# The reliability of acceptability judgments beyond English

**Introduction:** Introspective acceptability judgments are the principal source of data in theoretical syntax. The reliability of these judgments has come under criticism in recent years (Edelman & Christiansen, 2003; Gibson & Fedorenko, 2010). A series of studies have concluded that these concerns are largely misplaced (e.g., Sprouse & Almeida, 2012); in particular, a vast majority of the judgments in a Minimalist syntax textbook were validated in a large-scale experiment with native participants.

While these results are reassuring, they have so far been limited to English acceptability judgments. A commonly advanced argument for the reliability of published judgments is that questionable judgments are weeded out by means of **informal peer review**: "If a key judgment is questionable, this is likely to be pointed out by a colleague, or by audience members in a talk, or reviewers of an abstract or journal article" (Phillips, 2009). Informal peer review for English judgments is made possible by the large community of linguists who are native English speakers, as well as the fact that even those linguists who are not native speakers of English are highly proficient in that language.

Is the current level of informal peer sufficient to guarantee the validity of judgments in languages that are less widely spoken than English? To address this question, we investigated the reliability of acceptability judgments in Hebrew and Japanese. While both languages are relatively well-studied, work in those languages is unlikely to enjoy the same level of informal peer review as does work on English.

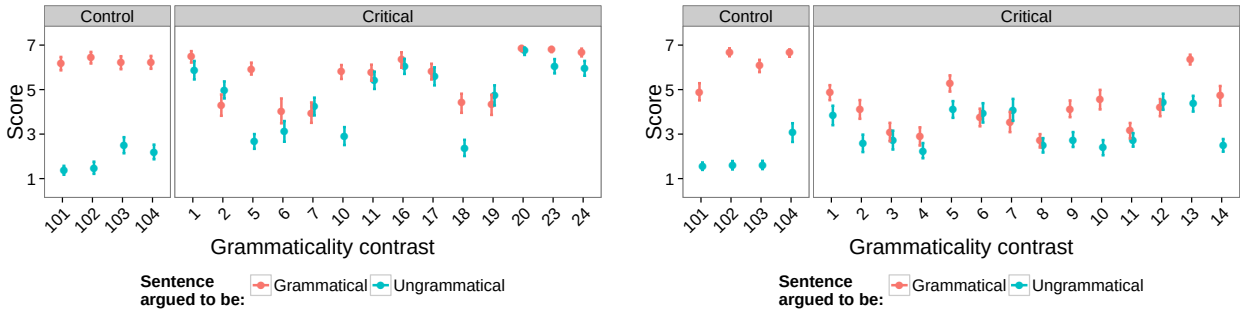
Questionable acceptability contrasts, selected from the literature by a native speaker of each of the languages, were rated for acceptability by a large sample of naive participants. Only approximately half of the controversial contrasts were replicated. We detail below the experimental design and results, and discuss the methdological implications of our findings.

**Methods:** Both experiments were administered on a website created for this purpose. Participants who were native speakers of Hebrew or Japanese were recruited on Facebook. The instructions were translations of those used by Sprouse and Almeida (2012). Participants rated each sentence on a scale of 1 (very bad) to 7 (very good). The intermediate steps were not labeled. Stimuli in each experiment were 14 questionable judgments ("Critical") selected from peer-reviewed journals and books, as well as 4 uncontroversial contrasts ("Control") included to make sure that the participants understood the task.

**Hebrew results:** A total of 76 participants completed the experiment. The control contrasts were robustly replicated: the grammatical sentences were rated higher than 6 on average and the ungrammatical sentences were rated lower than 2. The difference in acceptability was significant only in 7 out of the 14 questionable contrasts. Even in some of the contrasts that were successfully replicated, both the grammatical and the ungrammatical sentences were rated higher than 6, casting doubt on whether the lower acceptability sentence should be marked as ungrammatical. Participants' ratings of three of the contrasts went in the opposite direction from the original prediction, such as the following judgment (Shlonsky, 1992):

(1) elu ha-sfarim she-Dan tiyek otam bli likro *(otam).
these the-books that-Dan filed them without to-read them
'These are the books that Dan filed without reading.'

**Japanese results:** A total of 98 participants completed the experiment. The control contrasts were again robustly replicated. As in the Hebrew experiment, the mean score of the grammatical sentences was above 6 and one of the ungrammatical sentences was below 2. Repeated measures t-tests for each contrast revealed that only 8 out of 14 questionable contrasts were statistically significant. Many of the critical contrasts went in the predicted direction; in some cases the acceptability of both sentences of the contrast were rated lower




(a) (b)

Figure 1: Mean acceptability ratings: (a) Hebrew; (b) Japanese. Error bars represent bootstrapped 95% confidence intervals. Contrasts in which the two bars overlap substantially are in general not significantly replicated.

than 3. Once again, three contrasts were judged in the opposite direction from the original prediction, e.g. (Miyagawa, 2003):

(2) Hanako-ga Mary-no kanozyo-ga/*no kiitakotono-nai hihan-o sita.
Hanako-Nom Mary-Gen she-Nom/Gen heard-Neg criticism-Acc did
'Hanako did Mary's criticism that Mary/*Hanako hasn't heard.' (Free translation: 'Hanako made a criticism of Mary that Mary/*Hanako hasn't heard.')

**Discussion:** About half of the judgments that we (the authors) believed to be questionable did not stand up to an empirical test: some of the differences failed to reach significance even with a large sample size; others showed a numerical difference in the opposite direction from the predicted one. This result contrasts with English, where judgments appear to be robustly replicable (Sprouse & Almeida, 2012).

Some of the contrasts that were not replicated were drawn from papers that are widely cited (298 and 137 Google Scholar citations for (1) and (2) respectively), and may therefore have affected the development of syntactic theory. Many of the linguists who may have relied on those judgments are not speakers of the relevant languages; as such, they are likely to be unable to evaluate the quality of the judgments.

We find it encouraging that our own scepticism about particular judgments was often validated. This suggests that informal peer review can indeed be quite effective (Phillips, 2009), and that large-scale acceptability judgment collection experiments, which are difficult to conduct in smaller linguistic communities, are normally unnecessary. To extend the benefits of informal peer review to languages with a smaller community of linguists who are native speakers, we propose implementing an online platform that would enable native speakers to vet acceptability judgments or express concern about them. It may even be advisable to encourage authors of journal articles to use such a system before their article is accepted for publication.

Edelman, S., & Christiansen, M. H. (2003). How seriously should we take minimalist syntax? *Trends in Cognitive Sciences*, *7*, 60–61.

Gibson, E., & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, *14*, 233–234.

Miyagawa, S. (2003). A-movement scrambling and options without optionality. In K. Simin (Ed.), *Word Order and Scrambling* (pp. 177–200). Blackwell.

Phillips, C. (2009). Should we impeach armchair linguists? *Japanese/Korean Linguistics*, *17*, 49–64.

Shlonsky, U. (1992). Resumptive pronouns as a last resort. *Linguistic inquiry*, 443–468.

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, *48*, 609–652.