

Accepted for publication at Behavioral and Brain Sciences. Comment on target article: “An experimental approach to linguistic representation” by Holly Branigan and Martin Pickering.

Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable

Jon Sprouse^a and Diogo Almeida^b

^a*Department of Linguistics, University of Connecticut, Storrs*

^b*Department of Psychology, New York University Abu Dhabi, Abu Dhabi*

***Abstract:** Branigan and Pickering advocate the use of syntactic priming to investigate linguistic representations and argue it overcomes several purported deficiencies of acceptability judgments. While we recognize the merit of drawing attention to a potentially underexplored experimental methodology in language science, we do not believe the empirical evidence supports Branigan and Pickering’s claims about acceptability judgments. We present the relevant evidence.*

Branigan and Pickering (B&P) advocate the use of syntactic priming to investigate linguistic representations. We support the use of any data types that scientists find relevant for specific research questions, including syntactic priming. We regret, then, that B&P appear to repeat unsubstantiated claims that paint a relatively misleading picture of acceptability judgments (AJs), a data type that linguists have been using fruitfully for decades. From our perspective, much of the literature criticizing AJs has repeatedly focused on *logically possible* concerns about their use without investigating whether those concerns are *empirically attested*. This risks a vicious circle: Articles can cite each other for support, giving the illusion of empirical support. In this commentary, we highlight a number of studies that have pursued this issue head on, which we leverage to examine six of B&P’s claims about AJs in detail.

Claim 1: Linguists standardly ask a single informant about the acceptability of a few sentences (p. 12).

Claim 1 is a caricature of linguistic methodology that, to our knowledge, has never been supported by evidence. Nonetheless, a charitable interpretation of this claim reveals two separate concerns: (1) the routine use of small sample sizes, and (2) the susceptibility of AJs to investigator bias (Claim 2, below). An obvious consequence of using small samples sizes in research is an increase in errors (probably of all four types identified by Gelman & Carlin 2014: I, II, Sign, and Magnitude). By performing a large-scale comparison of the published results in linguistics with retests of those results using large samples of naïve participants, one can evaluate the quality of their convergence rate. This cannot identify specific errors, but it can tell us whether the differences between methods actually produce different results.

Sprouse and Almeida (2012) tested every English data point from a popular syntax textbook (Adger 2003) using large samples of naïve participants. Out of 365 phenomena, they conservatively estimate a minimum convergence rate of 98%. Sprouse et al. (2013) randomly sampled 148 phenomena from a leading linguistics journal (*Linguistic Inquiry*), and conservatively estimate a convergence rate of 95% ($\pm 5\%$ because of the random sampling). These high (*conservative*) convergence rates suggest that the sample sizes used by linguists (whatever they are) historically have introduced little error to the empirical record for any combination of the following reasons: (1) the samples are larger than what critics claim; (2) the effect sizes are so large that small samples still yield good statistical power; or (3) AJ results are highly replicated before and after publication (e.g., Phillips 2009).

Claim 2: Acceptability judgments are highly susceptible to theoretical cognitive bias because linguists tend to use professional linguists as participants (p. 13).

This can also be addressed by the studies discussed above. Cognitive bias should predict sign reversals between naïve and expert populations. Sprouse and Almeida (2012) found no sign reversals from the textbook data. Sprouse et al. (2013) reported a 1–3% sign reversal rate in the journal data. Mahowald et al. (2016) and Häussler et al. (2016) have replicated the latter without reporting an increased sign reversal rate (0–6%). Comparisons of naïve and expert populations also were conducted by Culbertson and Gross (2009), who report high inter- and intra-group correlations on 73 sentence types, and by Dąbrowska (2010). The latter found that, while experts gave less variable ratings than naïve participants on several sentence types, experts rated certain theoretically interesting syntactic violations as more *acceptable* than naïve participants, in apparent conflict with their theoretical commitments. Taken together, these results are not what one would expect if AJs were highly susceptible to cognitive bias.

Claim 3: Acceptability judgments are susceptible to differences in instructions (p. 13).

Claim 3 has been directly investigated by Cowart (1997), who reports that the systematic manipulation of instructions does not change the pattern of acceptability judgments for factorial designs.

Claim 4: Acceptability judgments are impacted by sentence processing effects (p. 13).

Claim 4 is technically true, but B&P exaggerate its consequences. First, many classic lexical and sentence processing effects have relatively small or negligible effects on acceptability (e.g., Featherston 2009; Phillips 2009; Sprouse 2008; Sprouse et al. 2012). Second, very few syntactic phenomena have been proposed to be fully reducible to sentence processing effects. The lone exceptions to this appear to be constraints on long-distance dependencies (e.g., Kluender & Kutas 1993; Hofmeister & Sag 2010), but in that case, a number of experimental studies have disproven the reductionist predictions

(Phillips 2006; Sprouse et al. 2012; Yoshida et al. 2014). Thus, to the extent that AJs are impacted by sentence processing, it appears as though the effects can be dealt with like any other source of noise in an experimental setting.

Claim 5: Acceptability judgments reveal only set membership (p. 14).

Claim 5 is confusing. It is false in the sense that, if one is interested in *set membership*, this property still needs to be *inferred* from acceptability data, using a logic that maps that data type back to the relevant cognitive computations. In this, AJs are like any other data type in cognitive science: No data types, including priming, directly reveal the underlying computations of the human brain, and all data types require a linking hypothesis between the observable data and the unobservable cognitive process.

Claim 6: Acceptability judgment have yielded no consensus theory among linguists (p. 15).

Claim 6 is a strange criticism to make of any data type, especially AJs. First, the beliefs of scientists are a subjective issue based on how they weigh different kinds of evidence. Second, AJs are, by all accounts, a robust and replicable data type. Whatever disagreements there are in linguistics literature, they appear to obtain mostly at the level of *interpreting*, not *establishing*, the data (e.g., Phillips 2009).

In conclusion, we support B&P's desire to bring new evidence to bear on questions about linguistic representation. We caution, however, that advocacy for one method should not be bolstered by misleading comparisons, especially with methods such as AJs, which yield data that is demonstrably robust, highly replicable, and comparatively convenient and inexpensive to collect.

References

- Adger, D. (2003) *Core syntax: A minimalist approach, vol. 33*. Oxford University Press.
- Cowart, W. (1997) *Experimental syntax: Applying objective methods to sentence judgments*. Sage.
- Culbertson, J. & Gross, S. (2009) Are linguists better subjects?. *The British Journal for the Philosophy of Science*, xpx032.
- Dąbrowska, E. (2010) Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27(1):1–23.
- Featherston, S. (2009) Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenschaft* 28(1):127–32.
- Gelman, A. & Carlin, J. (2014) Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9(6):641–51.
- Häussler, J., Juzek, T. & Wasow, T. (2016) To be grammatical or not to be grammatical – Is that the question? Poster presented at the Annual Meeting of the Linguistic Society of America.
- Hofmeister, P. & Sag, I. A. (2010) Cognitive constraints and island effects. *Language* 86(2):366.
- Kluender, R. & Kutas, M. (1993) Subjacency as a processing phenomenon. *Language and Cognitive Processes* 8(4):573–633.
- Mahowald, K., Graff, P., Hartman, J. & Gibson, E. (2016) SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language* 92(3):619–35.
- Phillips, C. (2006) The real-time status of island phenomena. *Language* 82(4):795–823.
- Phillips, C. (2009) Should we impeach armchair linguists? *Japanese/Korean Linguistics* 17:49–64.
- Sprouse, J. (2008) The differential sensitivity of acceptability judgments to processing effects. *Linguistic Inquiry* 39(4):686–94.
- Sprouse, J. & Almeida, D. (2012) Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics* 48(03):609–52.
- Sprouse, J., Wagers, M. & Phillips, C. (2012) A test of the relation between working-memory capacity and syntactic island effects. *Language* 88(1):82–123.

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* 134:219–48.

Yoshida, M., Kazanina, N., Pablos, L. & Sturt, P. (2014) On the origin of islands. *Language, Cognition and Neuroscience* 29(7):761–70.