Satiation and Uncertainty in the Mid-Zone of Sentence Acceptability Judgments

Arthur Stepanov*

Abstract

Brown, Fanselow, Hall and Kliegl (2021) suggest that the syntactic satiation effect arises irrespective of sentence type, for those sentences whose acceptability status falls in the mid-zone range of a discrete Likert scale. They further propose to treat it as a 'mere exposure' effect, but it remains unclear why repeated exposure only targets the stimuli in the mid-zone area. In this note, I argue that mid-scale ratings form a region of highest uncertainty as reflected in maximum variance in speakers' ratings compared to the other regions of the scale. Satiation may consequently be seen as an exposure effect targeting the most unstable or 'volatile' portion of the judgments.

1. Introduction

It has been known for some time that speakers' judgments of sentence acceptability tend to improve after being exposed to certain sentence types more than once, a phenomenon known as *syntactic satiation* (Snyder (2000)). In their recent experimental treatment of the syntactic satiation effect in German and English multiple wh-questions (as in *Wer hat was gekauft?* and *Who bought what?*) where order and animacy of the two wh-phrases was manipulated, Brown et al. (2021) report an increase of acceptability in these constructions across six subsequent blocks of trials irrespective of whether the order of wh-phrases was in line with the syntactic Superiority condition or not (cf. Chomsky (1973)). Both the initial and subsequent ratings on these questions fell in the range of 2.75-5.25 on a 1-to-7 discrete Likert scale in that study. Brown et al. propose that the satiation effect was not restricted to the target multiple wh-questions but extended even to filler items whose acceptability also fell in that acceptability range. This is illustrated in Figure 1 for their Experiment 1. Brown et al. propose that satiation may in fact be

*I thank Doug Saddy, Penka Stateva and the participants of the workshop for the useful feedback. This work has been financially supported by the Slovenian Research Agency (ARIS) project no. J6-4615.

Gisbert Fanselow's Contributions to Syntactic Theory, 323–336
Artemis Alexiadou, Doreen Georgi, Fabian Heck, Gereon Müller & Florian Schäfer (eds.)
LINGUISTISCHE ARBEITS BERICHTE 96, Universität Leipzig 2024

indiscriminate to syntactic construction; rather, it is *a property of the mid-zone ratings in general*.

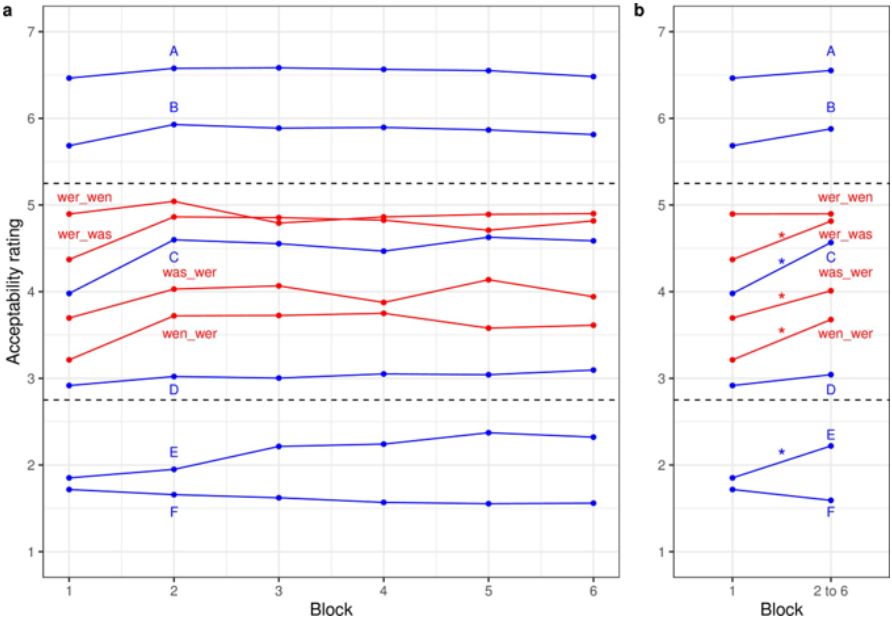

Figure 1: Well-formedness ratings on German multiple wh-questions (red) and unrelated fillers (blue) in Experiment 1 (a) by block (1-6) and (b) for first block and mean of later blocks. Dashed lines indicate the upper and lower bounds of the satiation zone. Asterisks indicate significant changes ($p<0.05$). Source: Brown et al. (2021).

This proposal shifts the focus of attention from the properties of specific syntactic configurations measured with a Likert scale to the speakers' perception of the scale itself as an instrument measuring subjective acceptability judgments, in particular, in the mid-zone area. Brown et al. (2021) suggest to view satiation as a 'mere exposure' effect in the sense of Zajonc (1968): repeated exposure of the individual to a stimulus enhances their attitude toward it without any further reinforcement. This still leaves a question why repeated exposure targets precisely the judgments in the mid-zone scale, in other words,

what is so special about it that makes it a source of subjective 'instability' of this kind. In this note I identify one way to approach this question.

2. Variance in the Mid-Zone

The null assumption that seems to underlie most syntactic studies using Likert scales is that speakers, including naive or non-trained linguists, have a priori the same access to all scale values. In other words, assuming a 1-to-*k* Likert scale, each value on that scale has an equal probability or chance of being selected. This can be represented using a probability mass function (PMF) for a discrete uniform distribution as follows:

$$p(X = i) = 1/((k-1)+1), \tag{1}$$

where *p(X = i)* is the probability of selecting Likert value *i*, $1 \leq i \leq k$. Such distribution of Likert values is characterized by the mean which is simply M= (*k*+1)/2 and the variance $\sigma^2 = (((k-1)+1)^2\text{-}1)/12$. For instance, for a 1-to-7 scale, the mean and variance is 4 (standard deviation = 2), and the corresponding PMF can be plotted as a series of 7 points all at the ordinate value of about 14%. This kind of distribution will ideally obtain if experimental participants are (for some reason) simply given the task of assigning a random Likert value to sentences they see or hear.

In actual sentence acceptability studies, of course, the underlying uniform distribution is modulated by the cognitive signal corresponding to the acceptability status of a particular sentence. In an experiment where it is a priori known that the evaluated sentences reflect the entire Likert value range, e.g. 1 to 7, in an equal proportion (an equal amount of sentences with the status "1", "2", and so on), the resulting distribution of speakers' ratings should not vary much from the original "carrier" uniform distribution (cf. (1)), because a more or less equal amount of each scale value will be given. This kind of result will demonstrate that speakers have an equal and unhindered access to each scale value. Any significant deviations from this pattern will indicate a bias in the usage of particular values.

Brown et al.'s (2021) study utilizes just the right setup to test this kind of prediction. In their experiments the authors use a set of 252 structurally diverse sentences that are grouped into six broadly balanced levels of acceptability ranging from fully acceptable and interpretable to fully unacceptable and uninterpretable, marked as types A-F in Figure 1. Most of these sentences

have been ranked in previous norming studies as occupying stable points in the 'judgment space' and used as 'calibrators' anchoring these points (cf. Featherston (2009) and Gerbrich et al. (2019)). This distribution of materials spanning the entire acceptability range has a good potential to counteract individual biases in the use of scale (usually dealt with by converting speakers' ratings into standard scores) and monitor the consistency of speakers' use of particular values of the scale at the population level via measures of central tendency, in particular, variance in the ratings.

Specifically, Brown et al. (2021) use the balanced set of fillers as acceptability 'calibrators' against which the target items are evaluated. Mean acceptability on the fillers marked as types A-F across trials are plotted with blue lines in Figure 1. In order to evaluate the prediction regarding uniformity of the scale use, we performed a series of additional post-hoc analyses of their rating data available at the article's OSF repository at https://osf.io/ge2db/. Figure 2 illustrates an aggregated post-hoc frequency analysis of specific scale values given by Brown et al.'s speakers on the filler sentences in Experiments 1 and 2.[1]

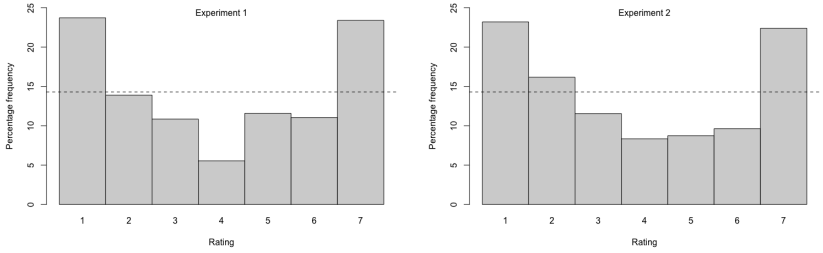

Figure 2: Percentage of occurrence of individual scale values on the filler sentences in Experiments 1-2 (Brown et al. (2021)). The dotted line represents the chance baseline.

The analysis shows that different scale values are used unequally across the entire acceptability range: there is a tendency for using values towards the extremes of the scale, whereas the mid-zone values, especially the center value '4', are under-used. This suggests that the mid-zone values tend to

[1] I omit the discussion of the authors' Experiment 3 here for reasons of space as it was a replication of Experiment 2 with respect to the target (English) sentences.

mark aggregate rather than individual participants' scores on the respective sentences. This implies that the actual acceptability profile of the sentences in question is characterized by an increased dispersion of values around '4' which should be reflected, in turn, in higher variance. This is indeed the case. We calculated the variance on the filler ratings for their Experiments 1-2 in order to compare it with the one predicted for a uniform 'carrier' distribution in (1). The results are shown in Figure 3.

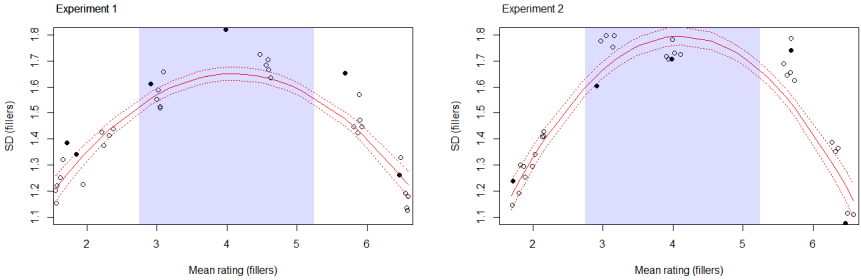

Figure 3: Post-hoc analysis of standard deviations against respective mean acceptability values performed on the 'pre-calibrated' filler sentences in Brown et al.'s Experiments 1 and 2, along with quadratic fit curves and confidence intervals (dotted lines). Filled circles indicate the first exposure to the sentences in respective acceptability classes, non-filled ones indicate subsequent exposures. The purple region demarcates the satiation area according to the authors.

Figure 3 plots the distribution of standard deviations of ratings against their means in the six pre-calibrated types of filler sentences in Brown et al.'s experiments. It shows that standard deviations increase and peak toward the mid-scale point and subside toward the extremes indicating that the subjects' reported mid-zone judgments are subject to the biggest fluctuations precisely at the mid-zone. This trend, observed over the entire scale range, can be described with a quadratic regression model and graphically represented by a parabola with a peak around mid-range (cf. Lipovetsky (2017)). Importantly, as Figure 3 illustrates, the mid-range of the scale characterized by the highest variance represents Brown et al.'s presumed satiation area.

3. Variance in Multiple Wh-Questions

Figure 4 shows standard deviations against mean values in target multiple wh-questions, in Brown et al.'s first two experiments.

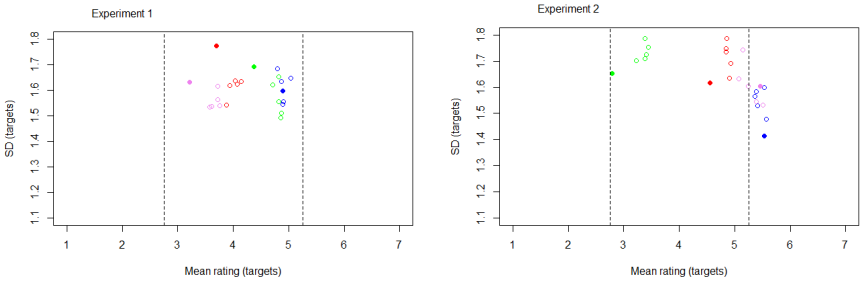

Figure 4: Post-hoc analysis of standard deviations against respective mean acceptability values on the target multiple wh-questions in Brown et al.'s Experiments 1 and 2. Filled circles represent the initial block of trials, non-filled circles represent subsequent trials. Colors indicate different combinations of wh-phrases. Experiment 1 (German): blue: *wer-wen* order, green: *wer-was* order, purple: *wen-wer* order, red: *was-wer* order. Experiment 2 (English): blue: *who-what* order, green: *what-who* order, purple: *which* N_{subj}*-which* N_{obj} order, red: *which* N_{obj}*-which* N_{subj} order. Dotted lines demarcate the satiation area.

The satiation effect in Figure 4 is seen wherever the entries corresponding to non-first trials (empty circles) are located to the right of the filled circle of the same color. The distribution of results in Figure 4 reveals that for those multiple wh-questions whose mean ratings fall outside the (demarcated) mid-range area, as in the blue-colored blocks corresponding to the *which* N_{subj} - *which* N_{obj} order in Experiment 2, these ratings display lesser variance, as opposed to the rest of the wh-questions which receive ratings within the mid-range area and most of which display a satiation effect marked by statistical significance (see Figure 1 for reference): the respective datapoints are located in the upper part of the plot, within the range of about 1.6-1.8. This pattern is more robust in Experiment 2 whereas in Experiment 1 it is seen with respect to the first trials

with somewhat lower variance in non-first trials than expected. Overall, this is in line with the analysis of the fillers above. Brown et al. (2021) establish that both fillers and targets in the mid-zone are subject to satiation: here we establish that they both are subject to higher variance.

4. Variance in Other Studies: Stateva et al. (2019)

The effect of the highest variance in the mid-zone is not artifactual to Brown et al.'s (2021) study. In fact, it is not limited to the satiation studies and even to strictly syntactic sentence acceptability tasks: it also pertains to interpretational aspects of sentence evaluation. Stateva et al. (2019) used a sentence-context verification task to explore approximate numerical boundaries associated with inherently vague quantifiers *few*, *some*, *half*, *most* and *almost all* in four unrelated languages, English, German, French and Slovenian. Subjects in this study were asked to evaluate how well sentences incorporating a quantified expression describe a situation where the actual number of the quantified individuals is given in relation to the total number of relevant individuals in a given context. For instance, the subjects were asked to evaluate a sentence like *Some men utilized an online dating site* in a context like *133 men sought a life partner. 41 of these men utilized an online dating site*, on a 1-to-5 Likert scale. The researchers manipulated the proportion of the quantified expression by varying the second number in the supporting context so as the resulting ratio would be between 1-99% of the total with an increment of 2% resulting in 50 data points per quantifier per subject (the actual contexts were all different across the study and presented in a random order to the subjects). The subjects were not given an opportunity to calculate the ratio explicitly by restricting a time window available for response; only an approximate estimation was possible in that situation. Representative results regarding the numerical boundaries of French quantifier *quelque* ('some') are shown in Figure 5 (see this work for similar results with respect to the other quantifiers in the four languages as mentioned above):

The left plot in Figure 5 shows the acceptability of sentences above the mid-point (=3) until the ratio of quantified individuals reaches about 50% after which it drops below the mid-point and lingers in the lower part of the scale. Notably, dispersion of judgments reflected in standard deviation (shown in green) varies practically as a mirror image of the rating pattern increasing toward the middle and decreasing when the rating approaches the upper scale extreme. In contrast, in the lower portion of the scale variance remains low

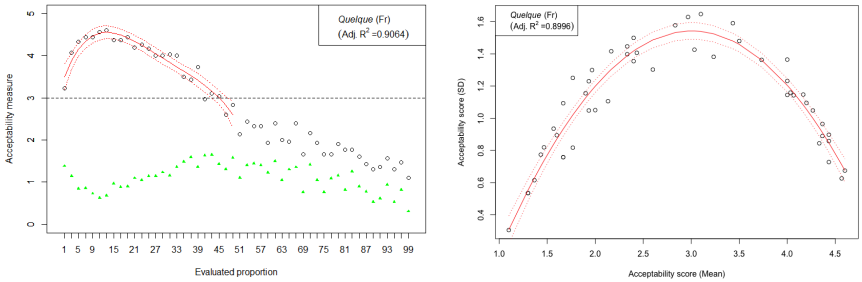

Figure 5: On the left: Mean acceptability scores on French sentences including the quantifier *quelque* ('some') against contexts with the percentage ratio of quantified individuals varying between 1-99% (standard deviations are in green), together with fit curves and confidence intervals predicting acceptability in the upper half of the Likert scale. On the right: standard deviations on mean acceptability values together with the fit curve and confidence intervals. Source: Stateva et al. (2019).

as well. The right plot makes that pattern more transparent. The latter is once again describable by a quadratic fit curve whose peak corresponding to the central point on the mean acceptability scale (in this case 3) is in the ballpark of the theoretical value of standard deviation for a discrete uniform distribution, that is $\sigma = \sqrt{\frac{(5-1+1)^2-1}{12}} \approx 1.41$. This means that for sentences whose mean acceptability is around the center of the scale, subjects' actual individual ratings were likely to be distributed more or less uniformly across the entire scale, with an approximately equal probability of occurrence (see Section 2).

An important feature unifying the two studies is the use of the set of stimuli that span the entire range of accepability in more or less discrete and evenly distributed increments. This kind of methodology allows one to monitor the dispersion of ratings at different portions of the scale in a more balanced and controlled manner than in cases when only a limited set of scale values is used.

5. The Uncertainty Factor

Let us now elaborate somewhat on the pattern of variance in Figure 3. The highest variance around the mid-scale point, 4 in our case, implies that for a sentence of type S^4 that has a mean acceptability status of 4 the actual ratings given by participants to this sentence will very likely be dispersed on both sides from the center, so that sentence will very likely receive actual ratings '3' as well as '5', or '2' as well or '6', or even '1' as well as '7'. In contrast, for a sentence S^2 the dispersion of ratings around the actual '2' value will presumably be much smaller, so, e.g. rating '7' is much less likely to appear. For sentences in S^1 and S^7 the respective scale value will presumably be the likeliest of all.

The gradient increase of variance toward the mid-zone maximum reflects a property of the population sample that we may refer to as *an uncertainty factor*. The dispersion of ratings around center value 4 suggests that this area of the scale is the region of highest speakers' uncertainty towards the actual acceptability status of the respective sentences.

There are two possible scenarios potentially able to account for this pattern of the scale use. One is that our baseline assumption in the beginning of Section 2 regarding speakers' a priori equal access to all scale values is wrong: speakers are, for some reason, not able to consciously access their judgment percepts that come from the mid-scale judgment zone as easily or straightforwardly as those associated with extreme scale values. The reason for that may be an increased number of factors that enter into determining the acceptability status, over and above its grammatical derivation, compared to the percepts that arise close to the scale extremes.

The other possibility (which does not necessarily exclude the first one) is that the highest uncertainty in the mid-scale is a task effect. In laying out the sentence acceptability task for their experimental subjects, Brown et al. (2021) mark all scale values explicitly using the same predicate meaning "well-formed", ranging from "not at all well formed" to "completely well-formed" (pp. 9,15). These definitions, presumably, prompt the subjects to treat the scale as a scale of degrees of well-formedness, appropriate for this type of study. Borrowing terminology from the psychometric literature, let us refer to this intended kind of scale as a *unipolar* scale. Unipolar scales have two characteristic properties: they (i) measure the degree of presence of some property and (ii) have no natural midpoint when going to a (single) extreme.

The alternative, *bipolar* or *bipolar reversal* scale (i) measures evaluation between two extreme opposites and (ii) provides a natural mid-point (see e.g. Shulman (1973) and Wang and Krosnick (2020) for details). In sociological surveys, the mid-point is often used to indicate some sort of neutral perspective such as "neither agree nor disagree" or "I don't know". While Brown et al.'s consistent use of a well-formedness predicate in their explicit definitions of scale points is characteristic of a unipolar scale, the authors at the same time define the scale's median value '4' as "kann man nicht zuordnen" / "cannot be classified as well-formed or ill-formed" (p. 9). It is possible that introduction of this explicit midpoint triggers the subjects' (unintended) perception of the entire scale as a bipolar reversal scale instead.

In fact, there is some indication that this is indeed so. In addition to the increased variance, the distribution of ratings in Brown et al.'s (2021) study manifests the so called *extremity response bias* known as the tendency to (over)use the extreme values of the scale. Indeed we see this in Figure 2 from Experiments 1 and 2 (recall that the filler sentences in these experiments were balanced across the entire acceptability range in the authors' experiments, thus all Likert scale values are expected to be used a priori). This tendency to use the extreme value is characteristic of bipolar reversal scales (see Shulman (1973) for discussion).

It should be noted that the explicit definition of the median value '4' as a scale-neutral midpoint in Brown et al.'s study may be responsible for its reduced usage by the subjects as depicted in Figure 2, but it alone does not explain the monotonic character of the pattern of variance peaking at the mid-zone manifested in Figure 3. The increased uncertainty seems to be a property of the mid-zone, not of the center-scale value alone.

The experimental design in Stateva et al.'s (2019) study, on the other hand, does not explicitly specify the center-scale value. Rather, the authors define a scale by only specifying extreme points (1 and 5 in their study) as "not well-formed" and "well-formed". In that study, therefore, the mid-point of the scale does not have an exclusive status. Nevertheless, the mid-range variance effect obtains in that study as well. Consequently, explicitly defining the mid-point cannot be the only reason for the increased variance.

What all this suggests in terms of a sentence acceptability task is that, when asked to evaluate acceptability on a discrete scale of gradience, speakers might tend to interpret the task as a binary task instead, implicitly solving it not in graded terms, but in categorical ones. In essence, they might be *dividing*

the scale into two halves corresponding to the opposite categorical values (acceptable, unacceptable). Under this scenario, speakers might be using the non-extreme values of the scale (other than the midpoint) in order not to express the gradient nature of acceptability, as intended by experimental design, but rather, to express a degree of subjective uncertainty toward one or the other categorical status of the sentence. This seems to be a promising direction to explore in further studies of this kind.

6. Variance and Satiation

Recall that Brown et al. (2021) suggest to see satiation as a 'mere exposure' effect in the sense of Zajonc (1968): repeated exposure of the individual to a stimulus object such as nonsense words or syllables enhances their attitude toward it without any further reinforcement. Zajonc's relevant measure of improvement is actually on a 'good-bad' scale in the sense of people's sentiment or stance toward the stimulus which is not quite the same as a percept of sentence acceptability status. Regardless of this concern, a 'mere exposure' kind of explanation still leaves a question why repeated exposure targets precisely the judgments in the mid-zone scale. The observed pattern of speakers' variance might provide the beginning of a clue: repeated exposure targets the most 'uncertain' or volatile part of the acceptability scale.

If the variance in ratings is indeed due to speakers' subjective uncertainty about the 'unacceptable' and 'acceptable' halves of the scale, then satiating at least in some cases means switching simply from the former to the latter half. It is also tempting to relate the speakers' scale split strategy in the context of the binary character of the sentence's grammaticality status (Schütze (1996)). Making these hypotheses more precise necessitates spelling out a model of relatedness between the graded and binary components of the speakers' sentence evaluation mechanism, which cannot be addressed in this short note (but see, e.g. Bader and Häussler (2010), for a discussion along these lines). If this line of inquiry is correct, it implies that satiation may involve a categorical decision at some point. Note that this explanation differs from the Brown et al.'s (2021) own perspective on the nature of satiation who note that "the size of initial rise of ratings is not very dramatic [...]; they increase only by roughly .5 on a 7-point scale, meaning that repeated exposure does not change the quality of the judgment." (p. 22). But if specific ratings for at least some sentences actually cross the mid-point on repeated exposure then the change

should be regarded as qualitative, no matter what the actual average values turn out to be.

To conclude, Brown et al. (2021) open an important venue of investigation into speakers' patterns of use of the acceptability scale. This kind of investigation of the 'global' use of the scale is only possible if a balanced set of stimuli spanning the entire range of acceptability is considered. I have sketched a way to explore this venue via tracking the variance in the ratings at the population level; variance is viewed here as a marker of speakers' rating uncertainty. We have seen that the mid-scale zone is a zone of both satiation and the higher uncertainty in speakers' ratings. Although this does not imply direct causality, it is conceivable that the same cognitive factor(s) contribute to each phenomenon. Their proper discovery still awaits its place and time.

References

Bader, Markus and Jana Häussler (2010): 'Toward a model of grammaticality judgments', *Journal of Linguistics* **46**(2), 273–330.

Brown, J. M. M., Gisbert Fanselow, Rebecca Hall and Reinhold Kliegl (2021): 'Middle ratings rise regardless of grammatical construction: Testing syntactic variability in a repeated exposure paradigm', *PLOS ONE* **16**(5), e0251280. **URL:** *https://dx.plos.org/10.1371/journal.pone.0251280*

Chomsky, Noam (1973): Conditions on transformations. *In:* S. R. Anderson and P. Kiparsky, eds., *A festschrift for Morris Halle*. Holt, Rinehart & Winston, New York, pp. 232–286.

Featherston, Sam (2009): A scale for measuring well-formedness: Why syntax needs boiling and freezing points. *In:* S. Featherston and S. Winkler, eds., *The Fruits of Empirical Linguistics*. Vol. 1, de Gruyter, Berlin, pp. 47–74.

Gerbrich, Hannah, Vivian Schreier and Sam Featherston (2019): Standard items for English judgement studies: syntax and semantics. *In:* S. Featherston, R. Hörnig, S. von Wietersheim and S. Winkler, eds., *Experiments in Focus: Information Structure and Semantic Processing*. de Gruyter, Berlin, pp. 305–328.

Lipovetsky, Stan (2017): 'Factor analysis by limited scales: which factors to analyze?', *Journal of Modern Applied Statistical Methods* **16**(1), 233–245.

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

Shulman, Art (1973): 'A Comparison of Two Scales on Extremity Response Bias', *Public Opinion Quarterly* **37**(3), 407.

Snyder, William (2000): 'An experimental investigation of syntactic satiation effects', *Linguistic Inquiry* **31**, 575–582.

Stateva, Penka, Arthur Stepanov, Viviane Déprez, Ludivine Emma Dupuy and Anne Colette Reboul (2019): 'Cross-Linguistic Variation in the Meaning of Quantifiers: Implications for Pragmatic Enrichment', *Frontiers in Psychology* **10**, 957.
URL: *https://www.frontiersin.org/article/10.3389/fpsyg.2019.00957/full*

Wang, Rui and Jon A. Krosnick (2020): 'Middle alternatives and measurement validity: a recommendation for survey researchers', *International Journal of Social Research Methodology* **23**(2), 169–184.

Zajonc, Robert B. (1968): 'Attitudinal effects of mere exposure', *Journal of Personality and Social Psychology* **9**(2, Pt.2), 1–27.