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

Kyle Mahowald, Peter Graff, Jeremy Hartman, Edward Gibson

Language, Volume 92, Number 3, September 2016, pp. 619-635 (Article)

Published by Linguistic Society of America
DOI: 10.1353/lan.2016.0052

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

Kyle Mahowald
Massachusetts Institute of Technology

Peter Graff
Intel Corporation

Jeremy Hartman
University of Massachusetts Amherst

Edward Gibson
Massachusetts Institute of Technology

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. In this article, we first present the results of a large-scale replication of the Sprouse et al. 2013 study 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 nonexistent 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 meaningful quantitative 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 at least five unique participants, using at least five unique sentences of each type. If all participants in the sample 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 reliable performance.*

Keywords: syntax, semantics, linguistic acceptability, quantitative linguistics

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.

Since the early days of generative grammar, researchers have been asking questions about methods in linguistics research, specifically, the relationship between grammaticality and acceptability (Chomsky 1986, Labov 1978, Levelt et al. 1977, McCawley

* We would like to thank Evelina Fedorenko, Roger Levy, Steve Piantadosi, Colin Phillips, and the audience at AMLaP 2014 for comments and discussion. We would further like to thank Ruth Abrams, Zuzanna Balewski, Michael Behr, Lauren Burke, Camille DeJarnett, Jessica Hamrick, Anthony Lu, Emily Lydic, Molly McShane, Philip Smith, Jacob Steinhardt, Stephanie Tong, Chelsea Voss, Tyrel Waagen, Xiaoran Xu, and Fangheng Zhou for constructing experimental items as well as Zoe Snape for help with the implementation of the experiment. We thank Richard Futrell, Alexander Paunov, and Caitlyn Hoeflin for coding the obviousness of judgments. KM was supported by an NDSEG graduate fellowship.

1982), the reliability of intuition as a method (Labov 1978, Levelt et al. 1977), and the practice of linguists using their own intuitions as opposed to consulting naive speakers (e.g. Birdsong 1989, Householder 1965, Spencer 1973). For a more detailed history of these issues, see Schütze 1996. In recent years, increases in the availability of data have led to further discussions of the weaknesses of the informal method (e.g. Arppe & Järvikivi 2007, Cowart 1997, Featherston 2005, Gibson & Fedorenko 2010, 2013, Gibson et al. 2013, Gross & Culbertson 2011, Lau et al. 2017, Linzen & Oseki 2015, Schütze 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, an advantage of formal methods is that they provide fellow researchers with quantitative information about the quality of the data 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, and Almeida (2013; henceforth SS&A) 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 SS&A, 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.

First, we briefly discuss the following arguments in favor of quantitative methods in syntax and semantics research.

(i) **The current error rate in informal linguistic judgments is not as low as it could be.** SS&A 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, SS&A experimentally analyzed 148 randomly sampled English acceptability judgments from *Linguistic Inquiry* (*LI*) 2001–2010. Of these 148 experiments, 127 (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 (all values here are obtained using mixed models, which are most appropriate for this type of data; Barr et al. 2013, Bates et al. 2014). Seven of the 148 experiments (5%) did not show predicted effects in any of the three experiments. SS&A generalize from the 95% rate that informal acceptability intuitions reported in research articles on generative syntax have statistical properties similar to those of quantitative experiments that compare acceptability ratings for various experimental items given 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. SS&A (p. 230) 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 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, and Fedorenko (2013), we believe that the current error rate in published linguistic judgments could be lower without 'crippl[ing] linguistic investigation' (Culicover & Jackendoff 2010:234). Much of the recent debate on quantitative methods in syntax and semantics has focused on whether a significant *p*-value ($p < 0.05$) has been obtained through a quantitative experiment (see Sprouse & Almeida 2012 for more discussion of effect size and statistical power in linguistic acceptability judgments). If the null hypothesis can be rejected, the syntactic judgment is said to 'replicate'. While SS&A 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.

On the one hand, a $p < 0.05$ false-positive THRESHOLD for NHST in behavioral experiments is not comparable to a 5% false-positive RATE in published acceptability judgments. The NHST paradigm assumes that one has performed statistical significance testing for each particular effect under consideration; the $p < 0.05$ 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 and the assumption that the particular sample being studied is drawn from a larger pool. On the other hand, a 5% error rate in linguistic acceptability judgments suggests that 5% of all judgments would diverge from the results of a formal experiment. But there is no sampling being done; the method provides no quantitative information about any individual effect. Therefore the notion of a statistical threshold across a body of judgments is problematic. If the average linguistics paper has thirty-three examples (the average number of US English examples found in the articles examined by SS&A), divergences are uniformly distributed, and if the divergence rate is 5%, then every paper is likely to contain a questionable judgment: 1.64 on average.

(ii) **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 useful for meaningfully interpreting individual acceptability contrasts.

(iii) **Informal linguistic experiments make it difficult for researchers who either are from other fields or do not 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 un-

familiar with the language in question. This concern may be especially relevant in light of recent evidence that reliability on acceptability judgments in non-English languages may be even lower than those in English (Linzen & Oseki 2015). Formal experiments using a consistent methodology make it easy to compare effects across languages.

(iv) **Formal experiments need not be costly or time consuming since even very few participants can sometimes give meaningful results.** Another common concern is that formal experiments in syntax and semantics are too costly in terms of 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 crowdsourcing platforms like Amazon.com's Mechanical Turk, which can provide robust results for cognitive behavioral experiments (Crump et al. 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. the Turkolizer software from Gibson et al. 2011). In a syntactic judgment experiment on Mechanical Turk, the researcher typically posts a survey consisting of a set of sentences (presented either visually or through audio) and asks 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, as well as those who do fieldwork 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. Following Myers's (2009) proposal of using mini-experiments to formalize and quantify the sort of informal syntactic exploration that happens anyway, 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 here 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. For our study, we sampled a new set of sentences from the same 2001–2010 *LI* issues that SS&A evaluated, and tested them in a rating experiment and a forced-choice experiment.[1]

2.1. Ratings experiments.

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

Stimuli. PG and an undergraduate assistant went through all of the *LI* articles from 2001–2010 in which US English contrasts were presented and that were sampled from

[1] SS&A also did a magnitude-estimation study, but since those results were very similar to those of the rating study, we did not include magnitude estimation here.

by SS&A. We selected only contrasts that (i) were English, (ii) directly compared a grammatical and an ungrammatical sentence (irrespective of the particular judgment reported, and assuming that OK > ? > ?? > *? > *), and (iii) were not primarily dependent on interpretation. This resulted in a total of 814 contrasts. Of these, forty-one contrasts had already been tested by SS&A. From the remaining 773 examples, we randomly sampled 101 contrasts.

For ninety-six 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 article. Six templates/sample item pairs were assigned to sixteen 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 ten experimental items (hypothesized grammatical/ungrammatical pairs; twenty 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 twelve experimental items per contrast. For the five other contrasts in our sample, eleven items were constructed by JH, which, together with the original sentence pair reported, also resulted in twelve items for those five contrasts. All 1,212 contrast pairs were divided into four 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 twenty-one of our 1,212 sentences had minor spelling mistakes or errors in punctuation. We excluded these twenty-one 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 participants using the rating scale differently from each other, 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 for 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 and 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 condition), divided by the standard deviation of the data. Using Cohen's recommended effect-size windows, we found 19/100 effects to be small effects ($d < 0.5$), 15/100 to be medium effects ($0.5 < d < 0.8$), and 66/100 to be large effects ($d > 0.8$). Of the nineteen small-effect contrasts, seven actually have a Cohen's *d* of < 0.2, which is the minimum value that Cohen posits for a 'small effect'.

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 the sentence is reported as 'acceptable' or 'unacceptable' in the source *LI* article) and random intercepts for both participant and item and random slopes for gram-

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

maticality grouped by both participant and item (random effects). The estimated coefficient for grammaticality is an 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-effects model. Despite the fact that many of the articles from which these examples were drawn talk about these contrasts categorically as either grammatical or ungrammatical, Fig. 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-effects model described above, fit using the R statistical programming language (R Core Team 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 hypothesized ungrammatical sentences), leaving the random-effects structure intact in both models (Agresti 2002, Baayen et al. 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 given 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 detected 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 $\alpha = 0.05$. Effect sizes at least this big are estimated for 81% of our contrasts. For a 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 ninety-two of 100 contrasts in our random sample show significant effects in the predicted direction (92%). The contrasts that do not show clear effects in either the rating experiment or the forced-choice experiment (discussed below) are reported in Table 1.

SOURCE ARTICLE	RATING	FORCED CHOICE
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 sign 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. An asterisk next to the plus or minus means that the effect was significant at $p < 0.05$.

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

DISCUSSION. Of the 100 contrasts, the majority showed the predicted effect robustly. Of the eight that did not show a significant result in the predicted direction, four showed clear trends in that 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 0.1 and ps < 0.15. 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 in the online supplementary materials.[3] It should not automatically be concluded that the inclusion of these sentences represents failures on the part of the researchers, although we do believe these experiments suggest that these sentences warrant further investigation.

2.2. FORCED-CHOICE EXPERIMENTS.

PARTICIPANTS. A total of 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 chose 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 those in the ratings experiment, except with minor spelling and punctuation corrections.

PROCEDURE. Participants were asked to read each pair of sentences out loud to themselves and choose which sentence 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, six of 100 contrasts trended in the opposite direction from that predicted. The remaining 94/100 sentences showed an effect in the predicted direction. An effect size greater than 0.75 was found for 81/100, and roughly half (52/100) had an effect size greater than 0.9. Overall, these results demonstrated smaller effects than those reported by SS&A, but they were qualitatively similar.

To control for individual variation by participant and item, we fit a logistic linear mixed-effects model predicting whether the participant preferred the hypothesized acceptable sentence over the hypothesized unacceptable one. We included a fixed-effect intercept (that is, whether 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 is essentially an 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 gives 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-effects model.[4] As with the rating study, we see no discrete jumps but rather a continuum of effect sizes.

[3] The online appendices referenced throughout this article can be accessed at http://muse.jhu.edu/article/628201/pdf.

[4] 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 does not matter here, because these effects were very large.

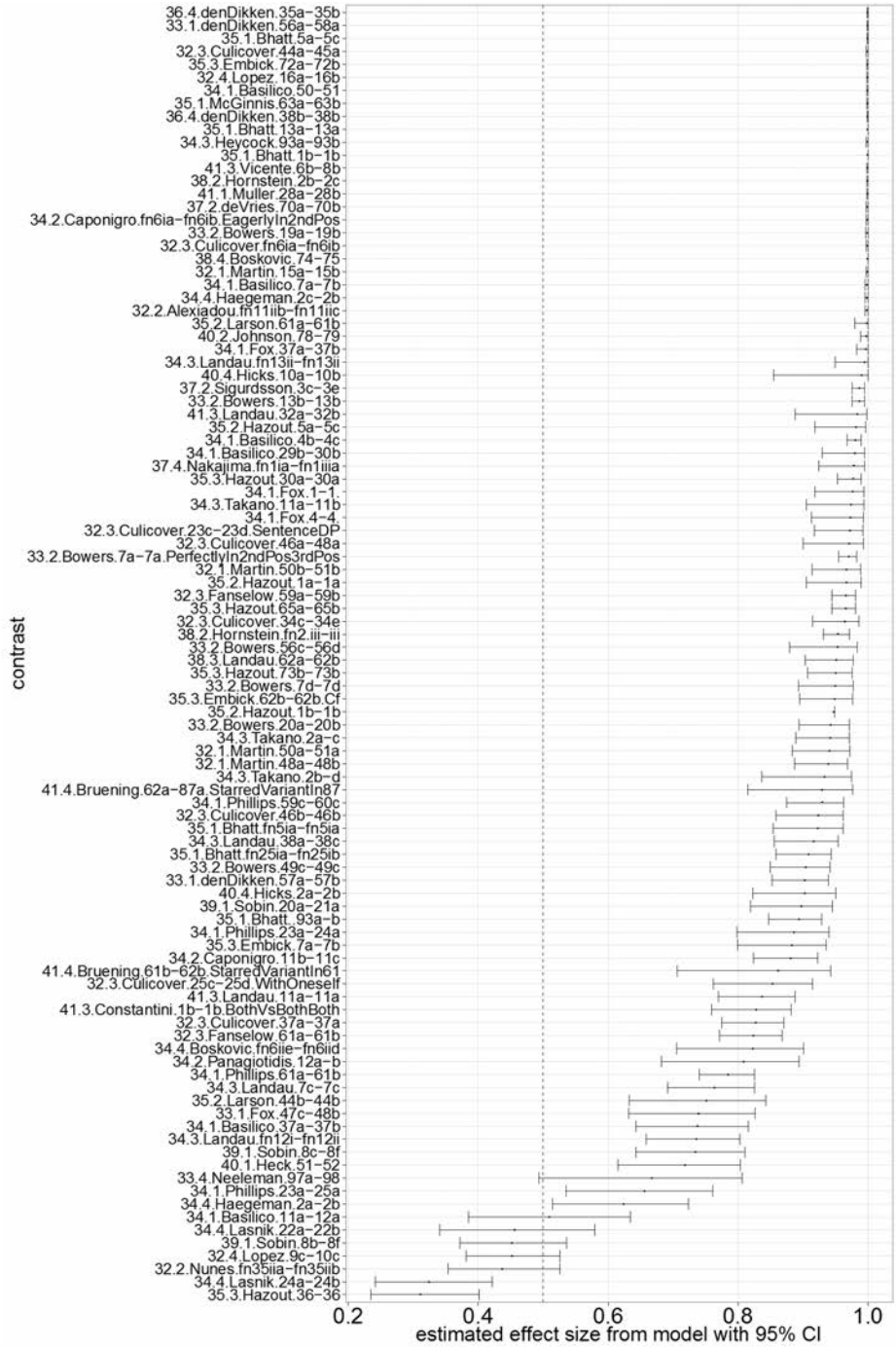

Figure 2. Each point is an effect size for the forced-choice experiment listed on the y-axis with 95% CIs estimated from the linear mixed-effects model. When the error bars extend through the line at .5, the effect is not significant. The error bars that appear entirely to the left of the line at .5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

To assess significance in an NHST framework, we used the logistic linear mixed-effects model described above and used the *z*-value to calculate a *p*-value. As before, we calculated statistical power for several possible true effect sizes. If the underlying effect 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 online Appendix A. Significant trends in favor of the hypothesized acceptable version were shown in 92/100 contrasts. Two items showed nonsignificant trends in favor of the hypothesized acceptable version, and four showed nonsignificant 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, the methods are critically different.

Our results are quite similar to those reported by SS&A. The contrasts that did not show significant effects in the expected direction in our experiments are listed in Table 1 above, and we briefly discuss several of these cases in online Appendix D.

3. SNAP JUDGMENTS. Across our two experimental paradigms, about half of the examples sampled from *LI* showed a strong effect (Cohen's $d > 0.8$). 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 since even strong intuitions are sometimes wrong. To increase 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. Of course, not all questions of acceptability are best investigated through forced-choice judgments, and we leave it to future work to extend this method to other paradigms.

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 about 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* is 0.75, this means that 75% of the time, sentence A would be preferred.

If we ask five people which sentence they prefer and they all prefer sentence A, can we conclude with certainty that everyone will prefer sentence A? No: it is possible that, if we ask a sixth person, she will prefer sentence B. We still do not know if the probability of someone in the larger population preferring sentence A is 90% or 70% or 10% (in which case getting five 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 results 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 0.5, we might still think the effect is near 0.5 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(5.9, 1.1). This means that, for a new contrast where a researcher has used informal methods to decide that sentence A is clearly better than sentence B, this is roughly equivalent to an experiment where six participants have said they prefer A and one prefers B. (For a much more detailed description, see online 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 0.90 and our lower 95% confidence bound at or above 0.75. We chose this threshold somewhat arbitrarily based on our perception of what constitutes clear linguistic evidence. But this method can be used to ask what sample size is needed to achieve some other threshold by using the data in Table 2 below. Depending on the sort of conclusion a researcher plans to draw, different levels of effect size and confidence may be warranted.

If we ask three 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 .89 with a 95% CI of [.70, 1]. If we asked five people and they all answered sentence A, the mean goes to .91 with a 95% CI [.75, 1]. Thus, where $n = 5$ and the results of the experiment are unanimous, we get an expected mean of .91 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] Of course, as the sample size goes up, the experiment will give more information. Consequently, if a larger sample is easily obtainable, we recommend that. We do not recommend using fewer than five participants with five unique items for a SNAP judgment.

3.1. TESTING SNAP JUDGMENTS. We can test the efficacy of the SNAP judgments proposal in the sample that we already have. First, we ask what the false-positive rate is for SNAP judgments: how often a SNAP judgment will give a unanimous result when the underlying experiment is inconclusive. For each experiment, we randomly sampled five 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 five 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, fifty-four of them produced a unanimous 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 experiments was .92 with a 95% CI of [.76, .99].[6] Compare this to our beta prior, which gave us an expected mean of .93 [.75, 1]. The empiri-

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

[6] Note that this sampling technique (where we sample unique subjects and items) is not the same as simply drawing from a binomial with some parameter p for each experiment, due to subject and item effects.

cal test is consistent with the results obtained using the beta prior: the means match exactly, and the lower bound for the 95% CI is also very similar (.75 compared to .76 in the empirical test). Of the 54% of contrasts that pass SNAP judgments, on average only 0.20% of them show results in the opposite direction of what is predicted, and none of them included a significant result in the opposite direction.

Next, we ask about the statistical power of SNAP judgments, specifically for large effects. That is, when an effect is very large, how often does the SNAP judgments paradigm fail to detect it and thus unnecessarily suggest a full experiment? We define a large effect here as one where the mean is greater than .90 (roughly half of the experiments in our sample have this property). We simulate experiments as above, drawing on the real data from the subset of experiments where the mean is above the .90 threshold. On average, 77% of the experiments with true means above .90 produce unanimous effects in our simulations using five data points. Despite the small sample size, this is near the 80% threshold recommended for appropriate statistical power in experiments. Of course, for smaller effects, the statistical power will be much lower.

One may still, at this point, wonder why we recommend at least five data points for a SNAP judgment as opposed to any arbitrary number. We believe that $n = 5$, with a unanimous result, provides a robust generalization sufficient for most linguistic judgments. And crucially, it does not give a significant result in the wrong direction for any of the samples that we tested from *LI*. In Table 2, we give means and 95% and 99% CIs for SNAP judgments that give unanimous results in the predicted direction.

n	MEAN WITH 95% CI		MEAN WITH 99% CI	
1/1	0.87	[0.64, 1]	0.87	[0.5, 1]
2/2	0.88	[0.67, 1]	0.88	[0.55, 1]
3/3	0.89	[0.7, 1]	0.89	[0.59, 1]
4/4	0.9	[0.73, 1]	0.9	[0.62, 1]
5/5	0.91	[0.75, 1]	0.91	[0.65, 1]
6/6	0.92	[0.77, 1]	0.92	[0.67, 1]
7/7	0.92	[0.78, 1]	0.92	[0.69, 1]
8/8	0.93	[0.8, 1]	0.93	[0.71, 1]
9/9	0.93	[0.81, 1]	0.93	[0.73, 1]
10/10	0.94	[0.82, 1]	0.94	[0.74, 1]

TABLE 2. Means with the 95% and 99% CIs where the number of unanimous participants in the experiment varies from one to ten. For instance, in an experiment with three participants, all of whom choose unanimously, one can estimate a mean of .89 with a 95% CI of .70 to 1 and a 99% CI of .59 to 1.

3.2. RECOMMENDATION FOR SNAP JUDGMENTS. Given the proposal and evaluation above, we make the following recommendations for SNAP judgments.

(i) **To ensure the applicability of our empirical estimates, SNAP judgments should only be used when the researcher believes, after informal investigation, that the effect is clear and likely to 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 & Almeida 2012). From a statistical perspective, it is important that the researcher has this belief since the recommendations here are based on data that were published in a journal. Thus, if a researcher only has an inkling that sentence A is better than sentence B and wants to run a test to be sure, SNAP judgments is not appropriate. In that case, we recommend a full, larger-scale experiment.

(ii) **Construct five 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 five contrasts vary in lexical content and whatever other factors may influence the acceptability of the sentences in question. Present each contrast to a unique naive participant and ask for a forced-choice judgment. This could be done using Amazon's Mechanical Turk (paying perhaps five cents for one judgment, such that the whole experiment will cost less than fifty cents) or by simply asking for judgments from students, friends, informants, or colleagues who are naive to the experiment in question.[7]

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 five participants. Table 2 lists the conclusions that can be drawn from even smaller SNAP judgments. For these extreme cases, we note that even the use of three independent data points substantially reduces the risk of a false positive. Thus, these recommendations need not dramatically slow or impede the pace of fieldwork: fieldworkers often do have three independent data points for a construction or contrast in question and can report those judgments quantitatively.

(iii) **If all five naive participants agree with the predicted result, one can conclude the following: the predicted mean for the full experiment is .91, with a 95% CI of [.75, 1] and a 99% CI of [.65, 1].** That is, one can be 95% confident that at least 75% of people would agree with the intuition. See Table 2 for guidelines when using participant sample sizes other than five.

(iv) **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 five items. Is there a pattern among the 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 judgments experiment, one can then test these hypotheses. At that point, we recommend a formal experiment with a larger number of items and participants in order 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.)

3.3. Limitations of SNAP judgments. Although the SNAP judgments paradigm addresses issues with statistical power in running linguistic experiments, there are a number of limitations of this technique that are worth considering and that can be explored in future research. For one, the guidelines presented here do not 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, he 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 and semantics.

[7] 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.

Moreover, the SNAP judgments proposal does not address what factors are involved in distinguishing the acceptability of two structures. In particular, this framework does not guarantee that any differences are due to syntax ('grammaticality') rather than some other factor that might systematically differ between the two structures, such as world knowledge ('plausibility') or lexical or discourse properties.

When SNAP judgments is used with in-person participants as opposed to over the internet, we recommend caution in making sure that the researcher does not bias the participants toward any particular answer. When done online using a crowdsourcing service like Mechanical Turk, it is good practice to include some checks to make sure that participants are engaged, doing the task, and competent users of the language of interest. For more on using Mechanical Turk for language research, see Crump et al. 2013, Gibson et al. 2011, Mason & Suri 2012, and Sprouse 2011.

Another limitation of the paradigm here is that we have so far tested it only for English sentences sampled from *LI*. More work is needed to see how well the recommendations here extend to other languages and other types of sentences that may be of theoretical interest. There is also more work needed in order to know whether *LI* sentences are typical of judgments of theoretical interest or whether they differ in meaningful ways from other peer-reviewed journals and whether journals differ from conference proceedings.

Finally, it is important to note that the judgments we trained on here are published judgments and thus not necessarily representative of the judgments that linguists are faced with in day-to-day research. In particular, because the recommendations here are based on published judgments that the authors presumably believe are correct, the model assumptions are not valid for questions where a researcher is uncertain. It is possible that the distribution observed here would be different using unpublished data and thus that there might be different SNAP judgments recommendations depending on where in the publication pipeline a particular judgment is.

4. General discussion. In this article, we have replicated the empirical findings of SS&A. 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. Of these syntactic judgments reported in *LI*, 89% 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 articles 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 we 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 five participants, we recommend a forced-choice experiment with five participants and five items. If the result is unanimous, the results of this small N experiment can be combined with a beta prior to give a predicted effect size of .93 with a 95% CI [.75, 1].

There is great value in being able to attain cheap and easy quantitative data in syntax and semantics—and in knowing when to run larger experiments. Marantz (2005) (and more recently Linzen and Oseki (2015)) proposes three distinct classes of acceptability judgment: judgments that contrast word salad with obviously grammatical language (*Ate rat cat the* vs. *The cat ate the rat*), judgments that test 'obvious' features of a grammar such as adding *-ed* to form a past tense in English, and judgments that explore more subtle features of a language like reference or long-distance dependency. Broadly, this three-way categorization classifies judgments into 'obvious' (Marantz's first two classes) and 'nonobvious' judgments (the third class). Marantz suggests that this third class of more subtle judgments (what we call 'nonobvious') are the ones that could benefit from more formal experimentation. Linzen and Oseki (2015) show that these intuitively nonobvious judgments are much less likely to be replicated in a formal experiment, with rates of failures to replicate of 50% in Hebrew and 32% in Japanese for these nonobvious judgments (where obviousness was coded by the experimenters).

But how does a researcher know a priori if she is dealing with a nonobvious judgment? Five independent coders (KM, EG, and three others) annotated the 100 *LI* contrasts from our sample and classified them as obvious or nonobvious and found that the average agreement between two coders was only 68%. Thus, it is not always obvious when a contrast is obvious. The authors discussed the disagreements among the five raters and arrived at a consensus set of obvious and nonobvious judgments, without consulting the experimental data for these contrasts. Here, the contrasts that were rated as 'obvious' had on average a mean of 92% in the forced-choice experiments, compared to 79% for the nonobvious judgments. Significant results in the predicted direction were shown for 98% of 'obvious' judgments in the forced-choice full experiment, compared to just 86% for the 'nonobvious' judgments. Because judgments with over 90% agreement in a full experiment were shown to be likely to pass SNAP judgments, the SNAP judgments paradigm gives researchers an effective way to know if they are dealing with a judgment that is in need of formal experimentation: judgments that do not produce unanimous SNAP results are likely not 'obvious' contrasts. To that end, one way of thinking of SNAP judgments is as an empirical procedure for determining whether a syntactic contrast falls into the nonobvious class of judgments that warrant further formal experimentation.

SNAP judgments will lead to more published quantitative data in two ways: (i) through the publication of small-sample SNAP judgments data and (ii) through the increased use of formal acceptability experiments for contrasts that do not meet the SNAP threshold. 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 on acceptability judgments. This body of data will make it easy for researchers to uncover new and interesting empirical phenomena and to 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 from the Open Science Foundation at http://osf.io/5wm2a), one can easily test a new contrast and plot it alongside the 100 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. 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 researcher 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 shared across disciplines, across languages, and by future generations of researchers.

REFERENCES

Agresti, Alan. 2002. *Categorical data analysis*. New York: John Wiley & Sons.

Arppe, Antti, and Juhani Järvikivi. 2007. Take empiricism seriously! In support of methodological diversity in linguistics. *Corpus Linguistics and Linguistic Theory* 3.99–109. DOI: 10.1515/CLLT.2006.007.

Baayen, R. Harald; Doug J. Davidson; and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59.390–412. DOI: 10.1016/j.jml.2007.12.005.

Barr, Dale J.; Roger Levy; Christoph Scheepers; and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68.255–78. DOI: 10.1016/j.jml.2012.11.001.

Bates, Douglas; Martin Maechler; Ben Bolker; and Steven Walker. 2014. lme4: Linear mixed-effects models using Eigen and S4. Online: http://CRAN.R-project.org/package=lme4.

Birdsong, David. 1989. *Metalinguistic performance and interlinguistic competence*. Dordrecht: Springer.

Chomsky, Noam. 1957. *Syntactic structures*. Berlin: Walter de Gruyter.

Chomsky, Noam. 1986. *Barriers*. Cambridge, MA: MIT Press.

Cohen, J. 1994. The earth is round ($p < .05$). *American Psychologist* 49.997–1003. DOI: 10.1037/0003-066X.49.12.997.

Cowart, Wayne. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

Crump, Matthew J. C.; John V. McDonnell; and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE* 8:e57410. DOI: 10.1371/journal.pone.0057410.

Culicover, Peter W., and Ray Jackendoff. 2010. Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences* 14.234–35. DOI: 10.1016/j.tics.2010.03.012.

Featherston, Sam. 2005. Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua* 115.1525–50. DOI: 10.1016/j.lingua.2004.07.003.

Gibson, Edward, and Evelina Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14.233–34. DOI: 10.1016/j.tics.2010.03.005.

Gibson, Edward, and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28.88–124. DOI: 10.1080/01690965.2010.515080.

Gibson, Edward; Steven T. Piantadosi; and Evelina Fedorenko. 2013. Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes* 28.229–40. DOI: 10.1080/01690965.2012.704385.

Gibson, Edward; Steve Piantadosi; and Kristina Fedorenko. 2011. Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass* 5.509–24. DOI: 10.1111/j.1749-818X.2011.00295.x.

Gross, Steven, and Jennifer Culbertson. 2011. Revisited linguistic intuitions. *The British Journal for the Philosophy of Science* 62.639–56. DOI: 10.1093/bjps/axr009.

Householder, Fred W., Jr. 1965. On some recent claims in phonological theory. *Journal of Linguistics* 1.13–34. DOI: 10.1017/S0022226700000992.

Labov, William. 1978. Sociolinguistics. *A survey of linguistic science*, ed. by William Orr Dingwall, 339–72. Stamford, CT: Greylock.

Lau, Jey; Alexander Clark; and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, to appear.

LEVELT, WILLEM J. M.; J. A. W. M. VAN GENT; A. F. J. HAANS; and A. J. A. MEIJERS. 1977. Grammaticality, paraphrase and imagery. *Acceptability in language*, ed. by Sidney Greenbaum, 87–101. The Hague: Mouton.
LINZEN, TAL, and YOHEI OSEKI. 2015. The reliability of acceptability judgments across languages. New York: New York University, MS.
MARANTZ, ALEC. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review* 22.429–45. DOI: 10.1515/tlir.2005.22.2-4.429.
MASON, WINTER, and SIDDHARTH SURI. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44.1–23. DOI: 10.3758/s13428-011-0124-6.
MCCAWLEY, JAMES D. 1982. How far can you trust a linguist? *Language, mind, and brain*, ed. by Thomas W. Simon and Robert J. Scholes, 75–87. Hillsdale, NJ: Ablex.
MYERS, JAMES. 2009. The design and analysis of small-scale syntactic judgment experiments. *Lingua* 119.425–44. DOI: 10.1016/j.lingua.2008.09.003.
R CORE TEAM. 2012. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Online: http://www.R-project.org.
SCHÜTZE, CARSON T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.
SORACE, ANTONELLA, and FRANK KELLER. 2005. Gradience in linguistic data. *Lingua* 115.1497–1524. DOI: 10.1016/j.lingua.2004.07.002.
SPENCER, N. J. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* 2.83–98. DOI: 10.1007/BF01067203.
SPROUSE, JON. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43.155–67. DOI: 10.3758/s13428-010-0039-7.
SPROUSE, JON, and DIOGO ALMEIDA. 2012. Power in acceptability judgment experiments and the reliability of data in syntax. Irvine: University of California, Irvine, and Ann Arbor: Michigan State University, MS. Online: http://ling.auf.net/lingbuzz/001520.
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.
WASOW, THOMAS, and JENNIFER ARNOLD. 2005. Intuitions in linguistic argumentation. *Lingua* 115.1481–96. DOI: 10.1016/j.lingua.2004.07.001.

Mahowald
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
77 Massachusetts Ave., 46-3037
Cambridge, MA 02139
[kylemaho@mit.edu]

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

Hartman
Department of Linguistics
University of Massachusetts at Amherst
650 N. Pleasant St., Integrative Learning Center 428
Amherst, MA 01003
[hartman@linguist.umass.edu]

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

[Received 17 June 2014;
revision invited 1 January 2015;
revision received 28 January 2015;
revision invited 18 September 2015;
revision received 16 October 2015;
accepted 11 November 2015]