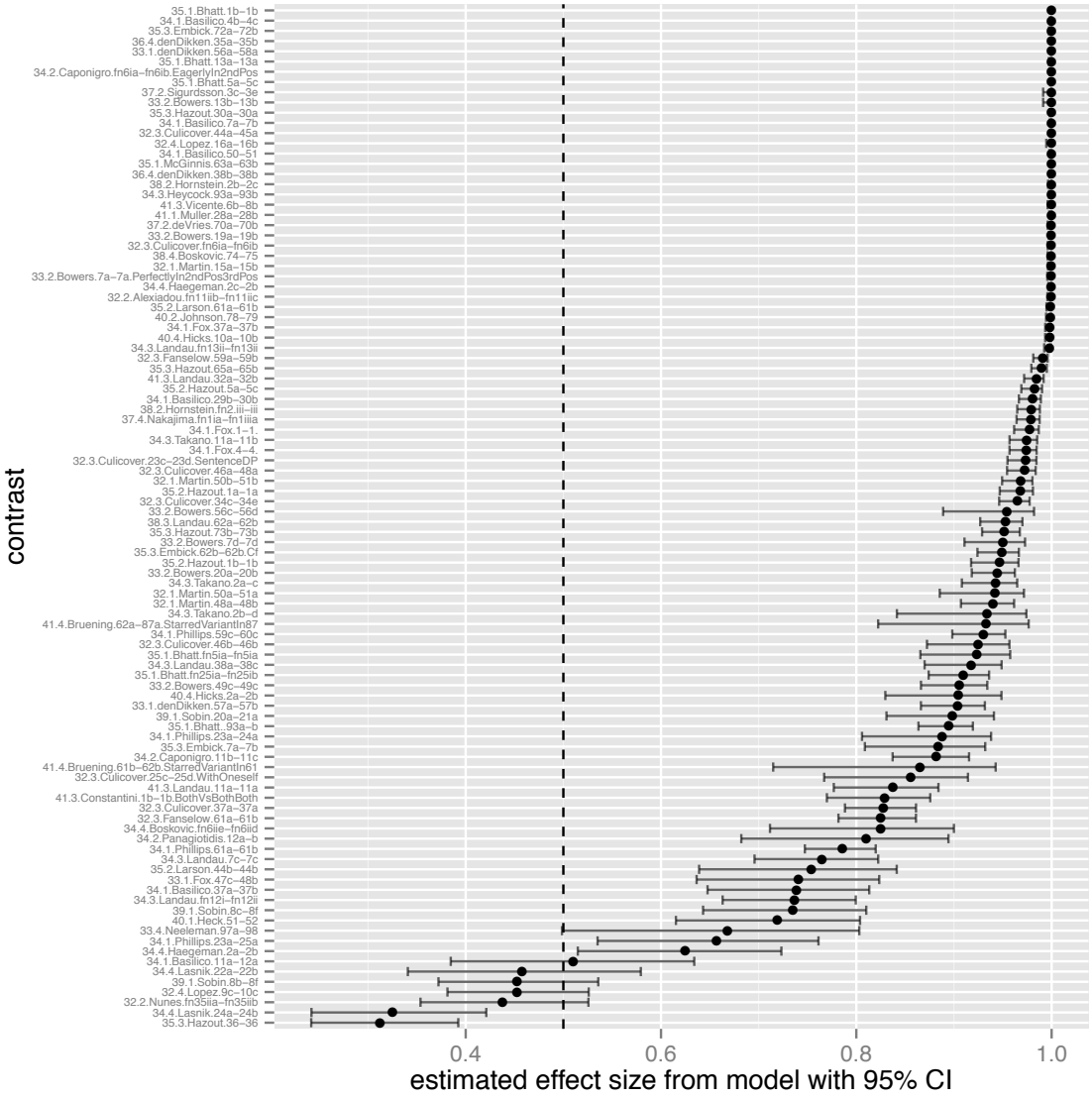

Figure 2: Each point in this figure is an effect size for the forced choice experiment listed on the y-axis with 95% CIs estimated from the linear mixed effect model. When the error bars extend through the line at .5, the effect is not significant. The error bars that appear entirely to the left of the line at .5 are significant in the opposite direction of what was predicted.

To assess significance in an NHST framework, we used the logistic linear mixed effect model described above and used the z-value to calculate a p-value.[4] As before, we calculated statistical power for several possible true effect sizes. If the underlying effect

[4] With only one fixed effect (the intercept) and thus no collinearity among fixed effects, it is unnecessary in

size was 0.7 (meaning 70% of participants prefer the "good" sentence), for instance, we would have an 80% chance of detecting a true effect. The results appear in Appendix A. 92/100 showed significant trends in favor of the hypothesized acceptable version. Two items showed non-significant trends in favor of the hypothesized acceptable version, and four showed non-significant trends in favor of the hypothesized unacceptable version. Finally, two contrasts showed significant effects in the opposite direction of the predicted effect.

Discussion

In summary, 89/100 contrasts showed a significant effect in the predicted direction in both experiments, and 95/100 contrasts showed a significant effect in the predicted direction in at least one of the two experiments. These results therefore suggest that, whereas most published linguistic judgments are consistent with the results found by formal experiments, they are critically different.

Our results are quite similar to the results reported by SSA. However, SSA interpret divergences between formal experimental results and researchers' intuitions differently than we do. In particular, when the judgments of the *Linguistic Inquiry* researchers diverge from those of the participants of a formal experiment, SSA raise the question of whether we should trust the researchers' intuitions or whether we should trust the results from the formal experiment. SSA suggest that many of the divergent judgments in their experiments do not in fact have meaningful theoretical implications, mostly because of differences between the materials that they constructed and the intended contrast in the original article. In the contrasts that we considered, we believe that the materials that we used are representative of the target contrasts. We therefore believe that we can trust the results from our formal experiments. At the very least, the fact that a large sample of native English speakers have a different judgment from the original researcher should be cause for concern to the original theoretical claims.

---

this case to do a likelihood ratio test to assess significance.

The contrasts that did not show significant effects in the expected direction in our experiments are listed in Table 1. We briefly discuss several of these cases in Appendix D.

Source article	Rating experiment	Forced choice experiment
35.3.Hazout.36-36	+	-*
34.4.Lasnik.24a-24b	+	-*
32.2.Nunes.fn35iia-fn35iib	+*	-
32.4.Lopez.9c-10c	+*	-
39.1.Sobin.8b-8f	+*	-
34.4.Lasnik.22a-22b	+	-
34.1.Basilico.11a-12a	+	+
35.3.Hazout:73b-73b	+	+*
33.1.Fox:47c-48b	+	+*
33.4.Neeleman:97a-98	+	+
34.3.Landau:7c-7c	+	+*

Table 1: Contrasts that did not show significant effects in the predicted direction in either the rating experiment or forced choice experiment. A plus means that the contrast showed an effect in the predicted direction, whereas a minus means that the effect went in the opposite direction from what was predicted. A * next to the plus or minus means that the effect was significant at $p < .05$. Yellow indicates effects that went in the predicted direction but were not significant; pink indicates effects that went in the opposite direction and were not significant; and red indicates effects that went significantly in the opposite direction.

3. SNAP Judgments

Across our two experimental paradigms, about half of the examples sampled from *LI* showed a strong effect (Cohen's d > .8). As many researchers have pointed out, it seems unnecessary to run a formal experiment in these cases, because they seem intuitively

obvious. But it is not good scientific practice to rely only on intuition. To improve efficiency while maintaining scientific rigor, we propose a method for not having to run a large experiment while still being able to reach a statistically justified conclusion. We call this method a Small N Acceptability Paradigm for linguistic judgments (SNAP Judgments). We focus on forced choice judgments because they are simpler, have greater statistical power, and correspond more closely to the sorts of binary judgments between two sentences that often appear in linguistics papers.

The basic idea is to be able to draw a statistically valid conclusion based on the data of just a few participants. To do so, we want to determine how many participants we would need to consult in order to be reasonably confident that we have a meaningful result. The simplest way to think of this is to think of each experimental trial as a flip of a weighted coin, where Heads corresponds to a preference for Sentence A and Tails corresponds to a preference for Sentence B. What we want to know is how often the coin will come up heads over a large sample of flips—that is, across many trials, how often Sentence A would be picked over Sentence B. We want to make inferences about the underlying parameter *p*, which is a probability between 0 and 1 that tells us how often Sentence A would be preferred over Sentence B. If *p* was .75, that means that 75% of the time, Sentence A would be preferred.

If we ask 5 people which sentence they prefer and they all prefer Sentence A, can we conclude with certainty that everyone will prefer Sentence A? No: it's possible that, if we ask a sixth person, she will prefer Sentence B. We still don't know if the probability of someone in the larger population preferring Sentence A is 90% or 70% or 10% (in which case getting 5 people who prefer Sentence A was an unlikely—but possible—accident). We want to try to infer *p*.

Here, we estimate *p* in a Bayesian framework by placing a beta prior distribution over the effect sizes found in our forced choice experiment above. In effect, this technique lets us supplement the result of our experiment by adding in prior information about what typical linguistic contrasts are like. If we were very confident that most effect sizes were

large, then even after collecting just one data point, we might conclude that the effect was likely large. If we were very confident that the effect size was near .5, we might still think the effect is near .5 even if we asked 100 people and they all said the same thing! To decide how we should set our prior for linguistic judgments, we empirically estimated the parameters in the beta distribution by fitting the data we obtained in the Forced Choice experiment above. We found that the best prior was Beta(.6, .6). (For a much more detailed description, see Appendix E.)

Using the prior, we can ask how many people we need to survey (while getting a unanimous result) in order to put our estimated effect size over .90 and our lower 95% confidence bound at or above .75. If we ask 3 people whether they prefer Sentence A or Sentence B and they all say Sentence A, the expected mean proportion of people who would pick Sentence A is .86 with a 95% confidence interval of [.53, 1]. If we asked 5 people and they all answered Sentence A, the mean goes to .9, with a 95% CI [.67, 1]. And if we ask 7, the mean is .93 [.75, 1]. Thus, where n=7 and the results of the experiment is unanimous, we get an expected mean of .93 and a lower 95% bound of .75. We believe that this is sufficient to be confident that the result would obtain in a larger experiment.[5]

Testing SNAP Judgments

We can test the efficacy of the SNAP Judgments proposal in the sample that we already have. For each experiment, we randomly sampled 7 data points, each with a unique participant and a unique item. We then focused on only the experiments that gave us a unanimous result among the 7 randomly sampled data points. We can think of this as a simulated outcome for a SNAP Judgment. We repeated this procedure 100 times for each of the 100 experiments. On average, 46 of them produced a unanimous 7-participant result. Of only those trials that produced a unanimous result, when we looked at the result of the full experiment with all participants and all items, the mean across those

[5] Note that, in a frequentist paradigm in which we do not use the prior information obtained from the existing data, we would have to ask 12 people in order to get a 95% CI of [.75, 1] (using the Wilson score

experiments was .93 with a 95% CI of [.79, .99]. The lowest value obtained in any of the 100 simulations was .72. Compare this to our Beta prior, which gave us an expected mean of .93 [.75, 1]. The empirical test is very consistent with the results obtained using the Beta prior: the means match exactly, and the lower bound for the 95% CI is slightly conservative (.75 compared to .80 in the empirical test), which suggests that the Beta(.6, .6) prior is a good model for this data.

To make sure that we were not overfitting the data, we also ran 100 simulations where we divided the data in half, such that we used 50 of the 100 contrasts to set the shape parameters for the beta prior and sampled 7 participants each from the other 50 contrasts to get our empirical estimate, as above. Across the 100 simulations, the mean α and β was .6, with standard deviation .04. In the empirical sample from the held-out data set of only those trials in which the judgment was unanimous across all 7 unique participants and items, the mean was .93 with a standard deviation of .004. Thus, even when the beta distribution shape parameters are learned and tested on independent data, we get similar results—which suggests that, as long as the samples from *LI* are representative of the sorts of linguistic judgments to which this method will be applied, the prior is robust.

One may still, at this point, wonder why we recommend 7 data points for a SNAP Judgment as opposed to any other number. We believe that n=7, with a unanimous result, provides a robust generalization sufficient for most linguistic judgments. And crucially, it would *never* give a significant result in the wrong direction in any of the samples that we tested from *LI.* Of course, this framework can also provide expected values and 95% CIs for any other n. For instance, from 3 unanimous judgments, the expected mean in a large experiment is .86 with a 95% CI of [.53, 1]. So, with only 3 judgments, we can be reasonably confident that the result of a large experiment would not be less than .50. This would be a statistically valid conclusion and would be better than just relying on intuition and not reporting any quantitative results at all.

Nonetheless, we reiterate that, in most cases, researchers are not interested in whether

---

interval for computing binomial CI's).

Sentence A is some epsilon better than Sentence B but want to make a broader generalization. By using n=7 as a field-wide standard for SNAP Judgments, we can begin to develop intuitions about how to interpret these results. Below, however, we discuss what to do when 7 participants are not readily available and give guidelines for conclusions that can be drawn based on even fewer data points.

Recommendation for SNAP Judgments

Given the proposal and evaluation above, we make the following recommendations for SNAP Judgments:

1. **To ensure the applicability of our empirical estimates, SNAP Judgments should only be used when the researcher is confident that the effect will be unanimous.** If one does not believe that the results of the survey will be unanimous, it is better to do a large N rating study, which gives more gradient information, or a forced-choice study, which has more statistical power (Sprouse and Almeida 2012).

2. **Construct 7 unique contrasts (each consisting of Sentence A vs. Sentence B, where one of the two sentences is hypothesized to be more acceptable than the other) and make sure that the 7 contrasts vary in lexical content and whatever other factors may influence the acceptability of the sentences in question. Present each contrast to a unique naïve participant and ask for a forced choice judgment.** This could be done using Amazon's Mechanical Turk (paying perhaps 5 cents for 1 judgment such that the whole experiment will cost less than 50 cents) or by simply asking for judgments from students, friends, informants, or colleagues who are naïve to the experiment in question.[6]

[6] Note that, for experiments run on Mechanical Turk, there are occasionally participants who click randomly or do not pay attention to the task. As a result, it is good practice to also include questions with known answers, such as simple comprehension questions about the target sentence. Participants who do not correctly answer these simple questions can be excluded from the analysis.

Researchers in the field can use the same procedure, and in extreme situations in which access to speakers is severely limited (as in the case of endangered languages), it may be necessary to poll fewer than 7 participants. Table 2 below lists the conclusions that can be drawn from even smaller SNAP Judgments. For these extreme cases, we note that even the use of 3 independent data points shrinks the 95% credible interval substantially. Thus, these recommendations need not dramatically slow or impede the pace of fieldwork: fieldworkers often do have 3 independent data points for a construction or contrast in question and can report those judgments quantitatively.

3. **If all 7 naïve participants agree with the predicted result, one can conclude the following: the predicted mean for the full experiment is .93, with a 95% CI of [.75, 1] and a 99% CI of [.62, 1].** That is, one can be 95% confident that at least 75% of people would agree with the intuition. See Table 2 below for guidelines when using participant numbers other than 7.

4. **If the judgments are not all in agreement, then the intuition that the result would be unanimous is wrong.** This is not a failure of SNAP Judgments but one of its major advantages: giving the opportunity to explore where there is variation among items. If there is not agreement among participants, look at the seven items. Is there a pattern among items that do not show the expected effect? Do variations in word choice or prosody or context seem to affect the results? If so, this might be an opportunity to further explore the nature and size of the effect in question. If new hypotheses are generated by the SNAP Judgment experiment, one can then test these hypotheses. At that point, we recommend a formal experiment with a larger number of items and participants to understand the size of the effect and possible sources of variation. (Note that, if one were to simply perform SNAP Judgments repeatedly on the same grammatical contrast, one might eventually find a string of unanimous responses just by chance. This is the multiple comparison fallacy, and, in that case, the statistical guidelines here would

not be directly applicable.)

# unanimous	mean with 95% CI	mean with 99% CI
1	0.73 [0.22, 1]	0.73 [0.08, 1]
2	0.81 [0.41, 1]	0.81 [0.23, 1]
3	0.86 [0.53, 1]	0.86 [0.35, 1]
4	0.88 [0.61, 1]	0.88 [0.44, 1]
5	0.90 [0.67, 1]	0.90 [0.51, 1]
6	0.92 [0.71, 1]	0.92 [0.57, 1]
7	0.93 [0.75, 1]	0.93 [0.61, 1]
8	0.93 [0.77, 1]	0.93 [0.65, 1]
9	0.94 [0.79, 1]	0.94 [0.68, 1]
10	0.95 [0.81, 1]	0.95 [0.71, 1]

Table 2 This table shows the mean with the (a) 95% credible interval and (b) 99% credible interval where the number of unanimous participants in the experiment varies from 1 to 10. For instance, in an experiment with 3 participants, all of whom choose unanimously, one can estimate a mean of .86 with a 95% credible interval of .53 to 1 and a 99% credible interval of .35 to 1.

General discussion

In this paper, we have replicated the empirical findings of SSA. In a sample of 100 contrasts from *Linguistic Inquiry,* we found a wide range of effect sizes in both a rating experiment and a forced choice experiment. Small, medium, and large effects are all well represented in the data set. 89% of these syntactic judgments reported in *LI* show significant effects in the predicted direction in two types of large-scale formal experiments. In the remaining 11%, there are varying levels of uncertainty about the judgments elicited. In all of these cases, we believe that the formal experiments uncover interesting sources of variation that could illuminate the theoretical questions at stake and improve the papers in which they appeared.

Moreover, we used the empirical results presented here as a foundation on which to build a prior distribution of what syntactic judgments can be expected to look like. Specifically, we found that a Beta distribution is a good fit to the distribution of probabilities found in

the forced-choice experiment and used it to recommend a new paradigm for small-sample acceptability experiments.

We believe that the SNAP Judgments paradigm will make it easier and cheaper for language researchers to obtain statistically justified linguistic acceptability judgments. Specifically, in instances where a researcher is confident that a judgment would produce a unanimous result across 7 participants, we recommend a forced-choice experiment with 7 participants and 7 items. If the result is unanimous, the results of this small N experiment can be combined with a Beta(.6, .6) prior to give a predicted effect size of .93 with a 95% CI [.75, 1].

Of course, SNAP Judgments cannot solve all open questions in linguistic methodology. The guidelines presented here do not, for instance, solve the problem of how to write good items that generalize to the contrast in question. Even if a researcher were to test 100 versions of some specific syntactic generalization, she may have overlooked some special case in which the generalization does not hold due to lexical, pragmatic, or contextual factors. Statistics should supplement, not replace, careful thought about syntax semantics. Moreover, we do not claim that the SNAP Judgment proposal solves all questions about how acceptability judgments relate to underlying questions about grammaticality.

We do, however, believe that there is abundant value in being able to attain cheap and easy quantitative data in syntax and semantics. Not only will the collection of more quantitative data help preclude erroneous analyses from entering the literature, but it will also enable us to continue building a body of empirical data. This body of data will make it easy for researchers to uncover new and interesting empirical phenomena and place those phenomena into a larger quantitative framework so as to understand gradient effects and sources of variation in linguistic data. For instance, using the data from this study (available for download on the Open Science Foundation at the URL osf.io/5wm2a), one can easily test a new contrast and plot it alongside the 101 phenomena tested here in order to ask what other sentences the contrast patterns like, how much variation there is

among participants for that contrast, and how sensitive the contrast is to variation in lexical items or context.[7] Knowing whether a particular proposed effect is small, medium or large--and knowing exactly what that means relative to other published judgments—is a worthwhile goal.

Given the ease with which SNAP Judgments can be attained, we believe that the time and effort required is not much more than what a linguist already spends when discussing judgments with friends, colleagues, and students or what a field linguist spends eliciting judgments from informants. By treating syntax and semantics questions empirically, we can develop standardized quantitative methods that can be understood across disciplines, across languages, and by future generations of researchers.

[7] The mixed model approach used in this paper is particularly useful for analyses of this sort by allowing researchers to extract parameters that estimate variance by participant and item.

References

Agresti, A. (2002). *Categorical data analysis* (Vol. 359). John Wiley & Sons.

Arppe, A., & Järvikivi, J. (2007). Take empiricism seriously! In support of methodological diversity in linguistics. *Corpus Linguistics and Linguistic Theory*, *3*(1), 99–109.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. Retrieved from http://CRAN.R-project.org/package=lme4

Chomsky, N. (1957). *Syntactic structures*. Walter de Gruyter.

Chomsky, N. (1986). *Barriers*. Cambridge MA: MIT Press.

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997.

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Sage Publications Thousand Oaks, CA.

Cox, D., & Mayo, D. (2011). Statistical Scientist Meets a Philosopher of Science: A Conversation. *Rationality, Markets and Morals*, *2*.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

Culicover, P. W., & Jackendoff, R. (2010). Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences*, *14*(6), 234–235.

Featherston, S. (2005). Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua*, *115*(11), 1525–1550.

Gelman, A. (2012). Ethics and Statistics: Ethics and the Statistical Use of Prior Information. *CHANCE*, *25*(4), 52–54.

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. doi:10.1177/1745691614551642

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. Boca Raton, FL: CRC press.

Gibson, E., & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, *14*(6), 233–234. doi:10.1016/j.tics.2010.03.005

Gibson, E., & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, *28*(1-2), 88–124. doi:10.1080/01690965.2010.515080

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, *5*(8), 509–524.

Gibson, E., Piantadosi, S. T., & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes*, *28*(3), 229–240. doi:10.1080/01690965.2012.704385

Gross, S., & Culbertson, J. (2011). Revisited linguistic intuitions. *The British Journal for the Philosophy of Science*, *62*(3), 639–656.

Hartman, J. (2011). (Non-)Intervention in A-movement: some cross-constructional and cross-linguistic consequences. *Linguistic Variation* 11.2: 121-148.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. doi:10.3758/s13428-011-0124-6

Myers, J. (2009). The design and analysis of small-scale syntactic judgment experiments. *Lingua*, *119*(3), 425–444. doi:10.1016/j.lingua.2008.09.003

Schutze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. University of Chicago Press.

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, *115*(11), 1497–1524. doi:10.1016/j.lingua.2004.07.002

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

Sprouse, J., & Almeida, D. (2012). Power in acceptability judgment experiments and the reliability of data in syntax. *LingBuzz*. Retrieved from lingbuzz/001362

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, *134*, 219–248. doi:10.1016/j.lingua.2013.07.002

Team, R. C., & others. (2012). R: A language and environment for statistical computing.

Wasow, T., & Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua*, *115*(11), 1481–1496.

Appendix A: Rating Study Results

zbad is the average z-score for the hypothesized "bad" option. zgood is the average z-score for the hypothesized good option. Z.diff is the difference between zgood and zbad and is the effect size. beta is the estimate from the linear mixed effect moel, which has a standard error se and a t value tvalu. chi2 is the chi-squared value comparing the full model to an intercept-only model, and chi2p 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.

exp	zbad	zgood	z.diff	beta	se	tvalue	chi2	chi2p	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:fn35iia-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	<.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	<.0001	<.0001	TRUE	TRUE
35.2.Larson:44b-44b	0.54	0.8	0.27	0.27	0.08	3.5	10.76	0.001	<.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	<.0001	<.0001	TRUE	TRUE
32.3.Fanselow:61a-	-0.96	-0.63	0.33	0.33	0.06	5.24	22	<.0001	<.0001	TRUE	TRUE

61b											
34.1.Phillips:23a-25a	-0.02	0.39	0.4	0.4	0.09	4.5	12.78	<.0001	<.0001	TRUE	TRUE
35.1.Bhatt.:93a-b	-0.69	-0.29	0.41	0.4	0.08	5.02	14.82	<.0001	<.0001	TRUE	TRUE
32.3.Culicover:46a-48a	-0.2	0.21	0.41	0.41	0.07	6.3	22.52	<.0001	<.0001	TRUE	TRUE
34.3.Landau:38a-38c	-0.28	0.14	0.42	0.42	0.07	5.68	18.67	<.0001	<.0001	TRUE	TRUE
39.1.Sobin:8b-8f	-0.36	0.06	0.42	0.42	0.07	6.32	25	<.0001	<.0001	TRUE	TRUE
34.4.Boskovic:fn6iie-fn6iid	-0.34	0.12	0.46	0.46	0.1	4.54	13.78	<.0001	<.0001	TRUE	TRUE
35.3.Embick:62b-62b.Cf	0.4	0.86	0.46	0.46	0.08	5.54	17.72	<.0001	<.0001	TRUE	TRUE
34.4.Haegeman:2a-2b	-0.15	0.3	0.46	0.46	0.06	7.27	25.89	<.0001	<.0001	TRUE	TRUE
34.3.Landau:fn12i-fn12ii	-0.88	-0.42	0.46	0.46	0.06	7.73	27.48	<.0001	<.0001	TRUE	TRUE
34.1.Basilico:37a-37b	-0.37	0.1	0.47	0.47	0.09	5.31	16.56	<.0001	<.0001	TRUE	TRUE
39.1.Sobin:8c-8f	-0.32	0.15	0.47	0.47	0.06	8.02	35.62	<.0001	<.0001	TRUE	TRUE
35.2.Hazout:1a-1a	-0.86	-0.34	0.52	0.52	0.06	8.77	34.21	<.0001	<.0001	TRUE	TRUE
33.2.Bowers:7d-7d	0.56	1.12	0.56	0.56	0.1	5.48	17.11	<.0001	<.0001	TRUE	TRUE
34.1.Phillips:61a-61b	-1	-0.41	0.59	0.59	0.07	7.86	25.39	<.0001	<.0001	TRUE	TRUE
39.1.Sobin:20a-21a	-0.18	0.41	0.6	0.6	0.08	7.41	23.31	<.0001	<.0001	TRUE	TRUE
35.3.Embick:7a-7b	-0.47	0.14	0.62	0.62	0.09	6.98	22.55	<.0001	<.0001	TRUE	TRUE
34.1.Fox:37a-37b	-0.17	0.45	0.62	0.62	0.08	8.26	27.1	<.0001	<.0001	TRUE	TRUE
35.1.Bhatt:fn25ia-fn25ib	-1.08	-0.43	0.65	0.65	0.08	7.78	26.37	<.0001	<.0001	TRUE	TRUE
34.3.Takano:2b-d	-0.52	0.14	0.66	0.66	0.12	5.71	16.37	<.0001	<.0001	TRUE	TRUE
41.3.Landau:32a-32b	-0.6	0.06	0.66	0.66	0.07	8.98	29.15	<.0001	<.0001	TRUE	TRUE
34.1.Phillips:23a-24a	-0.41	0.3	0.71	0.71	0.13	5.28	14.52	<.0001	<.0001	TRUE	TRUE
33.2.Bowers:20a-20b	-0.53	0.18	0.71	0.71	0.12	5.84	16.21	<.0001	<.0001	TRUE	TRUE
34.4.Haegeman:2c-2b	-0.73	0	0.73	0.73	0.08	8.8	28.07	<.0001	<.0001	TRUE	TRUE
32.3.Culicover:25c-25d.WithOneself	-0.25	0.52	0.77	0.77	0.1	7.76	22.87	<.0001	<.0001	TRUE	TRUE
41.4.Bruening:61b-62b.StarredVariantIn61	-0.24	0.54	0.78	0.78	0.16	4.89	13.65	<.0001	<.0001	TRUE	TRUE
35.1.Bhatt:fn5ia-fn5ia	0.03	0.8	0.78	0.78	0.09	8.89	25.75	<.0001	<.0001	TRUE	TRUE
32.1.Martin:50b-51b	-0.49	0.29	0.78	0.78	0.07	11.65	42.71	<.0001	<.0001	TRUE	TRUE
32.3.Culicover:46b-46b	-0.32	0.49	0.81	0.81	0.08	9.89	30.41	<.0001	<.0001	TRUE	TRUE

40.4.Hicks:2a-2b	0.21	1.04	0.82	0.83	0.12	6.65	19.3	<.0001	<.0001	TRUE	TRUE
34.2.Panagiotidis:12a-b	0.07	0.92	0.84	0.84	0.13	6.5	19.02	<.0001	<.0001	TRUE	TRUE
38.2.Hornstein:fn2.iii-iii	-0.24	0.6	0.84	0.84	0.1	8.68	28.84	<.0001	<.0001	TRUE	TRUE
32.3.Culicover:34c-34e	-0.28	0.56	0.84	0.84	0.08	10.76	34.15	<.0001	<.0001	TRUE	TRUE
32.1.Martin:50a-51a	-0.21	0.65	0.87	0.87	0.09	9.2	27	<.0001	<.0001	TRUE	TRUE
35.2.Hazout:5a-5c	-0.67	0.21	0.89	0.89	0.06	15.54	42.33	<.0001	<.0001	TRUE	TRUE
40.4.Hicks:10a-10b	-0.8	0.11	0.91	0.91	0.08	11.63	36.63	<.0001	<.0001	TRUE	TRUE
32.3.Culicover:23c-23d.SentenceDP	0.2	1.12	0.92	0.93	0.15	6.11	16.76	<.0001	<.0001	TRUE	TRUE
34.3.Takano:2a-c	-0.36	0.58	0.94	0.93	0.09	10.1	27.12	<.0001	<.0001	TRUE	TRUE
34.1.Fox:4-4.	-1.01	-0.09	0.93	0.93	0.08	11.24	36.45	<.0001	<.0001	TRUE	TRUE
41.4.Bruening:62a-87a.StarredVariantIn87	-0.31	0.65	0.95	0.95	0.12	8.01	23.2	<.0001	<.0001	TRUE	TRUE
34.3.Heycock:93a-93b	0.04	1.01	0.98	0.97	0.07	14	38.46	<.0001	<.0001	TRUE	TRUE
38.3.Landau:62a-62b	-0.12	0.87	0.98	0.98	0.1	10.32	29.29	<.0001	<.0001	TRUE	TRUE
32.3.Culicover:44a-45a	-0.61	0.4	1.01	1.01	0.06	16.01	44.47	<.0001	<.0001	TRUE	TRUE
40.2.Johnson:78-79	-0.57	0.48	1.05	1.04	0.08	12.3	31.13	<.0001	<.0001	TRUE	TRUE
41.3.Constantini:1b-1b.BothVsBothBoth	-0.14	0.91	1.04	1.04	0.08	13.69	37.01	<.0001	<.0001	TRUE	TRUE
34.2.Caponigro:fn6ia-fn6ib.EagerlyIn2ndPos	-0.06	0.98	1.03	1.04	0.07	15.02	37.98	<.0001	<.0001	TRUE	TRUE
34.1.Basilico:29b-30b	-0.97	0.07	1.05	1.05	0.1	10.71	30.49	<.0001	<.0001	TRUE	TRUE
34.3.Takano:11a-11b	-0.92	0.12	1.05	1.05	0.09	11.4	31.77	<.0001	<.0001	TRUE	TRUE
34.1.Fox:1-1.	-1.01	0.08	1.09	1.09	0.07	14.94	46.32	<.0001	<.0001	TRUE	TRUE
37.4.Nakajima:fn1ia-fn1iiia	-0.91	0.21	1.12	1.11	0.13	8.69	23.29	<.0001	<.0001	TRUE	TRUE
35.1.Bhatt:5a-5c	-0.4	0.72	1.12	1.12	0.07	15.57	43.81	<.0001	<.0001	TRUE	TRUE
33.2.Bowers:56c-56d	-0.37	0.77	1.14	1.14	0.16	7.09	20.34	<.0001	<.0001	TRUE	TRUE
32.2.Alexiadou:fn11iib-fn11iic	-0.56	0.58	1.14	1.14	0.1	11.71	32.22	<.0001	<.0001	TRUE	TRUE
33.1.denDikken:56a-58a	-0.51	0.66	1.17	1.17	0.09	13.22	35.62	<.0001	<.0001	TRUE	TRUE
32.3.Culicover:fn6ia-fn6ib	-0.77	0.41	1.18	1.18	0.07	17.1	48.58	<.0001	<.0001	TRUE	TRUE
36.4.denDikken:35a	-0.26	0.95	1.21	1.21	0.09	13.25	37.76	<.0001	<.0001	TRUE	TRUE

-35b											
35.1.Bhatt:1b-1b	-0.52	0.69	1.21	1.21	0.07	17.76	50.67	<.0001	<.0001	TRUE	TRUE
34.3.Landau:fn13ii-fn13ii	-0.49	0.73	1.22	1.23	0.1	12.21	34.15	<.0001	<.0001	TRUE	TRUE
41.3.Vicente:6b-8b	-0.98	0.26	1.24	1.24	0.08	15.94	42.74	<.0001	<.0001	TRUE	TRUE
33.2.Bowers:7a-7a.PerfectlyIn2ndPos3rdPos	-0.42	0.82	1.25	1.25	0.07	17.18	40.23	<.0001	<.0001	TRUE	TRUE
41.1.Muller:28a-28b	-0.86	0.42	1.28	1.28	0.11	11.3	31.48	<.0001	<.0001	TRUE	TRUE
35.1.McGinnis:63a-63b	-0.35	0.94	1.28	1.28	0.09	14.49	36.9	<.0001	<.0001	TRUE	TRUE
38.2.Hornstein:2b-2c	-0.12	1.24	1.35	1.35	0.09	14.74	45.23	<.0001	<.0001	TRUE	TRUE
33.2.Bowers:19a-19b	-0.35	1.02	1.37	1.37	0.1	13.18	39.73	<.0001	<.0001	TRUE	TRUE
35.3.Embick:72a-72b	-0.37	1.05	1.41	1.41	0.13	11.17	30.27	<.0001	<.0001	TRUE	TRUE
32.1.Martin:15a-15b	-0.33	1.12	1.45	1.45	0.13	11.35	29.79	<.0001	<.0001	TRUE	TRUE
32.4.Lopez:16a-16b	-0.44	1.03	1.48	1.48	0.07	20.04	55.9	<.0001	<.0001	TRUE	TRUE
34.1.Basilico:50-51	-0.81	0.68	1.49	1.49	0.14	10.89	29.51	<.0001	<.0001	TRUE	TRUE
33.1.denDikken:57a-57b	-0.65	0.87	1.51	1.52	0.1	15.16	39.21	<.0001	<.0001	TRUE	TRUE
34.1.Basilico:7a-7b	-0.46	1.06	1.52	1.52	0.09	16.29	40.28	<.0001	<.0001	TRUE	TRUE
32.1.Martin:48a-48b	-0.93	0.67	1.6	1.6	0.12	13.14	33.95	<.0001	<.0001	TRUE	TRUE
32.3.Fanselow:59a-59b	-0.49	1.12	1.61	1.61	0.08	20.91	48.82	<.0001	<.0001	TRUE	TRUE
35.3.Hazout:30a-30a	-0.67	0.98	1.64	1.64	0.15	10.76	27.84	<.0001	<.0001	TRUE	TRUE
37.2.deVries:70a-70b	-0.68	0.97	1.65	1.65	0.07	22.04	50.1	<.0001	<.0001	TRUE	TRUE
35.2.Larson:61a-61b	-0.81	0.85	1.66	1.66	0.09	17.71	43.62	<.0001	<.0001	TRUE	TRUE
38.4.Boskovic:74-75	-0.8	0.87	1.67	1.67	0.08	20.91	44.93	<.0001	<.0001	TRUE	TRUE
35.3.Hazout:65a-65b	-0.99	0.73	1.72	1.72	0.12	13.94	35	<.0001	<.0001	TRUE	TRUE
33.2.Bowers:13b-13b	-0.81	0.98	1.79	1.79	0.11	16.49	40.15	<.0001	<.0001	TRUE	TRUE
35.1.Bhatt:13a-13a	-1.03	0.79	1.82	1.82	0.07	27.59	61.45	<.0001	<.0001	TRUE	TRUE
34.1.Basilico:4b-4c	-0.81	1.03	1.84	1.84	0.06	28.8	64.27	<.0001	<.0001	TRUE	TRUE
36.4.denDikken:38b-38b	-1.02	0.89	1.91	1.91	0.08	24.38	58.24	<.0001	<.0001	TRUE	TRUE
37.2.Sigurdsson:3c-3e	-0.92	1.08	2	2	0.07	29.93	58.39	<.0001	<.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.

exp	gramm	beta	z	se	p	pred	sig
35.3.Hazout.36-36	0.39	-0.79	-4.03	0.2	<.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	<.0001	TRUE	TRUE
39.1.Sobin.8c-8f	0.7	1.02	4.6	0.22	<.0001	TRUE	TRUE
34.3.Landau.fn12i-fn12ii	0.71	1.03	5.47	0.19	<.0001	TRUE	TRUE
34.1.Basilico.37a-37b	0.72	1.04	4.59	0.23	<.0001	TRUE	TRUE
33.1.Fox.47c-48b	0.71	1.05	4.06	0.26	<.0001	TRUE	TRUE
35.2.Larson.44b-44b	0.69	1.11	3.85	0.29	<.0001	TRUE	TRUE

34.3.Landau.7c-7c	0.71	1.18	6.07	0.19	<.0001	TRUE	TRUE
34.1.Phillips.61a-61b	0.77	1.3	10.07	0.13	<.0001	TRUE	TRUE
34.2.Panagiotidis.12a-b	0.75	1.45	4.1	0.35	<.0001	TRUE	TRUE
34.4.Boskovic.fn6iie-fn6iid	0.73	1.54	4.56	0.34	<.0001	TRUE	TRUE
32.3.Fanselow.61a-61b	0.78	1.55	8.99	0.17	<.0001	TRUE	TRUE
32.3.Culicover.37a-37a	0.79	1.57	9.36	0.17	<.0001	TRUE	TRUE
41.3.Constantini.1b-1b.BothVsBothBoth	0.79	1.58	7.19	0.22	<.0001	TRUE	TRUE
41.3.Landau.11a-11a	0.8	1.64	7.28	0.22	<.0001	TRUE	TRUE
32.3.Culicover.25c-25d.WithOneself	0.8	1.77	5.75	0.31	<.0001	TRUE	TRUE
41.4.Bruening.61b-62b.StarredVariantIn61	0.74	1.84	3.77	0.49	<.0001	TRUE	TRUE
34.2.Caponigro.11b-11c	0.83	2.01	8.24	0.24	<.0001	TRUE	TRUE
35.3.Embick.7a-7b	0.83	2.03	6.12	0.33	<.0001	TRUE	TRUE
34.1.Phillips.23a-24a	0.83	2.06	5.95	0.35	<.0001	TRUE	TRUE
35.1.Bhatt..93a-b	0.85	2.14	9.59	0.22	<.0001	TRUE	TRUE
39.1.Sobin.20a-21a	0.83	2.18	6.41	0.34	<.0001	TRUE	TRUE
40.4.Hicks.2a-2b	0.86	2.24	6.22	0.36	<.0001	TRUE	TRUE
33.1.denDikken.57a-57b	0.87	2.24	8.85	0.25	<.0001	TRUE	TRUE
33.2.Bowers.49c-49c	0.85	2.26	8.48	0.27	<.0001	TRUE	TRUE
35.1.Bhatt.fn25ia-fn25ib	0.89	2.31	8.81	0.26	<.0001	TRUE	TRUE
34.3.Landau.38a-38c	0.88	2.41	7.65	0.32	<.0001	TRUE	TRUE
35.1.Bhatt.fn5ia-fn5ia	0.88	2.49	6.74	0.37	<.0001	TRUE	TRUE
32.3.Culicover.46b-46b	0.89	2.51	7.03	0.36	<.0001	TRUE	TRUE
34.1.Phillips.59c-60c	0.86	2.59	7.84	0.33	<.0001	TRUE	TRUE
41.4.Bruening.62a-87a.StarredVariantIn87	0.83	2.6	4.57	0.57	<.0001	TRUE	TRUE

34.3.Takano.2b-d	0.86	2.65	5.06	0.52	<.0001	TRUE	TRUE
32.1.Martin.48a-48b	0.89	2.75	7.9	0.35	<.0001	TRUE	TRUE
32.1.Martin.50a-51a	0.91	2.79	7.2	0.39	<.0001	TRUE	TRUE
34.3.Takano.2a-c	0.9	2.8	7.52	0.37	<.0001	TRUE	TRUE
33.2.Bowers.20a-20b	0.88	2.82	8.01	0.35	<.0001	TRUE	TRUE
35.2.Hazout.1b-1b	0.86	2.91	3811.11	0	<.0001	TRUE	TRUE
35.3.Embick.62b-62b.Cf	0.87	2.93	7.25	0.4	<.0001	TRUE	TRUE
33.2.Bowers.7d-7d	0.9	2.95	7	0.42	<.0001	TRUE	TRUE
35.3.Hazout.73b-73b	0.9	2.98	8.21	0.36	<.0001	TRUE	TRUE
38.3.Landau.62a-62b	0.92	3	7.66	0.39	<.0001	TRUE	TRUE
33.2.Bowers.56c-56d	0.89	3.04	5.6	0.54	<.0001	TRUE	TRUE
38.2.Hornstein.fn2.iii-iii	0.95	3.06	13.01	0.23	<.0001	TRUE	TRUE
32.3.Culicover.34c-34e	0.91	3.32	6.97	0.48	<.0001	TRUE	TRUE
35.3.Hazout.65a-65b	0.97	3.38	11.94	0.28	<.0001	TRUE	TRUE
32.3.Fanselow.59a-59b	0.97	3.38	12.16	0.28	<.0001	TRUE	TRUE
35.2.Hazout.1a-1a	0.9	3.41	5.78	0.59	<.0001	TRUE	TRUE
32.1.Martin.50b-51b	0.91	3.42	6.34	0.54	<.0001	TRUE	TRUE
33.2.Bowers.7a-7a.PerfectlyIn2ndPos3rdPos	0.97	3.54	14.39	0.25	<.0001	TRUE	TRUE
32.3.Culicover.46a-48a	0.9	3.57	5.13	0.7	<.0001	TRUE	TRUE
32.3.Culicover.23c-23d.SentenceDP	0.94	3.6	5.91	0.61	<.0001	TRUE	TRUE
34.1.Fox.4-4.	0.91	3.63	5.55	0.65	<.0001	TRUE	TRUE
34.3.Takano.11a-11b	0.92	3.65	5.1	0.71	<.0001	TRUE	TRUE
34.1.Fox.1-1.	0.93	3.77	5.43	0.69	<.0001	TRUE	TRUE
35.3.Hazout.30a-30a	0.97	3.79	9.45	0.4	<.0001	TRUE	TRUE
37.4.Nakajima.fn1ia-fn1iiia	0.93	3.84	5.63	0.68	<.0001	TRUE	TRUE
34.1.Basilico.29b-30b	0.94	3.93	5.71	0.69	<.0001	TRUE	TRUE

34.1.Basilico.4b-4c	0.98	3.99	13.1	0.3	<.0001	TRUE	TRUE
35.2.Hazout.5a-5c	0.93	4.03	4.92	0.82	<.0001	TRUE	TRUE
41.3.Landau.32a-32b	0.92	4.17	3.88	1.07	<.0001	TRUE	TRUE
33.2.Bowers.13b-13b	0.99	4.44	11.69	0.38	<.0001	TRUE	TRUE
37.2.Sigurdsson.3c-3e	0.99	4.44	11.69	0.38	<.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	<.0001	TRUE	TRUE
34.1.Fox.37a-37b	0.92	6.06	5.86	1.04	<.0001	TRUE	TRUE
40.2.Johnson.78-79	0.94	6.7	5.87	1.14	<.0001	TRUE	TRUE
35.2.Larson.61a-61b	0.95	6.76	4.62	1.46	<.0001	TRUE	TRUE
32.2.Alexiadou.fn11iib-fn11iic	0.95	7.45	7.53	0.99	<.0001	TRUE	TRUE
34.4.Haegeman.2c-2b	0.93	7.5	7.21	1.04	<.0001	TRUE	TRUE
34.1.Basilico.7a-7b	0.97	7.59	6.8	1.12	<.0001	TRUE	TRUE
32.1.Martin.15a-15b	0.97	7.88	7.75	1.02	<.0001	TRUE	TRUE
38.4.Boskovic.74-75	0.97	7.95	1318.99	0.01	<.0001	TRUE	TRUE
32.3.Culicover.fn6ia-fn6ib	0.96	7.96	7.7	1.03	<.0001	TRUE	TRUE
33.2.Bowers.19a-19b	0.95	8.04	6.54	1.23	<.0001	TRUE	TRUE
34.2.Caponigro.fn6ia-fn6ib.EagerlyIn2ndPos	0.96	8.17	7.6	1.08	<.0001	TRUE	TRUE
37.2.deVries.70a-70b	0.97	8.26	7.25	1.14	<.0001	TRUE	TRUE
41.1.Muller.28a-28b	0.96	8.4	8.05	1.04	<.0001	TRUE	TRUE
38.2.Hornstein.2b-2c	0.97	8.49	7.31	1.16	<.0001	TRUE	TRUE
41.3.Vicente.6b-8b	0.97	8.49	7.31	1.16	<.0001	TRUE	TRUE
35.1.Bhatt.1b-1b	0.98	8.51	1004.54	0.01	<.0001	TRUE	TRUE
34.3.Heycock.93a-93b	0.96	8.61	6.13	1.4	<.0001	TRUE	TRUE
35.1.Bhatt.13a-13a	0.98	8.69	1834.83	0	<.0001	TRUE	TRUE
36.4.denDikken.38b-38b	0.96	8.76	8.92	0.98	<.0001	TRUE	TRUE

35.1.McGinnis.63a-63b	0.93	8.78	6.95	1.26	<.0001	TRUE	TRUE
34.1.Basilico.50-51	0.96	8.82	7.08	1.25	<.0001	TRUE	TRUE
32.4.Lopez.16a-16b	0.98	9.1	6.59	1.38	<.0001	TRUE	TRUE
35.3.Embick.72a-72b	0.98	9.1	7.07	1.29	<.0001	TRUE	TRUE
32.3.Culicover.44a-45a	0.96	9.33	5.31	1.76	<.0001	TRUE	TRUE
35.1.Bhatt.5a-5c	0.96	9.74	7.12	1.37	<.0001	TRUE	TRUE
33.1.denDikken.56a-58a	0.95	10.35	6.8	1.52	<.0001	TRUE	TRUE
36.4.denDikken.35a-35b	0.96	10.61	6.56	1.62	<.0001	TRUE	TRUE

Appendix C

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

References for *Linguistic Inquiry* papers

Alexiadou, A., & Anagnostopoulou, E. (2001). The subject-in-situ generalization and the role of case in driving computations. *Linguistic Inquiry*, *32*(2), 193–231.

Basilico, D. (2003). The topic of small clauses. *Linguistic Inquiry*, *34*(1), 1–35.

Beck, S., & Johnson, K. (2004). Double objects again. *Linguistic Inquiry*, *35*(1), 97–123.

Becker, M. (2006). There began to be a learnability puzzle. *Linguistic Inquiry*, *37*(3), 441–456.

Bhatt, R., & Pancheva, R. (2004). Late merger of degree clauses. *Linguistic Inquiry*, *35*(1), 1–45.

Boeckx, C., & Stjepanović, S. (2001). Head-ing toward PF. *Linguistic Inquiry*, *32*(2), 345–355.

Bošković, Ž. (2002). On multiple wh-fronting. *Linguistic Inquiry*, *33*(3), 351–383.

Bošković, Ž. (2007). On the locality and motivation of Move and Agree: An even more minimal theory. *Linguistic Inquiry*, *38*(4), 589–644.

Bošković, Ž., & Lasnik, H. (2003). On the distribution of null complementizers. *Linguistic Inquiry*, *34*(4), 527–546.

Bowers, J. (2002). Transitivity. *Linguistic Inquiry*, *33*(2), 183–224.

Bruening, B. (2010a). Ditransitive asymmetries and a theory of idiom formation. *Linguistic Inquiry*, *41*(4), 519–562.

Bruening, B. (2010b). Double object constructions disguised as prepositional datives. *Linguistic Inquiry*, *41*(2), 287–305.

Caponigro, I., & Pearl, L. (2009). The nominal nature of where, when, and how: Evidence from free relatives. *Linguistic Inquiry*, *40*(1), 155–164.

Caponigro, I., & Schütze, C. T. (2003). Parameterizing passive participle movement. *Linguistic Inquiry*, *34*(2), 293–307.

Costantini, F. (2010). On Infinitives and Floating Quantification. *Linguistic Inquiry*, *41*(3), 487–496.

Culicover, P. W., & Jackendoff, R. (2001). Control is not movement. *Linguistic Inquiry*, *32*(3), 493–512.

De Vries, M. (2006). The syntax of appositive relativization: On specifying coordination, false free relatives, and promotion. *Linguistic Inquiry*, *37*(2), 229–270.

Den Dikken, M. (2005). Comparative correlatives comparatively. *Linguistic Inquiry*, *36*(4), 497–532.

Den Dikken, M., & Giannakidou, A. (2002). From hell to polarity:"Aggressively non-D-linked" wh-phrases as polarity items. *Linguistic Inquiry*, *33*(1), 31–61.

Embick, D. (2004). On the structure of resultative participles in English. *Linguistic Inquiry*, *35*(3), 355–392.

Fanselow, G. (2001). Features, $þeta$-roles, and free constituent order. *Linguistic Inquiry*, *32*(3), 405–437.

Fox, D. (2002). Antecedent-contained deletion and the copy theory of movement. *Linguistic Inquiry*, *33*(1), 63–96.

Fox, D., & Lasnik, H. (2003). Successive-cyclic movement and island repair: The difference between sluicing and VP-ellipsis. *Linguistic Inquiry*, *34*(1), 143–154.

Haddican, B. (2007). The structural deficiency of verbal pro-forms. *Linguistic Inquiry*, *38*(3), 539–547.

Haegeman, L. (2003). Notes on long adverbial fronting in English and the left periphery. *Linguistic Inquiry*, *34*(4), 640–649.

Haegeman, L. (2010). The movement derivation of conditional clauses. *Linguistic Inquiry*, *41*(4), 595–621.

Hazout, I. (2004a). Long-distance agreement and the syntax of for-to infinitives. *Linguistic Inquiry*, *35*(2), 338–343.

Hazout, I. (2004b). The syntax of existential constructions. *Linguistic Inquiry*, *35*(3), 393–430.

Heck, F. (2009). On certain properties of pied-piping. *Linguistic Inquiry*, *40*(1), 75–111.

Heycock, C., & Zamparelli, R. (2003). Coordinated bare definites. *Linguistic Inquiry*, *34*(3), 443–469.

Hicks, G. (2009). Tough-constructions and their derivation. *Linguistic Inquiry*, *40*(4), 535–566.

Hirose, T. (2007). Nominal paths and the head parameter. *Linguistic Inquiry*, *38*(3), 548–553.

Hornstein, N. (2007). A very short note on existential constructions. *Linguistic Inquiry*, *38*(2), 410–411.

Johnson, K. (2009). Gapping is not (VP-) ellipsis. *Linguistic Inquiry*, *40*(2), 289–328.

Kallulli, D. (2007). Rethinking the passive/anticausative distinction. *Linguistic Inquiry*, *38*(4), 770–780.

Landau, I. (2003). Movement out of control. *Linguistic Inquiry*, *34*(3), 471–498.

Landau, I. (2007). EPP extensions. *Linguistic Inquiry*, *38*(3), 485–483.

Landau, I. (2010). The explicit syntax of implicit arguments. *Linguistic Inquiry*, *41*(3), 357–388.

Larson, R. K., & Marušič, F. (2004). On indefinite pronoun structures with APs: Reply to Kishimoto. *Linguistic Inquiry*, *35*(2), 268–287.

Lasnik, H., & Park, M.-K. (2003). The EPP and the subject condition under sluicing. *Linguistic Inquiry*, *34*(4), 649–660.

López, L. (2001). On the (non) complementarity of $þeta$-theory and checking theory. *Linguistic Inquiry*, *32*(4), 694–716.

Martin, R. (2001). Null case and the distribution of PRO. *Linguistic Inquiry*, *32*(1), 141–166.

McGinnis, M. (2004). Lethal ambiguity. *Linguistic Inquiry*, *35*(1), 47–95.

Müller, G. (2010). On deriving CED effects from the PIC. *Linguistic Inquiry*, *41*(1), 35–82.

Nakajima, H. (2006). Adverbial cognate objects. *Linguistic Inquiry*, *37*(4), 674–684.

Neeleman, A., & Van de Koot, H. (2002). The configurational matrix. *Linguistic Inquiry*, *33*(4), 529–574.

Nunes, J. (2001). Sideward movement. *Linguistic Inquiry*, *32*(2), 303–344.

Panagiotidis, P. (2003). One, empty nouns, and $þeta$-assignment. *Linguistic Inquiry*, *34*(2), 281–292.

Phillips, C. (2003). Linear order and constituency. *Linguistic Inquiry*, *34*(1), 37–90.

Rezac, M. (2010). ϕ-Agree Versus ϕ-Feature Movement: Evidence From Floating Quantifiers. *Linguistic Inquiry*, *41*(3), 496–508.

Richards, N. (2004). Against bans on lowering. *Linguistic Inquiry*, *35*(3), 453–463.

Sigurḥsson, H. Á. (2006). The nominative puzzle and the low nominative hypothesis. *Linguistic Inquiry*, *37*(2), 289–308.

Sobin, N. (2004). Expletive constructions are not "lower right corner" movement constructions. *Linguistic Inquiry*, *35*(3), 503–508.

Sobin, N. (2008). Do so and VP. *Linguistic Inquiry*, *39*(1), 147–160.

Stepanov, A., & Stateva, P. (2009). When QR disobeys superiority. *Linguistic Inquiry*, *40*(1), 176–185.

Stroik, T. (2001). On the light verb hypothesis. *Linguistic Inquiry*, *32*(2), 362–369.

Takano, Y. (2003). How antisymmetric is syntax? *Linguistic Inquiry*, *34*(3), 516–526.

Vicente, L. (2010). A note on the movement analysis of gapping. *Linguistic Inquiry*, *41*(3), 509–517.

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 (β=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.

(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 non-significant trend in favor of (a) in the rating study and a non-significant trend towards (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.[8]

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

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 (A. Gelman, Carlin, Stern, & Rubin, 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 is *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 .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:

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

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 5 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 5 *a priori* A's and 5 *a priori* B's to our 5 experimentally obtained A's such that we imagine we have 10 A's and 5 B's, 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 A's + 5 A's) / (15 trials) = 66. If we were very confident that the sentences are equally acceptable (i.e., the coin is fair; $p \sim .5$), we could use a Beta(100, 100) prior. With a prior like that, we'd have to conduct many more trials in order to move our estimate substantially away from .50. After getting 5 A's, we would still have an estimate of 51%.

If we thought it was very likely that one of the sentences was better, but we didn't know which, we might instead use a Beta prior of Beta(.1,.1). This would mean that, after asking 5 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 2 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.

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

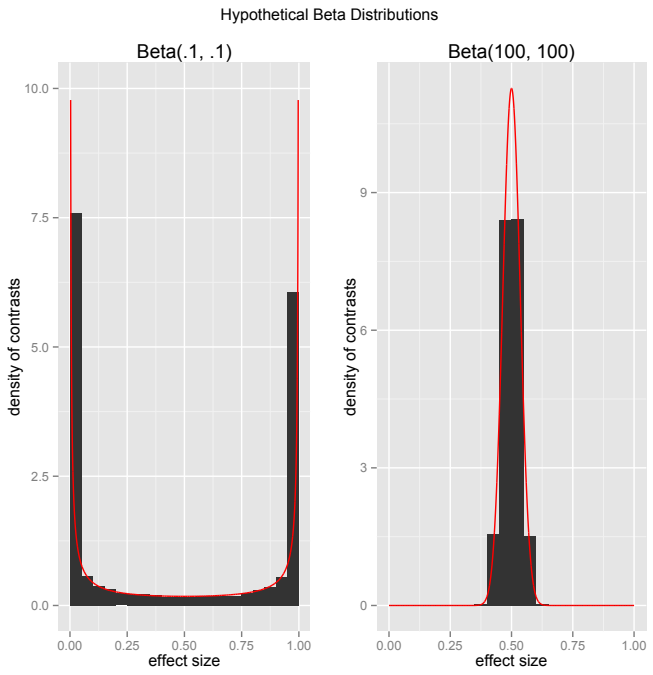

Figure 3: The histograms represent a density map of a draw from a beta distribution with the shape parameters indicated. The red line is the probability density of the beta distribution at each value for *p* between 0 and 1. The plot on the left conforms to an instance in which, most of the time, the probability *p* is extreme (towards 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 .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.[9] In other words, when we have a new contrast for which we don't have much data but which we believe likely to

[9] The prior that is obtained by our experimental results ends up very similar to what is obtained from the results from SSA's data (available on Jon Sprouse's webpage).

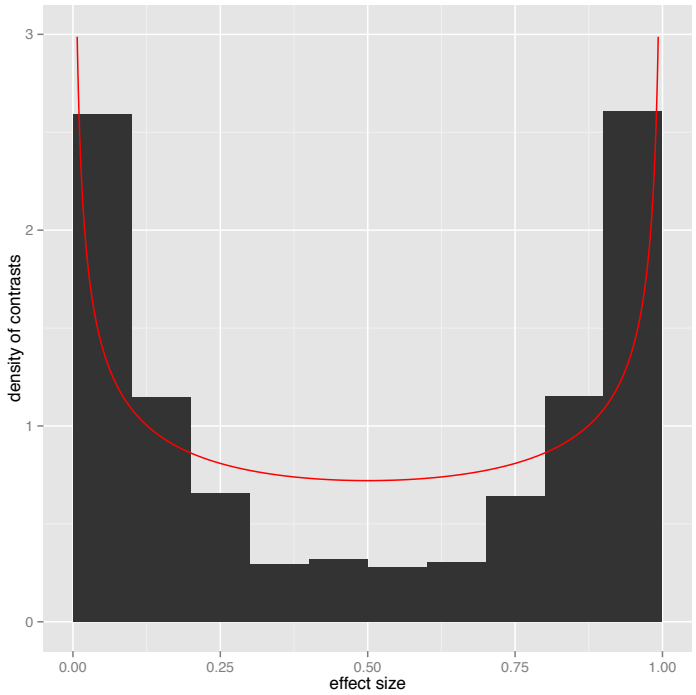

Figure 4: This plot corresponds to a smoothed histogram (averaged over many trials) of the data from our forced choice experiment where, for each contrast, one variant is randomly assigned to be Sentence A and one to be Sentence B. Most of the time, there is a strong preference for one sentence or the other. The best fit for the beta distribution is Beta(.6, .6)—which is shown by the red line.

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 many of the effects that we tested were hypothesized to be very strong.

In order to determine the prior empirically, for each contrast in our experiment, we randomly assign one sentence to be Sentence A and one to be Sentence B. We then draw a histogram of the effect sizes and fit the beta distribution to the histogram (as seen in Figure 4). Averaging over 100 samples, the best fit is Beta(.6, .6). 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. Because we want to remain unbiased and conservative, we chose not to use the information that we know which sentence the researcher has hypothesized as Sentence A and which as Sentence B. That 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 which are consistent with pure noise is actually a result in favor of the hypothesis. So, while it is often advisable to use prior information to design experiments and build models, here we want to be conservative so as not to bias the analysis towards the desired outcome.[10] Even so, if we took advantage of our knowledge that one sentence is preferred, the result would not be dramatically different. In particular, we obtain a best fit beta prior of Beta(5.9, 1.1). For 7 unanimous participants, this gives us a mean of .92 with a 95% CI of [.79, 1]. So the CI's lower bound is only slightly higher than if the prior did not include information about which sentence is preferred. To get the lower bound to .75 when we use this prior, we would need to include 5 participants in the experiment (as compared to 7, with the more conservative prior). 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 locally uniform. Hence, the outcome is similar under other plausible alternative priors. We prefer the more conservative [Beta(.6, .6)] but the recommendations made here are robust and not overly sensitive to the precise choice of prior.

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

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 .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 effect 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 effect model to the real data as described in the main text.
b) Use the random effect 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 effect 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 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 below tables 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.

Table F.1 Ratings study (all values where significance is defined by $p < .05$)

***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 F.2 Forced choice study (all values where significance is defined by $p < .05$)

***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

*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 $p < .05$, 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).