Glossa a journal of general linguistics

Linzen, Tal and Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics* 3(1): 100. 1–25, DOI: https://doi.org/10.5334/gjgl.528

RESEARCH

The reliability of acceptability judgments across languages

Tal Linzen[1] and Yohei Oseki[2]

[1] Johns Hopkins University, 3400 N. Charles St., Baltimore, MD, US

[2] New York University, 10 Washington Place, New York, US

Corresponding author: Tal Linzen (tal.linzen@jhu.edu)

The reliability of acceptability judgments made by individual linguists has often been called into question. Recent large-scale replication studies conducted in response to this criticism have shown that the majority of published English acceptability judgments are robust. We make two observations about these replication studies. First, we raise the concern that English acceptability judgments may be more reliable than judgments in other languages. Second, we argue that it is unnecessary to replicate judgments that illustrate uncontroversial descriptive facts; rather, candidates for replication can emerge during formal or informal peer review. We present two experiments motivated by these arguments. Published Hebrew and Japanese acceptability contrasts considered questionable by the authors of the present paper were rated for acceptability by a large sample of naive participants. Approximately half of the contrasts did not replicate. We suggest that the reliability of acceptability judgments, especially in languages other than English, can be improved using a simple open review system, and that formal experiments are only necessary in controversial cases.

Keywords: acceptability judgments; reliability; experimental syntax; Hebrew; Japanese

1 Introduction

Acceptability judgments are a major source of data in linguistics. Most of the acceptability contrasts reported in the literature reflect the judgment of a single individual — the author of the article — occasionally with feedback from colleagues. The reliability of such judgments has repeatedly come under criticism (Langendoen et al. 1973; Schütze 1996; Edelman & Christiansen 2003; Gibson & Fedorenko 2010; Gibson et al. 2013). It has been argued, for example, that "the journals are full of papers containing highly questionable data, as readers can verify simply by perusing the examples in nearly any syntax article about a familiar language" (Wasow & Arnold 2005: 1484). If this criticism turned out to be correct, decades of syntactic theory would appear to be standing on shaky empirical ground. Although some of the critics have targeted generative syntacticians in particular for criticism, this issue applies to other research communities as well, as the debate around the use of introspective judgments in The Cambridge Grammar of the English Language illustrates (Huddleston & Pullum 2002a; b).

Other authors have defended the field's reliance on individual linguists' judgments (Phillips & Lasnik 2003; Featherston 2009; Phillips 2010). Proponents of this methodology have pointed out that while acceptability judgments are initially made by a single author, they are subsequently subjected to several stages of formal and informal peer review before being published, and as such are likely to be robust. This defense of judgments given by an individual author is supported by the results of a number of recent judgment

collection experiments, which replicated the overwhelming majority of English judgments from a Minimalist syntax textbook and from articles published in the generative linguistics journal *Linguistic Inquiry* (Sprouse & Almeida 2012; Sprouse et al. 2013; Mahowald et al. 2016).

The debate surrounding the reliability of acceptability judgments has so far been limited to English. While English is the source of a sizable proportion of the judgments in the literature — for better or worse, it accounted for about half of the syntactic acceptability judgments reported between 2001 and 2010 in *Linguistic Inquiry* (Sprouse et al. 2013) — theoretical developments in linguistics are often driven by data from other languages. The current study makes the first step in expanding the debate on judgment reliability beyond English by conducting acceptability judgment replication studies in Hebrew and Japanese. Although there are fairly active communities of syntacticians working on both of these languages, those communities are undoubtedly smaller than the community of syntacticians working on English; the English syntax community includes not only a large number of native English speakers but also an even larger number of linguists who do not speak English as a native language but are highly proficient in that language.

The contrasts selected for replication in previous experiments were sampled at random. In those studies, a large representative sample was necessary to estimate the proportion of reliable judgments in a particular body of work (Sprouse et al. 2013; Mahowald et al. 2016). As we argue in Section 2, however, many of the judgments in the literature are self-evident; for example, a syntax textbook might include the example **the bear snuffleds* to illustrate that past tense forms in English do not carry overt person agreement marking. These are not the judgments that critics take issue with; short of deliberate fraud by the author, such judgments are very likely to be replicable. Since such judgments are fairly common, the replication rate in a study based on a random sample of sentences conflates the proportion of self-evident judgments in the sample with the reliability of potentially questionable judgments, and is therefore difficult to interpret.

Our goal is to show that questionable contrasts can be reliably identified by a linguist, and consequently a more extensive review process, such as the one to which English judgments tend to be subjected, would have kept those judgments out of the published record. We therefore do not attempt to construct a representative sample of judgments; instead, we undersample self-evident contrasts (four in each language) and oversample ones that we as native speakers of Hebrew and Japanese deemed to be questionable (14 in each language).

The rest of this paper is organized as follows. Section 2 describes the methods we used to conduct acceptability judgment replication experiments in Hebrew and Japanese. In Section 3, we show that half of the Hebrew contrasts and a third of the Japanese contrasts that we deemed to be questionable failed to replicate in formal experiments. In Section 4, we discuss the interpretation of these results and suggest ways in which the benefits of informal peer review can be extended beyond English. Section 5 concludes the paper.

2 Methods

2.1 Participants

We conducted two acceptability rating experiments: one in Hebrew and one in Japanese. The Hebrew experiment was completed by 76 participants, and the Japanese experiment by 98 participants. All participants were volunteers recruited through Facebook (it is difficult to recruit a large enough sample of Hebrew and Japanese speaking subjects on paid platforms such as Amazon Mechanical Turk). We asked participants not to participate in the study if they did not satisfy the following two conditions: (1) they lived in

Israel/Japan in the first 13 years of their lives, except for short breaks; and (2) their parents spoke Hebrew/Japanese to them.

2.2 Materials

As we mentioned above, we did not attempt to select a representative sample of judgments from the literature. We illustrate the motivation for this decision using the three-way classification of syntactic judgments proposed by Marantz (2005). The first category of judgments discussed by Marantz, which we refer to as Class I judgments, consists of "word salads" — sequences of words that are so far from the grammar of the language that they cannot even be assigned a phonological representation. The following "word salad", for example, illustrates what English sentences would look like if English were a head-final language like Japanese (Marantz 2005: 433):

(1) *Man the book a woman those to given has.

The second category (Class II) includes judgments that illustrate uncontroversial facts about the grammar of the language, facts of the sort that might be presupposed in a theoretical analysis. The following contrast, for example, shows that English verbs agree in number with their subject (Marantz 2005: 434):

(2) a. The men are leaving.
b. *The men is leaving.

The third category (Class III) includes more subtle contrasts, such as constraints on *wh*-movement or on possible coreference relations across noun phrases. The judgments that critics take issue with typically fall into this category. Gibson et al. (2013) refer to this class as "theoretically meaningful contrasts". We prefer to use the more neutral term Class III judgments: it is often difficult to assess the theoretical significance of a particular contrast, and it is not clear whether there is a relationship between the "obviousness" of a judgment and its theoretical import.

The work on English by Sprouse and colleagues attempted to replicate a random sample of all published judgments, regardless of their class. One of the contrasts from Adger (2003) replicated by Sprouse & Almeida (2012) is shown in (3a):

(3) a. The bear snuffled.
b. *The bear snuffleds.

This contrast was replicated by a large margin, as were other Class II judgments. One might object that textbooks such as Adger (2003) contain more Class II judgments for pedagogical reasons. In reality, however, such judgments are also fairly common in the sample from *Linguistic Inquiry* investigated by Sprouse et al. (2013), e.g.:

(4) a. I hate eating sushi.
b. *I seem eating sushi.

(5) a. Wallace and Greg like each other.
b. *Each other like Wallace and Greg.

We can be fairly confident that judgments of this type will be reliably replicated with naive subjects (Mahowald et al. 2016). Consequently, our study focuses on Class III judgments. For the experiments reported below, we — linguists who are native speakers of Hebrew or Japanese — selected 18 acceptability contrasts in each language, 14

of which were Class III contrasts from the literature that we believed were potentially questionable (henceforth "critical items") and four were uncontroversial Class II contrasts (henceforth "control items").

We limited the total number of contrasts in each language to 18 to keep the experiment short (this was necessary since all of our participants were volunteers). While we did not record the overall number of judgments (questionable and unquestionable) in the articles we examined,[1] it is clear that there were many more unquestionable judgments than questionable ones. For example, Borer (1995), which was the source of three of our questionable Hebrew contrasts, contains more than a hundred Hebrew examples. At the same time, we did not attempt to compile an exhaustive list of all questionable judgments in the articles we have examined; the particular paper mentioned above, for example, contains additional questionable judgments that are relatively similar to the three judgments we tested and so were not included in our experiment.

The full list of materials is given in Section 2.3 (for Hebrew) and Section 2.4 (for Japanese); these sections can be safely skipped in a first reading of this article.

2.3 Hebrew contrasts

The Hebrew judgments were primarily drawn from peer-reviewed articles, in particular the Special Hebrew Issue of *Natural Language and Linguistic Theory* (August 1995) and other issues of *Natural Language and Linguistic Theory* and *Linguistic Inquiry*, as well as from two books: a collection of articles (Armon-Lotem et al. 2008) and a frequently cited dissertation published as a book (Shlonsky 1997). Some of the Hebrew judgments concerned DPs (noun phrases) rather than entire sentences, such as the following contrast (Belletti & Shlonsky 1995: 517):

(6) a. ha-haxzara shel ha-shtaxim la-falastinim
the-handing.over of the-territories to.the-Palestinians

b. *ha-haxzara la-falastinim shel ha-shtaxim
the-handing.over to.the-Palestinians of the-territories

We embedded these DPs in simple sentences; the added material is represented in the following section using squared brackets. All of the contrasts involved judgments on strings (is this sentence acceptable?), rather than judgments under an interpretation (can this sentence have this particular meaning?). The Hebrew judgments concerned word order (H3, H8, H9, H10, H13), argument optionality (H12) and omissibility of elements in coordination (H5, H14), among other phenomena.

The articles that the judgments were drawn from used a variety of romanization schemes; here we use a unified scheme that reflects modern pronunciation (*x* represents the voiceless velar fricative [x]). We kept the glosses used in the original articles even when our own judgments about the meaning of certain words diverged from the original authors'.

2.3.1 Critical items

(7) H1 (Arad 2003)
a. Dani tsava et ha-kirot be-laka.
Dani painted ACC the-walls in-varnish
'Dani painted the walls with varnish.'

[1] Indeed, it is difficult to count the number of contrasts in a given paper exactly since many of the examples are not given as explicit contrasts (e.g, a single starred sentence is given and the reader is expected to infer its unstarred counterpart or counterparts from the context).

b. *Dani siyed et ha-kirot be-laka.
Dani whitewashed ACC the-walls in-varnish
'Dani whitewashed the walls with varnish.'

(8) H2 (Shlonsky 1992)
a. 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.'
b. *Elu ha-sfarim she-Dan tiyek otam bli likro.
these the-books that-Dan filed them without to-read
'These are the books that Dan filed without reading.'

(9) H3 (Belletti & Shlonsky 1995)
a. [Hitvakaxnu al] ha-haxzara shel ha-shtaxim la-falastinim.
[we.argued about] the-return of the-territories to.the-Palestinians
'We argued about the return of the territories to the Palestinians.'
b. *[Hitvakaxnu al] ha-haxzara la-falastinim shel ha-shtaxim.
[we.argued about] the-return to.the-Palestinians of the-territories
'We argued about the return to the Palestinians of the territories.'

(10) H4 (Shlonsky 1990)
a. Et mi lo yadata im ha-xayalim atsru?
ACC who no you.knew if the-soldiers detained
'Who didn't you know whether the soldiers detained?'
b. *Mi lo yadata im ne'etsar al-yedei ha-xayalim?
who no you.knew if was.detained by the-soldiers
'Who didn't you know whether [he] was arrested by the soldiers?'

(11) H5 (Borer 1995)
a. Dani muxan haya ba-zman ve-Rina muxana hayta gam.
Dani ready was on-time and-Rina ready was too
'Dani was ready on time and Rina was too.'
b. *Dani muxan haya ba-zman ve-Rina hayta gam.
Dani ready was on-time and-Rina was too
'Dani was ready on time and Rina was too.'

(12) H6 (Borer 1995)
a. Ha-ne'arot shalxu kulan mixtavey mexa'a la-memshala.
the-women sent all letters protest to.the-government
'The women all sent protest letters to the government.'
b. ??Eize mixtavim ha-ne'arot shalxu kulan la-memshala?
which letters the-women sent all to.the-government
'Which letters did the women all send to the government?'

(13) H7 (Borer 1995)
a. Ratsinu lish'ol eifo ha-ma'avak ha-axaron ihiye.
we.wanted to.ask where the-struggle the-last will.be
'We wanted to ask where the last struggle will be.'

b. *Eifo ha-ma'avak ha-axaron ihiye?
where the-struggle the-last will.be
'Where will the last struggle be?'

(14) H8 (Shlonsky 2004)
a. [Ra'iti] para shveitsarit xuma.
[I.saw] cow Swiss brown
'I saw a brown Swiss cow.'
b. *[Ra'iti] para xuma shveitsarit.
[I.saw] cow brown Swiss
'I saw a Swiss brown cow.'

(15) H9 (Shlonsky 2004)
a. [Hu lakax et] ha-shulxan ha-shaxor ha-arox [sheli].
[he took ACC] the-table the-black the-long [my]
'He took my long black table.'
b. *[Hu lakax et] ha-shulxan ha-arox ha-shaxor [sheli].
[he took ACC] the-table the-long the-black [my]
'He took my black long table.'

(16) H10 (Siloni 1996)
a. [Shamanu al] ha-harisa shel ha-tsava et ha-ir.
[we.heard about] the-destruction of the-army ACC the-city
'We heard about the army's destruction of the city.'
b. *[Shamanu al] ha-harisa et ha-ir shel ha-tsava.
[we.heard about] the-destruction ACC the-city of the-army
'We heard about the army's destruction of the city.'

(17) H11 (Siloni 1995)
a. Ad ha-shana she-avra kol ha-klavim ha-noshxim et ba'aley-hem
until the-year last all the-dogs the-biting ACC owners-their
hayu mumatim.
were killed
'Until last year, all the dogs biting their owners were killed.'
b. *Ad ha-shana she-avra kol ha-klavim she-noshxim et ba'aley-hem
until the-year last all the-dogs that-bit ACC owners-their
hayu mumatim.
were killed
'Until last year, all of the dogs who bit their owners were killed.'

(18) H12 (Preminger 2009)
a. Dan masar et ha-ma'atafa la-mefakeax.
Dan handed ACC the-envelope to.the-supervisor
'Dan handed the envelope to the supervisor.'
b. *Dan masar et ha-ma'atafa.
Dan handed ACC the-envelope
'Dan handed the envelope.'

(19) H13 (Shlonsky 1997)
a. Ha-sfarim ne'elmu me-ha-sifriya.
the-books disappeared from-the-library
'The books disappeared from the library.'

b. *Ne'elmu ha-sfarim me-ha-sifriya.
disappeared the-books from-the-library
'The books disappeared from the library.'

(20) H14 (Botwinik-Rotem 2008)
a. Ha-sefer kal li-kri'a ve-le-nituax.
the-book easy to-reading and-to-analyzing
b. *Ha-sefer kal li-kri'a ve-nituax.
the-book easy to-reading and-analyzing
'The book is easy to read and to analyze.'

2.3.2 Control items

(21) H101 (anaphora binding)
a. Im-o shel ha-tinok ra'ata oto.
mother-his of the-baby saw him
'The baby's mother saw him.'
b. *Im-o shel ha-tinok ra'ata et atsmo.
mother-his of the-baby saw ACC himself
'The baby's mother saw himself.'

(22) H102 (number agreement)
a. Naflu le-Dani ha-maftexot.
fell.3PL to-Dani the-keys
b. *Ha-maftexot nafal le-Dani.
the-keys fell.3SG to-Dani
'Dani dropped his keys.'

(23) H103 (the relativizer *ha* can only be used directly before the present participle)
a. Hine ha-ish she-lo xoshev al kesef.
here the-man who-not think about money
b. *Hine ha-ish ha-lo xoshev al kesef.
here the-man the-not think about money
'Here is the man who doesn't think about money.'

(24) H104 (resumptive pronouns are not allowed in the subject position)
a. Ze ha-ish she-ohev le-daber al politika.
this the-man who-likes to-talk about politics
b. *Ze ha-ish she-hu ohev le-daber al politika.
this the-man who-he likes to-talk about politics
'This is the man who likes to talk about politics.'

2.4 Japanese contrasts

The Japanese judgments were selected from a number of sources: peer-reviewed papers published in *Natural Language and Linguistic Theory*, *Linguistic Inquiry*, and *Journal of East Asian Linguistics*, as well as in Japanese-specific journals; a dissertation published as a book (Miyagawa 1989); and three unpublished but widely cited dissertations (Farmer 1980; Hoji 1985; Oku 1998). Some of the Japanese judgments were bound to particular semantic interpretations (e.g., scope interpretations). In those cases in which the acceptability of the sentences was to be evaluated given a particular interpretation, explicit contexts were given to the participants; the participants were asked to rate the sentences

under those contexts. The English translations of the context sentences are indicated with parentheses.

2.4.1 Critical items

(25) J1 (Miyagawa 1989)
a. Kuruma-ga dorobou-ni ni-dai nusum-are-ta.
car-NOM thief-by two-CLF steal-PASS-PST
'Two cars were stolen by a thief.'
b. *Kodomo-ga geragerato san-nin warat-ta.
children-NOM loudly three-CLF laugh-PST
'Three children laughed loudly.'

(26) J2 (Kishimoto 2001)
a. Taro-wa nani-o kai-mo si-nakat-ta.
Taro-TOP anything-ACC buy-Q do-NEG-PST
'Taro did not buy anything.'
b. *Dare-ga warai-mo si-nakat-ta.
anyone-NOM laugh-Q do-NEG-PST
'Anyone did not laugh.'

(27) J3 (Miyagawa 2001)
(Taro took the exam, while Jiro didn't.)
a. Sono tesuto-o zen'in-ga uke-nakat-ta.
that exam-ACC everyone-NOM take-NEG-PST
'Everyone did not take that exam.' (not > all)
b. *Zen'in-ga sono tesuto-o uke-nakat-ta.
everyone-NOM that exam-ACC take-NEG-PST
'Everyone did not take that exam.' (not > all)

(28) J4 (Saito 1994)
a. Dare-ga naze nani-o kat-ta no?
who-NOM why what-ACC buy-PST Q
'Who bought what why?'
b. *John-ga naze nani-o kat-ta no?
John-NOM why what-ACC buy-PST Q
'What did John buy why?'

(29) J5 (Tada 1992)
(Taro winks.)
a. Taro-wa migime-dake-o tumur-e-ru.
Taro-TOP right.eye-only-ACC close-can-PRS
'Taro can wink his right eye.' (can > only)
b. *Taro-wa migime-dake-ga tumur-e-ru.
Taro-TOP right.eye-only-NOM close-can-PRS
'Taro can wink his right eye.' (can > only)

(30) J6 (Sakai 1994)
a. Hanako-ga Mary$_i$-no, kanozyo$_i$-ga kii-takotono-nai hihan-o sita.
Hahako-NOM Mary-GEN she-NOM hear-PRF-NEG critic-ACC did
'Hanako made Mary's criticism that she has not heard.'

b. *Hanako-ga Mary$_i$-no, kanozyo$_i$-no kii-takotono-nai hihan-o sita.
Hahako-NOM Mary-GEN she-GEN hear-PRF-NEG critic-ACC did
'Hanako made Mary's criticism that she has not heard.'

(31) J7 (Oku 1998)
(John's letter is in the garbage can./John's car is still dirty.)
a. Bill-wa zibun-no tegami-o sute, John-mo sute-ta.
Bill-TOP self-GEN letter-ACC throw John-Q throw-PST
'Bill threw Bill's letter, and John also threw John's letter.'
b. *Bill-wa kuruma-o teineini arat-ta-ga, John-wa araw-anakat-ta.
Bill-TOP car-ACC carefully wash-PST-but John-TOP wash-NEG-PST
'Taro washed a car carefully, but John did not wash a car carefully.'

(32) J8 (Hiraiwa 2010)
a. Naomi-o Ken-ga omoikkiri atama-o tatai-ta.
Naomi-ACC Ken-NOM hard head-ACC hit-PST
'Ken hit Naomi hard on the head.'
b. *Ken-ga omoikkiri Naomi-o atama-o tatai-ta.
Ken-NOM hard Naomi-ACC head-ACC hit-PST
'Ken hit Naomi hard on the head.'

(33) J9 (Farmer 1980)
a. Hanako-wa Taro-niyotte sono inu-o kaw-ase-rare-ta.
Hanako-TOP Taro-by that dog-ACC buy-CAUS-Pass-PST
'Hanako was made by Taro to buy that dog.'
b. *Sono inu-wa Taro-niyotte Hanako-ni kaw-ase-rare-ta.
that dog-TOP Taro-by Hanako-DAT buy-CAUS-Pass-PST
'That dog was made by Taro Hanako to buy.'

(34) J10 (Hoji 1985)
(Taro, Jiro, and Hanako ate a peach, a pear, and an orange, respectively.)
a. Dono kudamono-mo dareka-ga tabeta.
every fruit-Q someone-NOM ate
'Someone ate every fruit.' (every > some)
b. *Dareka-ga dono kudamono-mo tabeta.
someone-NOM every fruit-Q ate
'Someone ate every fruit.' (every > some)

(35) J11 (Miyagawa & Tsujioka 2004)
a. Taro-ga Hanako-ni Tokyo-ni nimotu-o okut-ta.
Taro-NOM Hanako-DAT Tokyo-DAT package-ACC send-PST
'Taro sent Hanako a package to Tokyo.'
b. *Taro-ga Tokyo-ni Hanako-ni nimotu-o okut-ta.
Taro-NOM Tokyo-DAT Hanako-DAT package-ACC send-PST
'Taro sent Hanako a package to Tokyo.'

(36) J12 (Boeckx & Niinuma 2004)
a. Hanako-ga Tanaka-sensei-ni Mary-o go-syookai-si-ta.
Hanako-NOM Tanaka-Prof.-DAT Mary-ACC HON-introduce-HON-PST
'Hanako introduced Mary to Prof. Tanaka.'

b. *Hanako-ga Mary-ni Tanaka-sensei-o go-syookai-si-ta.
Hanako-NOM Mary-DAT Tanaka-Prof.-ACC HON-introduce-HON-PST
'Hanako introduced Prof. Tanaka to Mary.'

(37) J13 (Saito 1992)
a. Taro to Hanako$_i$-ga otagai$_i$-o hihansi-ta.
Taro and Hanako-NOM each.other-ACC criticize-PST
'Taro and Hanako criticized each other.'

b. *Otagai$_i$-no sensei-ga Taro to Hanako$_i$-o hihans-ita.
each.other-GEN teacher-NOM Taro and Hanako-ACC criticize-PST
'Each other's teachers criticized Taro and Hanako.'

(38) J14 (Watanabe 2006)
a. Roger-wa donburi-ni yon-hai-no gohan-o tabe-ta.
Roger-TOP big.bowl-DAT four-CLF-GEN rice-ACC eat-PST
'Roger ate four big bowls of rice.'

b. *Roger-wa yon-hai-no gohan-o donburi-ni tabe-ta.
Roger-TOP four-CLF-GEN rice-ACC big.bowl-DAT eat-PST
'Roger ate four big bowls of rice.'

2.4.2 Control items

(39) J101 (quantifier floating)
a. Gakusei-ga san-nin sake-o non-da.
student-NOM three-CLF sake-ACC drink-PST
'Three students drank sake.'

b. *Gakusei-ga sake-o san-nin non-da.
student-NOM sake-ACC three-CLF drink-PST
'Three students drank sake.'

(40) J102 (nominative-genitive conversion)
a. Ame-ga hut-ta.
rain-NOM fall-PST
'It rained.'

b. *Ame-no hut-ta.
rain-GEN fall-PST
'It rained.'

(41) J103 (double accusative constraint)
a. Ken-ga Naomi-ni sono hon-o yom-ase-ta.
Ken-NOM Naomi-DAT that book-ACC read-CAUS-PST
'Ken made Naomi read that book.'

b. *Ken-ga Naomi-o sono hon-o yom-ase-ta.
Ken-NOM Naomi-ACC that book-ACC read-CAUS-PST
'Ken made Hanako read that book.'

(42) J104 (nominal structure)
a. John-wa hon-o takusan kat-ta.
John-TOP book-ACC many buy-PST
'John bought many books.'

b. *John-wa hon takusan-o kat-ta.
John-TOP book many-ACC buy-PST
'John bought many books.'

2.5 Procedure

The experiments were administered using a website created for this purpose. The instructions were based on those used by Sprouse & Almeida (2012). The participants were requested to rate each sentence on a scale from 1 (very bad) to 7 (very good). We emphasized that an acceptable sentence was not necessarily one that would be approved by official language institutions, but rather one that would not sound out of place when uttered by a native speaker in a conversation. Only a single lexicalization of each contrast was presented to participants.

Our participants rated both members of each contrast separately; other sentences were presented between the two members of the contrast, as we describe below. This design differs from standard practice in cognitive psychology, where care is taken to ensure that the same participant is not exposed to multiple versions of the same item. We believe that the concerns that motivate this practice in cognitive psychology do not apply to the case of acceptability judgments, because the original data point is itself an explicit comparison between two sentences; in fact, in some judgment replication studies both members of the contrast are displayed simultaneously and participants are instructed to make a forced choice between them (Sprouse et al. 2013).

The stimuli were divided into two blocks; each block contained one of the members of each contrast. Participants were not made aware of this division. The assignment of contrast members to blocks was counterbalanced across participants: for a given contrast, approximately half of the participants rated the unstarred member of the contrast first, and the other half read the starred member first. The allocation of contrast members to blocks was performed such that each block contained an equal number of unstarred and starred sentences, to avoid response bias (Sprouse 2009). The order of sentences within each block was pseudo-randomized such that no more than three consecutive sentences had the same acceptability annotation (starred or unstarred). The uncontroversial judgments were presented first in each block, to familiarize the participants with the task. Finally, we ensured that the first three sentences presented to a participant always included both starred and unstarred sentences (presented without the stars, of course).

Participants in the Hebrew experiment also rated a few unpaired sentences for acceptability; these sentences appeared in a middle block, between the two blocks reserved for acceptability contrasts. The ratings of these sentences are not analyzed in the current paper.

3 Results

The mean acceptability ratings for each of the sentences are shown in Figure 1 (for Hebrew) and Figure 2 (for Japanese). We assessed the statistical significance of the results using a two-tailed paired *t* test for each contrast separately (see Sprouse et al. 2013 for a discussion of analysis methods for this paradigm). Before the ratings were entered into the *t* test they were normalized ("*z* transformed") within each participant by subtracting the participant's mean rating and dividing the result by the standard deviation of the participant's ratings. This transformation, whose aim is to correct for differences between participants in their use of the scale, affected the resulting *t* statistics only slightly; none of the qualitative results for an individual contrast depended on whether or not it was applied. The full numerical results are reported in Table 1 for Hebrew and Table 2 for Japanese.

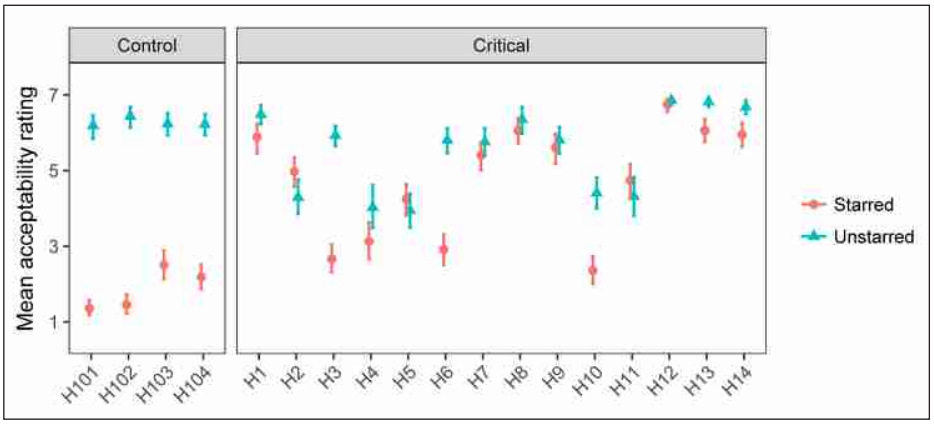

Figure 1: Results of the Hebrew experiment. Error bars represent bootstrapped 95% confidence intervals.

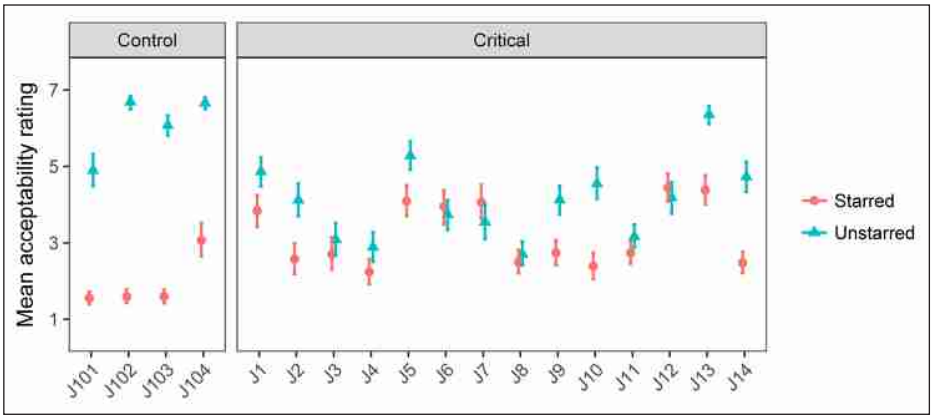

Figure 2: Results of the Japanese experiment. Error bars represent bootstrapped 95% confidence intervals.

3.1 Control contrasts

The control contrasts in both languages were robustly replicated (for all control contrasts, $t > 15, p < 0.001$). The average rating of each of the unstarred sentences was 5 or higher, whereas the starred sentences were rated 3 or lower.

3.2 Critical contrasts

3.2.1 Hebrew

Seven of the 14 Hebrew contrasts were replicated at the conventional statistical threshold of $p < 0.05$. Two contrasts showed a significant difference in the opposite direction than expected; in other words, the starred sentence was rated more highly than the unstarred one (H2: $p = 0.003$; H11: $p = 0.04$). The difference in ratings within the remaining five contrasts failed to reach significance. The sign of the difference in four of those contrasts was consistent with the originally reported judgments. This suggests that a larger sample size may result in a higher replication rate, though it should be kept in mind that our sample was already quite large ($n = 76$). Based on the variability of the responses, we estimate that the experiment was sensitive enough on average to detect a difference of 0.55 in ratings (see sensitivity analysis below).

3.2.2 Japanese

Ten of the 14 Japanese contrasts were replicated at the $p < 0.05$ level. The remaining four contrasts did not reach significance in either direction; the numerical difference in three

Table 1: Hebrew results: all participants. Legend: *0.01 < = p < 0.05; **p < 0.01; ! indicates that the starred member of the contrast received a higher average ranking than the unstarred member.

Contrast	Unstarred	Starred	Difference	t	p	Sig.
H1	6.49	5.89	0.60	3.20	0.002	**
H2	4.29	4.97	−0.68	−3.03	0.003	**!
H3	5.93	2.66	3.27	17.07	<0.001	**
H4	4.03	3.13	0.90	3.14	0.002	**
H5	3.94	4.25	−0.30	−1.45	0.152	!
H6	5.81	2.91	2.90	13.01	<0.001	**
H7	5.76	5.40	0.36	1.60	0.115	
H8	6.35	6.06	0.29	1.69	0.096	
H9	5.81	5.60	0.21	1.23	0.225	
H10	4.41	2.36	2.06	8.75	<0.001	**
H11	4.32	4.74	−0.42	−2.13	0.036	*!
H12	6.86	6.74	0.11	1.46	0.149	
H13	6.81	6.06	0.76	5.05	<0.001	**
H14	6.68	5.96	0.72	4.10	<0.001	**
H101	6.19	1.36	4.83	24.78	<0.001	**
H102	6.44	1.45	4.99	27.92	<0.001	**
H103	6.24	2.50	3.74	15.20	<0.001	**
H104	6.22	2.18	4.04	16.40	<0.001	**

Table 2: Japanese results: all participants.

Contrast	Unstarred	Starred	Difference	t	p	Sig.
J1	4.86	3.84	1.02	4.42	<0.001	**
J2	4.12	2.58	1.54	6.25	<0.001	**
J3	3.09	2.70	0.38	2.11	0.037	*
J4	2.89	2.24	0.65	3.22	0.002	**
J5	5.28	4.09	1.19	5.70	<0.001	**
J6	3.74	3.95	−0.21	−0.72	0.476	!
J7	3.54	4.06	−0.52	−1.88	0.064	!
J8	2.70	2.49	0.22	1.30	0.196	
J9	4.12	2.74	1.39	6.68	<0.001	**
J10	4.55	2.39	2.16	9.20	<0.001	**
J11	3.16	2.74	0.42	2.86	0.005	**
J12	4.18	4.45	−0.26	−1.39	0.169	!
J13	6.36	4.38	1.98	9.39	<0.001	**
J14	4.73	2.48	2.26	10.20	<0.001	**
J101	4.90	1.55	3.34	16.09	<0.001	**
J102	6.69	1.59	5.10	38.75	<0.001	**
J103	6.08	1.60	4.48	28.49	<0.001	**
J104	6.66	3.07	3.59	16.77	<0.001	**

of these contrasts went in the opposite direction than predicted. The higher proportion of replicated Japanese contrasts was not due to the larger sample of participants — the sensitivity of the Japanese experiment was almost identical to the Hebrew experiment, with

an average detectable difference of 0.54 — but rather to somewhat larger effect sizes: the average difference in ratings between unstarred and starred sentence in Japanese was 0.87 compared to 0.77 in Hebrew.

3.3 Variability across participants

Each of the participants rated both of the members of each contrast in their language. This makes it possible to investigate the distribution of the differences in ratings between the unstarred and starred member of each contrast (shown in Figure 3). In an ideal replication, all of the difference scores would be positive: every participant would rate the unstarred member higher than the starred one. This was only the case for one contrast (J102), though the other control contrasts approached this ideal picture; for example, only one

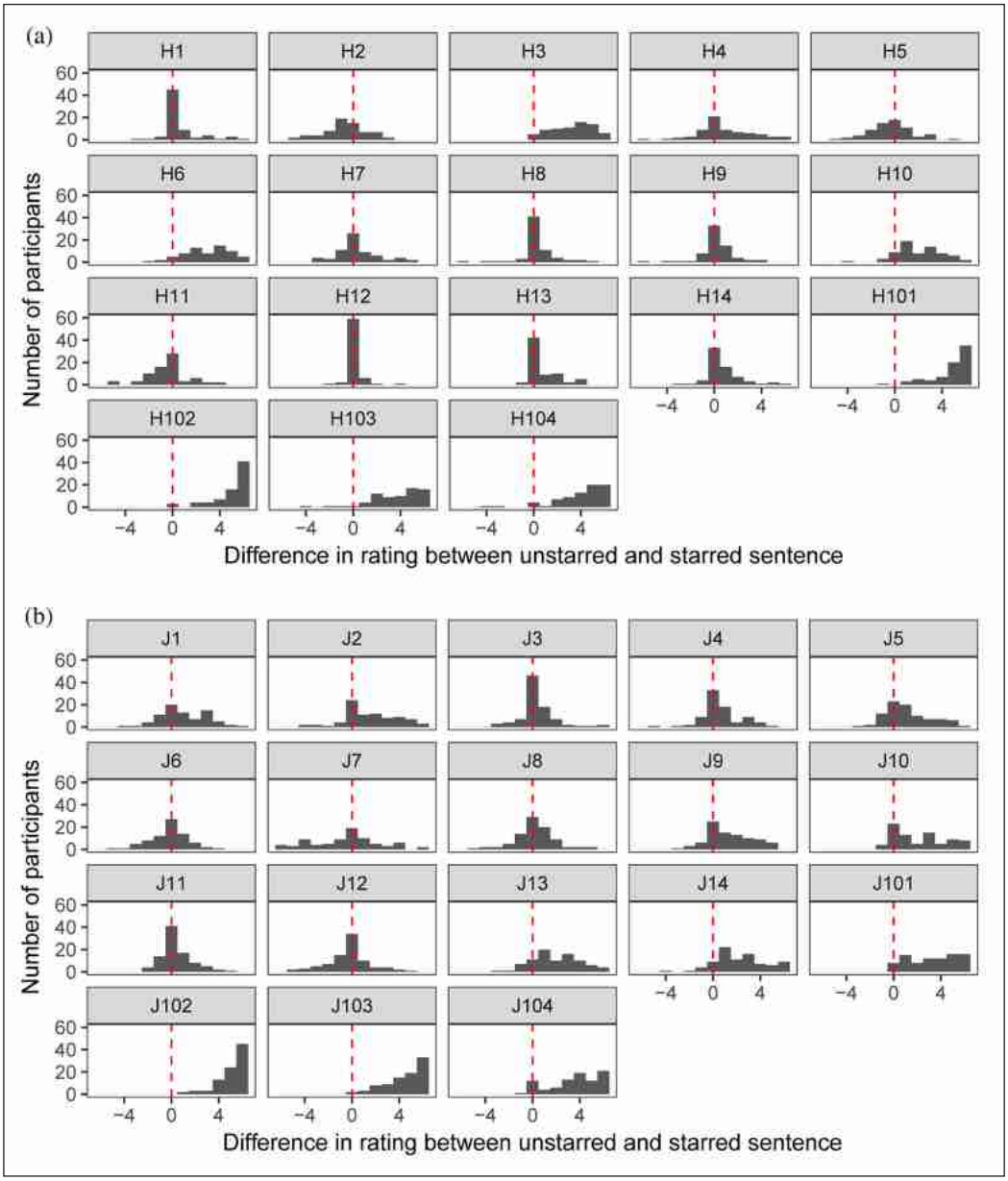

Figure 3: Histograms of the differences in ratings between the unstarred and starred members of each contrast in **(a)** Hebrew and **(b)** Japanese. H101–H104 and J101–J104 are the control contrasts. The bars at 0 (marked with dashed vertical lines) indicate the number of participants who gave the two members of the contrast the same rating. In a replicated contrast, most of the responses are expected to be to the right of this line.

out of 76 participants rated the starred member of H101 higher than the unstarred one. Most of the critical contrasts showed considerable variability; the most common difference score was often 0, indicating that a plurality of the participants gave the same rating to both members of the contrast.

In principle, the absence of a significant difference between the average ratings of the unstarred and starred members of a contrast could reflect dialectal differences (see Section 4.5.1): if there are two dialects, one consistent with the original judgment and the other consistent with its opposite, the two dialects could cancel out when the average is computed across all participants. The histograms in Figure 3 do not provide clear evidence for such an interpretation: the distributions appear to be unimodal (having a single peak), rather than bimodal as the dialectal differences hypothesis would predict.

3.4 Variability across contrasts

From the perspective of linguistic theory, only differences in rating within each contrast are relevant to the replicability debate. Syntactic theories typically do not make predictions about differences across unrelated contrasts; such differences could in principle be due to any number of non-syntactic factors (e.g., plausibility or lexical predictability). We nevertheless comment on the striking variability in ratings across contrasts. For instance, while contrast H8 was replicated ($p < 0.001$), its starred version received an average rating of 6.06 — higher than the rating of nine of the 14 unstarred critical sentences in the Hebrew experiment (e.g., the rating of the unstarred member of the replicated contrast H4 was 4.03). In Japanese, while contrast J4 was replicated ($p < 0.001$), its unstarred version was rated 2.89 on average — lower than six of the 14 starred critical sentences in the Japanese experiment.

This pattern of results illustrates the care that should be taken in relating acceptability to grammaticality; clearly, it makes little sense to interpret a mean acceptability rating of 6.06 as showing that a sentence is ungrammatical if a mean acceptability rating of 2.89 is taken to show that a sentence is grammatical. At the same time, precisely because acceptability ratings reflect the influence of a multitude of factors, the fact that both sentences in a replicated contrast were rated very highly casts doubt on the conclusion that the slightly diminished acceptability of the starred sentence was due to ungrammaticality (the fact that it categorically cannot be generated by the grammar) rather than due to other factors, such as pragmatics, frequency or gradient preferences. In cases of syntactic variation, for example, one variant may be moderately but systematically preferred to another, even though neither variant is ungrammatical in the categorical sense.

3.5 Sensitivity of tests

We defined the sensitivity of our tests as the minimal mean difference in ratings between the unstarred and starred member of a contrast for which our power to detect the difference with a $p < 0.05$ threshold was at the standard level of 0.8 (calculated using the `power.t.test` function in R; Cohen 1992). We estimated the standard deviation of the difference in ratings by averaging the empirical standard deviations across all contrasts within a given language; the resulting estimated standard deviation was 1.7 for Hebrew and 1.88 for Japanese. The corresponding sensitivity estimates were 0.55 and 0.54, respectively.

3.6 Reanalysis with a smaller sample size

The number of participants in our experiments was fairly large (around twice that of Sprouse & Almeida 2012, for example). Such samples are not always easy to obtain for less widely spoken languages. To determine whether the statistical significance of our findings crucially depended on the large sample size, we repeated our analysis, this

time restricting ourselves to the responses given by an arbitrarily selected subset of 20 participants (without collecting any new data).

With the smaller sample size, only four of the Hebrew contrasts were replicated at the conventional statistical threshold of $p < 0.05$. None of the differences in the remaining ten contrasts reached significance; four out of these were negative, and the other six were positive. In Japanese, seven of the 14 contrasts were replicated, one (J6) showed a significant difference in the opposite direction than predicted, and the remaining six contrasts did not reach significance. The detailed results of the subset analysis are shown in Table 3 for Hebrew and in Table 4 for Japanese. Since the participants in the smaller sample were selected at random, any differences in the pattern of results between the two analyses (e.g., the reversal of the sign of the nonsignificant contrast H7) are due only to sampling noise.

We conclude that given the small effect sizes of the differences in rating in contrasts such as the one we tested, a large number of participants (perhaps 100) is necessary to obtain clear results. The sensitivity of the paradigm can be increased by including multiple lexicalizations of the same contrast for each subject — i.e. creating multiple versions of the same contrast by replacing some lexical items with equivalent items — or by presenting both members of each contrast simultaneously (Sprouse & Almeida 2017).

4 Discussion

Half of the Hebrew contrasts and a third of the Japanese contrasts that we identified as potentially controversial did not replicate in formal experiments. The experiments included a relatively large number of participants, and were sufficiently powerful to detect a difference in rating of approximately half a point on a 7-point Likert scale. Our participants rated six of the controversial contrasts (three in each language) in the opposite direction from the originally reported judgments: the starred sentences received higher

Table 3: Hebrew results: a sample of 20 participants.

Contrast	Unstarred	Starred	Difference	*t*	*p*	Sig.
H1	6.58	6.16	0.42	0.97	0.345	
H2	5.31	5.50	−0.19	−0.31	0.762	!
H3	6.00	2.65	3.35	8.20	<0.001	**
H4	3.84	3.21	0.63	1.25	0.227	
H5	3.47	3.89	−0.42	−1.21	0.242	!
H6	5.89	3.50	2.39	4.99	<0.001	**
H7	6.11	6.39	−0.28	−0.98	0.339	!
H8	6.37	5.58	0.79	1.95	0.067	
H9	6.20	6.00	0.20	0.75	0.468	
H10	4.44	1.78	2.67	5.69	<0.001	**
H11	4.11	4.58	−0.47	−1.05	0.309	!
H12	6.95	6.74	0.21	1.68	0.110	
H13	6.89	6.37	0.53	1.90	0.073	
H14	6.79	5.95	0.84	2.38	0.028	*
H101	6.47	1.37	5.11	16.13	<0.001	**
H102	6.35	1.45	4.90	15.06	<0.001	**
H103	6.10	2.15	3.95	9.77	<0.001	**
H104	6.80	1.75	5.05	16.46	<0.001	**

Table 4: Japanese results: a sample of 20 participants.

Contrast	Unstarred	Starred	Difference	*t*	*p*	Sig.
J1	5.43	3.86	1.57	3.92	0.002	**
J2	4.19	2.31	1.88	4.03	0.001	**
J3	3.40	3.65	−0.25	−1.29	0.211	!
J4	3.19	1.88	1.31	3.86	0.002	**
J5	5.72	4.56	1.17	2.51	0.023	*
J6	3.17	4.67	−1.50	−2.58	0.025	*!
J7	3.50	3.50	0.00	0.10	0.924	
J8	2.84	2.68	0.16	0.45	0.659	
J9	3.70	3.05	0.65	1.41	0.176	
J10	4.85	2.90	1.95	4.51	<0.001	**
J11	3.40	2.60	0.80	2.10	0.049	*
J12	4.42	4.21	0.21	0.28	0.783	
J13	6.50	4.45	2.05	4.58	<0.001	**
J14	5.05	2.53	2.53	5.10	<0.001	**
J101	4.71	1.41	3.29	6.24	<0.001	**
J102	6.78	1.50	5.28	24.23	<0.001	**
J103	6.00	1.67	4.33	18.14	<0.001	**
J104	6.56	3.00	3.56	7.96	<0.001	**

ratings than their unstarred counterparts (the difference in the unexpected direction was significant in two of these cases, and nonsignificant in the remaining four). By contrast, cases that we judged to be Class II contrasts (control items) were consistently replicated by a comfortable margin.

The results of the experiments presented in this paper indicate that individual linguists can identify controversial contrasts with considerable accuracy. If between a third and a half of the controversial judgments that a single linguist was able to identify did not replicate, the field as a whole is indeed likely to be able to identify the majority of questionable judgments (keeping in mind, of course, that we do not have an estimate of the number of questionable judgments that we *failed* to identify).[2] This validates the intuition that informal peer review can effectively weed out such judgments from the literature (Phillips 2010).

Our results reinforce the concern that some Class III contrasts in the literature may not be replicable, and suggest that replicability issues may be more common in languages other than English. Gibson et al. (2013) propose an uncompromising approach to addressing this concern: they argue that every acceptability judgment must be validated in a formal experiment. Our view is that given that the robustness of Class II contrasts is obvious to any native speaker of the language, it would be a waste of resources to test each and every judgment in a formal experiment (Culicover & Jackendoff 2010; Poeppel 2010), especially in smaller language communities where a large sample of participants would be difficult to recruit. We suggest that linguists concerned with data quality should focus on the small minority of potentially questionable Class III contrasts; formal acceptability

[2] An anonymous reviewer asks: if problematic judgments are so easy to identify, why do linguists include them in their articles in the first place? One explanation may be confirmation bias, where the linguist who advances a theoretical position tends to take a favorable view of judgments that support their position (Gibson et al. 2013).

rating experiments are necessary only in the cases in which there is disagreement among linguists about a particular Class III judgment.

4.1 The peer review process

Our results suggest that the peer review of Hebrew and Japanese judgments may be insufficient. To understand why, it is instructive to divide the review mechanisms discussed by Phillips (2010) into three stages.

The first stage is pre-publication peer review, which takes place primarily at conferences. As we have pointed out, pre-publication peer review is likely to be less rigorous in languages other than English: most conference are likely to have few, if any, native speakers of the language in question, with the exception of conferences that focus on particular language families.

The second stage is the formal review that takes place as part of the journal publication process. Many papers do not undergo this process at all (e.g., book chapters, dissertations and conference proceeding papers). Questionable judgments can slip even into papers that are formally peer-reviewed; indeed, some of the judgments that failed to replicate in our experiments were drawn from journal articles. This issue is likely to be more acute in articles published in journals that are not language-specific; the editors of those journals may not be able to find reviewers who are simultaneously native speakers of the language and experts on the theoretical topic of the article. Anecdotally, the four Japanese judgments in our sample that were drawn from peer-reviewed East Asian linguistics journals (*Journal of East Asian Linguistics* and *Journal of Japanese Linguistics*), where the reviewers were more likely to be native speakers, were replicated in our experiment; the four Japanese judgments that did not replicate were taken from books or general journals (*Natural Language and Linguistic Theory*).[3] Finally, judgments are even less likely to be vetted by a reviewer who is a native speaker when the paper is not predominately about a particular language but includes one or two judgments from each of several languages (such judgments are typically elicited from the authors' colleagues via email or in hallway conversations).

The third stage in the process outlined by Phillips (2010) can be referred to as historical peer review. Phillips argues that questionable judgments do not make it into the "lore" of the discipline: they are ignored by subsequent researchers. Yet it is unclear whether this process could be effective if those researchers do not speak the language and are therefore incapable of evaluating the original judgments. As an example, after conducting our experiments we discovered that contrast H8 in the Hebrew experiment has been challenged by Siloni (2001: Footnote 15), but this fact does not seem to have undermined the influence of the analysis motivated in part by that contrast (Shlonsky 2004). Indeed, it is unclear whether the field is aware of Siloni's challenge to the validity of the contrast: a later paper on Welsh cites the Hebrew contrast without noting the disagreement about its status (Willis 2006).

4.2 Improving the peer review process

The weaknesses of the peer review process for less widely spoken languages can be remedied in a straightforward way. We propose an online crowdsourced database of published acceptability judgments, modeled after existing community resources such as Stack Overflow and Urban Dictionary. To help linguists who are not experts on a particular language to discover existing post-publication criticisms of published judgments in that language, links between different papers that discuss a given judgment will be automatically generated to the extent possible (some manual annotation may be necessary

[3] We thank an anonymous reviewer for suggesting this analysis.

to complement this automatic process). Users will be given the option to comment on judgments online. Such comments might specify the set of contexts in which the judgment is valid, or provide attested examples that challenge it. A voting mechanism will allow users to quickly evaluate a judgment without commenting on it. Such "upvotes" and "downvotes" have been successful in weeding out uninformed answers to questions on websites such as Stack Overflow. The website could also provide facilities for collecting judgments from a large sample of naive participants, in the infrequent cases in which this will be found to be necessary.

Some of the issues with peer review processes as currently implemented apply to widely studied languages as well: a questionable English judgment that has made it into a published paper may mislead linguists who are not native English speakers and are not aware of the controversy surrounding the judgment. We therefore believe that work on English will also benefit from the online crowdsourced database we have sketched.

In a recently published paper, Mahowald et al. (2016) recognized that many contrasts are very robust and that large-scale experiments are not always necessary to validate them. They calculated that unanimous judgments from seven participants on seven unique lexicalizations of a contrast are sufficient to establish the robustness of the contrast (see also Myers 2009). While we are not convinced that even a lightweight experiment is necessary to establish the robustness of judgments such as **the bear snuffleds*, Mahowald et al.'s (2016) proposal strikes us as a reasonable middle ground between traditional methdology and the formal-experiments-only position expressed in Gibson et al. (2013); in fact, their proposal can be straightforwardly implemented using the platform we sketched above.

4.3 What is a failure to replicate?

A reviewer correctly points out that a failure to replicate a judgment does not necessarily indicate that the original judgment was incorrect; in particular, statistical analysis of the results of an experiment can fail to reach statistical significance due to insufficient statistical power (a Type II error) rather than because the underlying effect is exactly 0. The reviewer suggests that only sign reversals constitute evidence against a contrast provided in the literature. We disagree with this argument: even significant sign reversals can occur by chance, and are not always more informative than nonsignificant results (Gelman & Carlin 2014). In fact, the argument can be made that in the social sciences, including linguistics, no empirical effect is exactly zero. Given an extremely large sample size (say, five million subjects), any judgment would either be significantly replicated or yield a significant sign reversal; indeed, a randomly generated contrast would be "replicated" about half of the time.

In practice, the sample size of a replication experiment should be based on the minimal effect size that is seen as robust enough to inform theory formation in syntax. According to our sensitivity analysis, our experiments were able to detect a difference in ratings of 0.5 on a 7-point scale (at the conventional threshold of $p < 0.05$), with a sample size of 76 participants in Hebrew and 98 in Japanese. If linguists believe (1) that the difference in acceptability between a sentence that is generated by the grammar and a minimally different sentence that is not generated by the grammar can be much smaller than 0.5 on a 7-point scale, and (2) that an individual linguist can detect such a small effect by introspection, many thousands of participants will be necessary for adequately powered replications. Of course, the combination of these two assumptions creates a significant burden-of-proof asymmetry: the intuition of the original linguist is privileged over that of the dissenting linguist, who is required to provide evidence from an enormous number of subjects to support their position.

4.4 *Should we expect judgments to replicate?*

The notion that acceptability judgments given by linguists are expected to replicate in a representative sample of the population is not without its opponents. Some linguists have argued that judgments always reflect a particular linguist's idiolect, in which case replication studies with naive participants are entirely irrelevant — those participants may well have a different idiolect from the original author (Den Dikken et al. 2007). In a more nuanced criticism, Hoji (2015) argues that replication does have value, but only if it has been established that the participants' idiolect is similar to the original author's in the relevant respect. We cannot conclusively refute this objection. The lack of bimodality in the pattern of responses does not provide evidence for idiolectal variability in our sample, but it is of course possible that only a handful of participants shared the original author's idiolect; such a small subset of participants will not show up as a discernible second peak in the distribution.

4.5 *Limitations and future work*

4.5.1 Dialectal and generational variation

Dialectal variation in the population is hard to rule out definitively as an explanation for replication failures. We did not find evidence for different patterns of responses among our participants that could be ascribed to different dialects (see Section 3.3), but there may certainly be systematic dialectal or generational differences between our participants as a group and the original authors. We recruited our participants on the internet in the 2010s; it is quite plausible that they were at least one generation younger than the authors of the original papers, most of which were published in the 1980s and 1990s. This objection raises interesting questions about our ability to rely for theory construction on contrasts that have accumulated in the literature across different generations; these questions cannot be addressed by our data and are not specific to the interpretation of judgment replication experiments.

4.5.2 Comparison to English

We conjectured that English judgments are more reliable than judgments from other languages. Our empirical results are consistent with this conjecture but do not prove it. Our experiments included an intentionally biased sample of sentences; in contrast to previous experiments that attempted to replicate a random sample of English judgments, our design does not provide us with a simple way to estimate the proportion of judgments in Hebrew and Japanese that are difficult to replicate. We can nevertheless attempt to assess the difference between our results and the results of Sprouse and Almeida's work on English, in two ways (we thank Diogo Almeida for these suggestions).

First, out of the 148 sentence types that Sprouse et al. attempted to replicate, 13 were originally reported with a question mark (? or *?; Sprouse et al. 2013: 233), a sample size that is quite similar to ours. The linguists who originally provided these acceptability judgments may have anticipated objections to these judgments and used the question marks as a hedge indicating that these are "subtler" contrasts (presumably, with a smaller effect size). Yet only one out of these 13 test cases failed to replicate in the Sprouse et al. survey, a much lower rate than in our experiments (seven in Hebrew and four in Japanese); if the question mark annotation is a reliable guide to how subtle the judgment is, then, this analysis indicates that English judgments are more reliable than Hebrew and Japanese ones.

Alternatively, we can examine English contrasts that are more likely to be controversial based on the numerical effect size estimated in Sprouse et al.'s experiment. There were 20 contrasts for which the effect size was medium (0.5) or smaller in the Sprouse et al. sample of 148 sentence types. In the Likert Scale task of Sprouse et al. (the task we

used), five results went in the opposite direction than argued in the original articles (see their Table 3), and six went in the predicted direction but did not reach significance. This pattern is more similar to our Hebrew and Japanese results. The conflicting results of these two indirect methods suggest that a more direct comparison between Hebrew, Japanese and English, and perhaps additional languages, would be an important direction for future work.

4.5.3 Sample of languages

While our study expands the number of languages in which judgments replication studies have been conducted from one (English) to three, we did not test a representative sample of languages. In particular, neither of the languages we have investigated is as underrepresented in linguistics as Estonian, Maltese or Chichewa might be. Our choice of Hebrew and Japanese was a matter of convenience: we are native speakers of those languages; the existence of a medium-sized community of linguists working on those languages made it possible to test a diverse range of acceptability judgments made by multiple authors; and the fact that millions of people speak each of those languages facilitated recruiting a satisfactory number of experimental participants. If our concerns about the peer review process turn out to be well-founded, however, we expect replication failures in languages with even smaller research communities to be at least as common as in the languages we have examined here.

4.5.4 Theoretical import of judgments

As in earlier studies by Sprouse and colleagues, we did not attempt to trace the influence of each acceptability judgment (replicated or not) on subsequent linguistic theory building: our goal was to evaluate the quality of the *data* reported in linguistics papers rather than the quality of the *theories* constructed based on that data. This kind of detective work could be quite informative: it is not unreasonable to conjecture that data points that crucially support one theory over another face greater scrutiny, especially if those theories themselves are widely cited; such theoretically critical data points are likely to be more robust (Phillips 2010). Yet such an exercise, while certainly worthwhile, would not be straightforward. Theories are rarely constructed based on a single data point, and it is often unclear which particular data points are seen as crucially supporting a theory. This work is best left to experts on the theoretical domains that have been informed by those data points.

5 Conclusion

The vast majority of published English judgments can be replicated with naive participants (Sprouse & Almeida 2012; Sprouse et al. 2013). We argued that this is due to two reasons. First, a large proportion of acceptability judgments illustrate obvious and uncontroversial contrasts (Class I/II judgments). Second, more subtle contrasts (Class III judgments) are informally vetted by a large community of linguists who are native English speakers. While not foolproof, this informal peer review process weeds out most questionable judgments (Phillips 2010).

To examine the efficacy of the peer review process in languages other than English, we selected acceptability judgments in Hebrew and Japanese that we deemed to be questionable. Half (in Hebrew) or a third (in Japanese) of the Class III contrasts we selected failed to replicate, while all Class II judgments were robustly replicated. These results suggest that (1) formal acceptability rating experiments are not necessary for each and every judgment, (2) linguists can effectively identify questionable contrasts, and (3) informal peer review mechanisms may be less effective for languages spoken by a smaller number

of linguists. We proposed an online community resource that can extend the benefits of informal peer review to less widely spoken languages.

We stress that our results do not suggest that there is a "replicability crisis" in Hebrew or Japanese linguistics. Although we did not explicitly count the number of contrasts that we did *not* consider to be questionable, we estimate that there were dozens of such contrasts for each potentially questionable judgment. In other words, most judgments are not controversial. Our goal in this study was to point out that some potentially unreplicable judgments do exist in the literature, and those can be identified by linguists.

Abbreviations

ACC = accusative, CAUS = causative, CLF = classifier, DAT = dative, GEN = genitive, HON = honorific, NEG = negation, NOM = nominative, PASS = passive, PL = plural, PST = past, PRF = perfect, PRS = present, Q = question, SG = singular, TOP = topic.

Acknowledgements

We thank Alec Marantz for feedback and the seminar *Linguistics as Cognitive Science*, which led to this project. We also thank Jeremy Kuhn for inspiring the idea of a crowdsourced acceptability judgment database, and Diogo Almeida, Ariel Cohen-Goldberg, Ted Gibson, Kyle Mahowald, Omer Preminger, and Glossa reviewers for comments. Previous versions of this work were presented at the 2016 and 2018 Annual Meetings of the Linguistic Society of America.

Competing Interests

The authors have no competing interests to declare.

References

Adger, David. 2003. *Core Syntax*. Oxford: Oxford University Press.

Arad, Maya. 2003. Locality constraints on the interpretation of roots: The case of Hebrew denominal verbs. *Natural Language & Linguistic Theory* 21(4). 737–778. DOI: https://doi.org/10.1023/A:1025533719905

Armon-Lotem, Sharon, Gabi Danon & Susan Rothstein. 2008. *Current issues in generative Hebrew linguistics*. Amsterdam and Philadelphia: John Benjamins. DOI: https://doi.org/10.1075/la.134

Belletti, Andrea & Ur. Shlonsky. 1995. The order of verbal complements: A comparative study. *Natural Language & Linguistic Theory* 13(3). 489–526. DOI: https://doi.org/10.1007/BF00992739

Boeckx, Cedric & Fumikazu Niinuma. 2004. Conditions on Agreement in Japanese. *Natural Language & Linguistic Theory* 22(3). 453–480. DOI: https://doi.org/10.1023/B:NALA.0000027669.59667.c5

Borer, Hagit. 1995. The ups and downs of Hebrew verb movement. *Natural Language & Linguistic Theory* 13(3). 527–606. DOI: https://doi.org/10.1007/BF00992740

Botwinik-Rotem, Irena. 2008. Object gap constructions. In Sharon Armon-Lotem, Gabi Danon & Susan D. Rothstein (eds.), *Current issues in generative Hebrew linguistics*, 77–104. Amsterdam and Philadelphia: John Benjamins Publishing. DOI: https://doi.org/10.1075/la.134.04obj

Cohen, Jacob. 1992. A power primer. *Psychological Bulletin* 112(1). 155. DOI: https://doi.org/10.1037/0033-2909.112.1.155

Culicover, Peter W. & Ray Jackendoff. 2010. Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences* 14(6). 234–235. DOI: https://doi.org/10.1016/j.tics.2010.03.012

Den Dikken, Marcel, Judy B. Bernstein, Christina Tortora & Raffaella Zanuttini. 2007. Data and grammar: Means and individuals. *Theoretical Linguistics* 33(3). 335–352. DOI: https://doi.org/10.1515/TL.2007.022

Edelman, Shimon & Morten H. Christiansen. 2003. How seriously should we take minimalist syntax? *Trends in Cognitive Sciences* 7(2). 60–61. DOI: https://doi.org/10.1016/S1364-6613(02)00045-1

Farmer, Ann. 1980. *On the interaction of morphology and syntax*. Cambridge, MA: MIT dissertation.

Featherston, Sam. 2009. Relax, lean back, and be a linguist. *Zeitschrift für Sprachwissenschaft* 28(1). 127–132. DOI: https://doi.org/10.1515/ZFSW.2009.014

Gelman, Andrew & John Carlin. 2014. Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* 9(6). 641–651. DOI: https://doi.org/10.1177/1745691614551642

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

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

Hiraiwa, Ken. 2010. Spelling-out the Double-*o* Constraint. *Natural Language & Linguistic Theory* 28. 229–240. DOI: https://doi.org/10.1007/s11049-010-9098-9

Hoji, Hajime. 1985. *Logical form constraints and configurational structures in Japanese*. Seattle, WA: University of Washington dissertation.

Hoji, Hajime. 2015. *Language faculty science*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781107110731

Huddleston, Rodney & Geoffrey K. Pullum. 2002a. A response concerning The Cambridge Grammar. http://linguistlist.org/issues/13/13-1932.html.

Huddleston, Rodney & Geoffrey K. Pullum. 2002b. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781316423530

Kishimoto, Hideki. 2001. Binding of indeterminate pronouns and clause structure in Japanese. *Linguistic Inquiry* 32. 597–633. DOI: https://doi.org/10.1162/002438901753373014

Langendoen, D. Terence, Nancy Kalish-Landon & John Dore. 1973. Dative questions: A study in the relation of acceptability to grammaticality of an English sentence type. *Cognition* 2(4). 451–478. DOI: https://doi.org/10.1016/0010-0277(73)90004-8

Mahowald, Kyle, Peter Graff, Jeremy Hartman & Edward Gibson. 2016. SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language* 92(3). 619–635. DOI: https://doi.org/10.1353/lan.2016.0052

Marantz, Alec. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review* 22(2–4). 429–445. DOI: https://doi.org/10.1515/tlir.2005.22.2-4.429

Miyagawa, Shigeru. 1989. *Structure and case marking in Japanese*. San Diego: Academic Press.

Miyagawa, Shigeru. 2001. EPP, Scrambling, and Wh-in-situ. In Michael Kenstowicz (ed.), *Ken Hale: A life in language*, 293–338. Cambridge, MA: MIT Press.

Miyagawa, Shigeru & Takae Tsujioka. 2004. Argument structure and ditransitive verbs in Japanese. *Journal of East Asian Linguistics* 13(1). 1–38. DOI: https://doi.org/10.1023/B:JEAL.0000007345.64336.84

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

Oku, Satoshi. 1998. *A theory of selection and reconstruction in the Minimalist perspective*. Storrs, CT: University of Connecticut dissertation.

Phillips, Collin. 2010. Should we impeach armchair linguists? *Japanese/Korean Linguistics* 17. 49–64.

Phillips, Collin & Howard Lasnik. 2003. Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Sciences* 7(2). 61–62. DOI: https://doi.org/10.1016/S1364-6613(02)00046-3

Poeppel, David. 2010. An egregious act of methodological imperialism. http://www.talkingbrains.org/2010/06/egregious-act-of-methodological.html.

Preminger, Omer. 2009. Failure to agree is not a failure – *φ*-Agreement with postverbal subjects in Hebrew. In Jeroen van Craenenbroeck (ed.), *Linguistic Variation Yearbook 2009*, 241–278. Amsterdam and Philadelphia: John Benjamins Publishing.

Saito, Mamoru. 1992. Long distance scrambling in Japanese. *Journal of East Asian Linguistics* 1(1). 69–118. DOI: https://doi.org/10.1007/BF00129574

Saito, Mamoru. 1994. Additional-*wh* effects and the adjunction site theory. *Journal of East Asian Linguistics* 3(3). 195–240. DOI: https://doi.org/10.1007/BF01733064

Sakai, Hiromu. 1994. Complex NP Constraint and case conversion in Japanese. In Masaru Nakamura (ed.), *Current topics in English and Japanese*, 179–200. Tokyo: Hitsuji Shobo.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Shlonsky, Ur. 1990. *Pro* in Hebrew subject inversion. *Linguistic Inquiry* 21(2). 263–275.

Shlonsky, Ur. 1992. Resumptive pronouns as a last resort. *Linguistic Inquiry* 23(3). 443–468.

Shlonsky, Ur. 1997. *Clause structure and word order in Hebrew and Arabic: An essay in comparative Semitic syntax*. New York and Oxford: Oxford University Press.

Shlonsky, Ur. 2004. The form of Semitic noun phrases. *Lingua* 114(12). 1465–1526. DOI: https://doi.org/10.1016/j.lingua.2003.09.019

Siloni, Tal. 1995. On participial relatives and complementizer D^0: A case study in Hebrew and French. *Natural Language & Linguistic Theory* 13(3). 445–487. DOI: https://doi.org/10.1007/BF00992738

Siloni, Tal. 1996. Hebrew noun phrases: Generalized noun raising. In Andrea Beletti & Luiggi Rizzi (eds.), *Parameters and functional heads: Essays on comparative syntax*, 239–267. Oxford: Oxford University Press.

Siloni, Tal. 2001. Construct states at the PF interface. *Linguistic Variation Yearbook* 1. 229–266. DOI: https://doi.org/10.1075/livy.1.10sil

Sprouse, Jon. 2009. Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry* 40(2). 329–341. DOI: https://doi.org/10.1162/ling.2009.40.2.329

Sprouse, Jon, Carson T. Schütze & Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua* 134. 219–248. DOI: https://doi.org/10.1016/j.lingua.2013.07.002

Sprouse, Jon & Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48(3). 609–652. DOI: https://doi.org/10.1017/S0022226712000011

Sprouse, Jon & Diogo Almeida. 2017. Design sensitivity and statistical power in acceptability judgment experiments. *Glossa* 2(1). 1–32. DOI: https://doi.org/10.5334/gjgl.236

Tada, Hiroaki. 1992. Nominative objects in Japanese. *Journal of Japanese Linguistics* 14. 91–108. DOI: https://doi.org/10.1515/jjl-1992-0105

Wasow, Thomas & Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115(11). 1481–1496. DOI: https://doi.org/10.1016/j.lingua.2004.07.001

Watanabe, Akira. 2006. Functional projections of nominals in Japanese: Syntax of classifiers. *Natural Language & Linguistic Theory* 24(1). 241–306. DOI: https://doi.org/10.1007/s11049-005-3042-4

Willis, David. 2006. Against N-raising and NP-raising analyses of Welsh noun phrases. *Lingua* 116(11). 1807–1839. DOI: https://doi.org/10.1016/j.lingua.2004.09.004

How to cite this article: Linzen, Tal and Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics* 3(1): 100.1–25, DOI: https://doi.org/10.5334/gjgl.528

Submitted: 19 September 2017 **Accepted:** 28 May 2018 **Published:** 13 September 2018

]u[*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS