Satiation effects generalize across island types

Jiayi Lu[1], **Nicholas Wright**[2], **Judith Degen**[1]
[1]{jiayi.lu, jdegen}@stanford.edu
Department of Linguistics, Stanford University
Stanford, CA 94305, USA
[2]njwright01@wm.edu
Department of English, College of William & Mary
Williamsburg, VA 23185, USA

Abstract

A recent proposal of syntactic satiation claims that it is driven by adaptation: comprehenders track and update their beliefs about the probability of observing certain sentences, leading to subsequent increases in the acceptability of those sentences. This leaves open what the representational targets of satiation are, that is: what is the tracked information that belief update is based on? In two acceptability judgment experiments, we show that exposure to one type of island violation can lead to the satiation of another island type, suggesting that island type-general representations are tracked by comprehenders in addition to island type-specific representations. The same experimental paradigm can be used for further exploration of the representational targets of satiation.

Keywords: psycholinguistics; island effect; adaptation; satiation; acceptability judgments

Introduction

In experimental syntax, a commonly employed measure of a sentence's grammaticality is acceptability judgments (Schütze, 1996). Sentence acceptability is affected by a widely observed phenomenon called **the satiation effect**: after repeated exposure to unacceptable sentences, people usually find these sentences increasingly acceptable (Brown, Fanselow, Hall, & Kliegl, 2021; Chaves & Dery, 2019; Francom, 2009; Goodall, 2011; Hiramatsu, 2001; Lu, Lassiter, & Degen, 2021; Snyder, 2000, 2021). Despite the abundance of studies on satiation, an important question is left mostly unanswered: when sentences satiate, what are the representational targets of satiation? That is, which latent or overt features of the linguistic signal do comprehenders track and adapt to? In this study, we take a first step towards addressing this question by studying the generalization of satiation across two different island-violation constructions: subject islands, and *whether*-islands.

Satiation mechanism

There are various proposals for the mechanism of satiation.[1] Under the memory-bottleneck account, satiation is the result of processing facilitation of memory-demanding sentences (Francom, 2009; Hofmeister & Sag, 2010). Under the priming account, satiation is an instance of structural priming (Francom, 2009; Do & Kaiser, 2017). Building on the priming-based account, a recently proposed adaptation account construes the satiation effect as an instance of adaptation (Lu et al., 2021): throughout exposure, comprehenders update their beliefs about the occurrence probability of certain linguistic forms (Kleinschmidt & Jaeger, 2011; Fine, Jaeger, Farmer, & Qian, 2013; Schuster & Degen, 2020); the more expected a form is, the more acceptable it is. In this work, we do not intend to argue for or against any of the aforementioned accounts of satiation. Instead, we aim to answer a question that affects all these accounts alike: which features are the representational targets of satiation? In the remainder of this section, we motivate this research question within the adaptation account of satiation (Lu et al., 2021) for illustrative purposes.

Under the adaptation account of satiation, comprehenders have and maintain uncertainty about the speaker's generative language model θ, which assigns contextual probabilities to the production of various utterances u with underlying linguistic representation LR (e.g., syntactic structures, lexical items etc.). Fig. 1 shows the causal model of utterance production assumed by Lu et al. (2021).

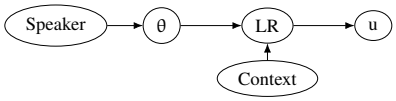

Figure 1: Causal model of utterance production (Lu et al., 2021)

When the comprehender observes an utterance by a speaker (s), the probability of each generative language model $\theta \in \Theta$ given s is updated via Bayesian belief updating. Assuming that context is fixed and given, and that

[1]Sprouse (2009) claims that there is no genuine satiation, and that increased acceptability after exposure is driven by an experimental confound: the result of an "equalization strategy", participants tend to balance their positive and negative responses when answering surveys. Therefore, in an acceptability judgment experiment with a high number of ungrammatical stimuli, participants give increasingly higher ratings to balance the overall responses. However, we will see clear evidence against this claim in the reported experimental results (see also Francom, 2009; Goodall, 2011; Crawford, 2012; Lu et al., 2021).

the comprehender recovers a set of linguistic representations (r) from the utterance (u), we have the following equation by Bayes rule:

$$p(\theta \mid r,s) \propto p(r \mid \theta)\, p(\theta \mid s) \tag{1}$$

From (1), we know that for any two generative models θ_i and θ_j where $p(r \mid \theta_i) > p(r \mid \theta_j)$, the posterior probability of θ_i increases compared to θ_j upon observing r :

$$\frac{p(\theta_i \mid r,s)}{p(\theta_j \mid r,s)} > \frac{p(\theta_i \mid s)}{p(\theta_j \mid s)} \tag{2}$$

As a result of this belief updating process, the comprehender's expectation for the linguistic representation r, expressed in (3) as a marginal probability over all possible generative models, should increase with exposure to r.

$$p(r \mid s) = \sum_{\theta \in \Theta} p(r \mid \theta) p(\theta \mid s) \tag{3}$$

In Lu et al. (2021)'s formulation, r in the above equations is assumed to be specific syntactic structures. That is, it is assumed that comprehenders only track and update their beliefs about different syntactic structures during adaptation. Increased expectation for a syntactic structure r would lead to increased expectation for utterances with that particular structure, and eventually result in increased acceptability ratings for such utterances.

The choice to assume syntactic structures as the representational target of satiation is in line with the rhetoric adopted in previous literature on satiation: past studies (Snyder, 2000; Sprouse, 2009; Chaves & Dery, 2019, *inter alia*) often discuss the satiation of various syntactic constructions (e.g., satiation of *whether*-island sentences, complex-NP island sentences, etc.). However, there is no evidence suggesting that the type of linguistic representation tracked by participants during satiation (the 'LR' node in Fig. 1, and r in equations 1, 2, and 3) needs to be particular syntactic structures or constructions. While past studies reporting satiation effects observed that exposure to one syntactic structure leads to acceptability increase in sentences of the same structure, such observations only suggest that the representational targets of satiation can be any linguistic representation that can be abstracted away from the satiated sentences (e.g. filler-gap dependencies, embedded clauses, etc.). These features may or may not be shared with other syntactic structures. In the following section, we shall see how we can pinpoint the representational targets of satiation in a generalization paradigm.

The generalization paradigm

One way to pinpoint the representational targets of the satiation effect is to use a generalization paradigm, i.e., to test whether satiation of one sentence type generalizes to others. If repeated exposure to sentence type A increases the perceived acceptability of not only A but also another sentence type B, we could conclude that comprehenders adapt to linguistic representations that are shared across sentence types A and B (see Bock (1989); Bott & Chemla (2016), among others, for examples of this type of reasoning). For example, both relative clauses and wh-question sentences contain filler-gap dependencies. If comprehenders track and update their beliefs about the probability of observing filler-gap dependencies, exposure to sentences with relative clauses should lead to increased acceptability of wh-question sentences. In contrast, if comprehenders track and update their beliefs about the distribution of specific syntactic tree structures, exposure to sentences with relative clauses should not affect the acceptability of wh-question sentences, since the two sentence types involve different syntactic structures.

To summarize, different hypotheses about the representational target of satiation make different predictions about whether satiation should generalize between sentence types. Fig. 2 sketches the hypothesis space in the context of the satiation of sentences with island violations, a widely studied class of syntactic constraints in the past literature on satiation (Snyder, 2000; Francom, 2009; Chaves & Dery, 2019, *inter alia*). In Fig. 2, colored boxes contain sets of sentence types that should exhibit satiation generalization under different assumptions of the representational targets of satiation. Exposure to a sentence type inside the box should lead to an acceptability increase for any other sentence type inside the same box. For example, if the representational target of satiation is a specific island-violation type (A), exposure to such sentences should only lead to satiation of sentences of the same type of island-violation sentence. If, instead, the representational target is the violation of an island constraint in general, exposure to one type of island-violation sentence should also lead to satiation of sentences with other types of island violations (B).

In this study, we employed the generalization paradigm to examine the generalization of satiation across subject island and *whether*-island violations. The former refers to ungrammatical syntactic movements from within complex subjects, and the latter refers to ungrammatical syntactic movements from *whether*-clauses (Ross, 1967). Example sentences are shown below. These two constructions have both been previously shown to satiate (Snyder, 2000; Chaves & Dery, 2019; Lu et al., 2021).

(1) Subject island violation
*Who$_i$ did Mary think the brother of t_i came to the party?

(2) *Whether*-island violation
*What$_i$ did Mary wonder whether John ate t_i?

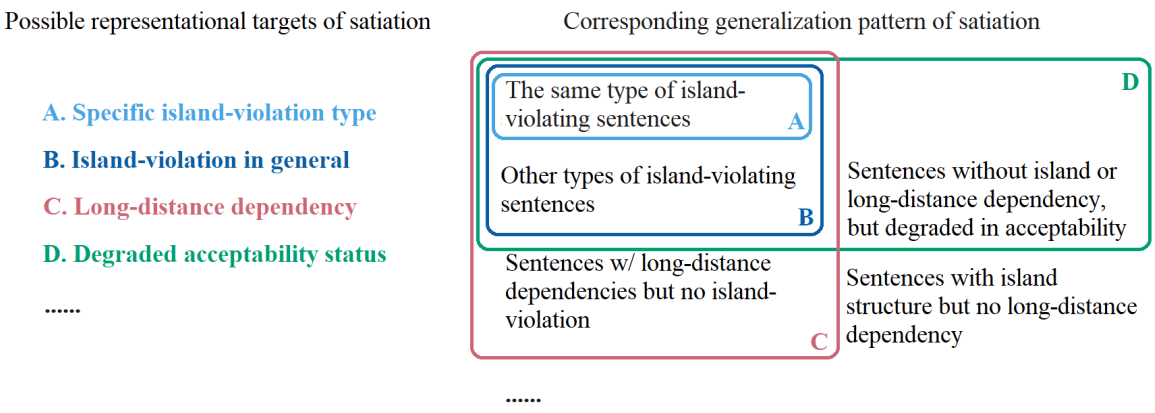

Figure 2: Possible representational targets of island satiation and generalization patterns. The lists of representational targets and sentence types are non-exhaustive, indicated by '. . . '.

By testing for generalization of satiation effects from one island-violation type to another, we aim to tease apart two hypotheses: one possibility is that comprehenders track the specific type of island violation during satiation (A in Fig. 2). Another possibility is that comprehenders track features that are not linked to any specific island violation type (e.g., island violations in general (B), the existence of a filler-gap dependency (C), degraded acceptability in general (D)). We leave distinguishing between B, C, and D for future work.

If comprehenders track only island type-specific representations, no generalization across island types is expected. If comprehenders track only island type-general information, the magnitude of satiation generalization across island types should be the equal to satiation resulting from exposure to the same island type. Finally, if comprehenders track both island type-specific and island type-general information, satiation generalization across island types should happen, but to a lesser extent than satiation resulting from exposure to the same island type.

Experiment 1

In Exp. 1, we tested whether satiation to *whether*-island sentences generalizes to subject island sentences using an exposure-test paradigm acceptability judgment experiment. If satiation generalizes across the two island types, the acceptability of subject island sentences should increase after exposure to *whether*-island sentences.[2]

Method

Participants We recruited 973 participants on Prolific, with 52 excluded because they met at least one of the following exclusion criteria: their primary language was not English, the 95% confidence intervals of responses to grammatical and ungrammatical fillers overlapped, or they answered a practice trial incorrectly more than once.

Materials and procedure Participants were asked to read sentences and give acceptability ratings on a sliding-scale with the lower end labeled 'completely unacceptable' and the higher end labeled 'completely acceptable'. The responses were recorded as numeric values between 0 and 1, with 0 representing the lower end and 1 representing the higher end of the scale.

For each trial, the participants saw a target sentence preceded by a context sentence. The experiment consisted of 44 trials divided into an exposure phase containing 12 exposure sentences and 12 filler sentences, and a test phase containing 10 test sentences and 10 filler sentences. The participants were randomly assigned to three exposure groups: the within-category group, the between-category group, and the control group. In the exposure phase, the within-category group saw subject island sentences, the between-category group saw *whether*-island sentences, and the control group participants saw polar questions. We used polar questions as a control because they are interrogative sentences like the other exposure sentences, but they do not possess most of the linguistic representations that are shared between *whether*-island and subject island sentences (e.g. island violations, long distance dependencies, degraded acceptability, etc.). In the test phase, all participants saw subject island sentences as test sentences. Example stimuli are shown in Table 1.

Among the three exposure groups, the within-category group served as a positive control where maximal satiation was expected, since the participants were exposed to the same sentence type in both the exposure phase and the test phase. The control group served as a negative control where no satiation was expected, given that

[2]Pre-registrations are available at `osf.io/hwk7g`. Experimental materials, data, and analysis scripts are available at `github.com/wright-nicholas/satiation-generalization`.

Table 1: Example stimuli.

Condition	Context	Target
Polar question	Jack thinks that Mary spilled a bottle of water.	Did Jack think that Mary spilled a bottle of water?
Subject island	Jack thinks that a bottle of water was spilled by Mary.	What does Jack think that a bottle of was spilled by Mary?
Whether-island	Jack wonders whether Mary spilled a bottle of water.	What does Jack wonder whether Mary spilled?
Grammatical filler	The journalist thought that the politician wrote a book.	What did the journalist think that the politician wrote?
Ungrammatical filler	The priest of the local church saw a man sleeping under the bridge.	What bridge the under saw church local the of did priest the?

the polar questions in the exposure phase do not share even island type-general representations (e.g. the existence of island-violation regardless of type, the existence of a long distance dependency, etc.) with the subject island sentences in the test phase, except for the interrogative force. Generalization of satiation from *whether*-islands to subject islands is detected as a positive difference between the between-category exposure group and the negative control group in the test phase. Satiation to island type-specific representations is detected as a negative difference between the between-category group and the within-category group in the test phase.

Results and discussion

Mean acceptability ratings of the test sentences by exposure group are shown in Fig. 3. The mean acceptability ratings of all exposure groups are plotted against trial number in Fig. 4.

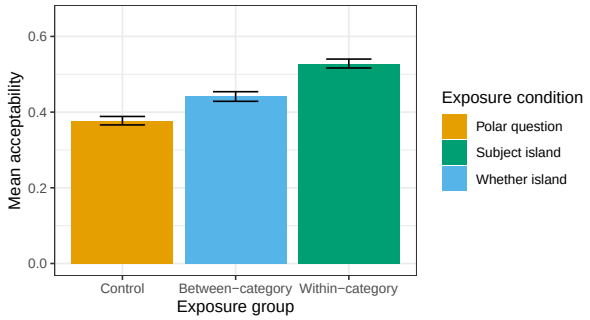

Figure 3: Test phase mean acceptability by exposure group (Exp. 1).

A linear mixed-effects model predicted acceptability ratings from dummy-coded fixed effects of experimental phase (reference level: test), exposure group (reference level: between-category), and their interaction. The model included random by-participant and by-item intercepts, by-participant slopes for experimental phase, and by-item slopes for both fixed effects and their interaction.

There was a significant exposure group effect in the test phase: compared to the between-category group ratings, the control group ratings were significantly lower (β=-.064, SE=.019, t=-3.35, p<.001), and the within-category group ratings were significantly higher (β=.085, SE=.020, t=4.25, p<.001).

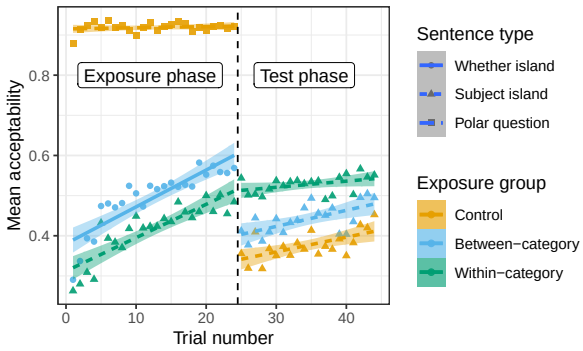

Figure 4: Mean acceptability ratings over time (Exp. 1).

Our design assumed that the control group's exposure to polar questions has little influence on their ratings for the subject-island sentences in the test phase. We verified this assumption by comparing the control group's responses for the first six trials of the test phase with the within-category group's responses for the first six trials of the exposure phase[3]. Using a linear mixed-effect model with a fixed effect of exposure group (control vs. within-category) predicting the acceptability ratings from the first six trials of both phases, we found no significant difference between the two groups (β=-.006, SE=.022, t=-.027, p=0.79).

Compared to the control group, the between-category group who were previously exposed to *whether*-island sentences rated the subject island sentences significantly higher. This suggests that the satiation to *whether*-island sentences generalized to subject island sentences, which supports the hypothesis that participants track and adapt to island type-general representations during satiation to island-violating sentences. Furthermore, the within-category group rated the test sentences signifi-

[3]We used the first six trials to represent the beginning of an experimental phase, following Lu et al. (2021).

cantly higher than the between-category group. This shows that the amount of between-category generalization is smaller than within-category satiation, which suggests that participants also track and adapt to island type-specific representations.

Finally, to test whether the acceptability increases in Exp. 1 indeed reflect satiation rather than the result of an equalization response strategy whereby participants try to balance their high and low responses (Sprouse, 2009), we show the cumulative mean of ratings on each trial in Fig. 5. If only the equalization response strategy is at play and no satiation took place, the cumulative mean should drift towards the midpoint of the scale. However, we see that the cumulative mean crosses the midpoint (0.5, marked by the dashed line) during the exposure phase. Thus, the changes in sentence acceptability in Exp. 1 cannot simply be explained as a task artifact.

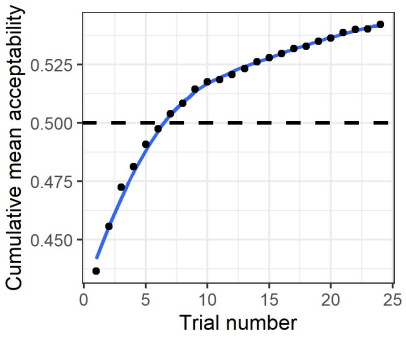

Figure 5: Cumulative mean acceptability ratings during exposure phase (Exp. 1). Dashed line represents the midpoint of the scale (0.5).

Experiment 2

In Exp. 1, we observed that satiation to *whether*-island sentences can generalize to subject island sentences. If this generalization effect is driven by participants adapting to island type-general representations, we expect this satiation generalization effect to also hold in the reverse direction: satiation to subject island sentences should generalize to *whether*-island sentences. We tested this prediction in Exp. 2.

Method

Participants A total of 968 participants were recruited on Prolific, with 23 excluded based on the same exclusion criteria as in Exp. 1.

Materials and procedures Exp. 2 used the same set of stimuli as Exp. 1, examples of which are shown in Table 1. The same experimental design was used, except that in Exp. 2 the within-category group participants saw *whether*-island sentences as exposure sentences, the between-category group participants saw subject island sentences as exposure sentences, and all test sentences were *whether*-island sentences.

Results and discussion

Mean acceptability ratings of the test sentences by exposure group are shown in Fig. 3. The mean acceptability ratings of all exposure groups are plotted against trial number in Fig. 4.

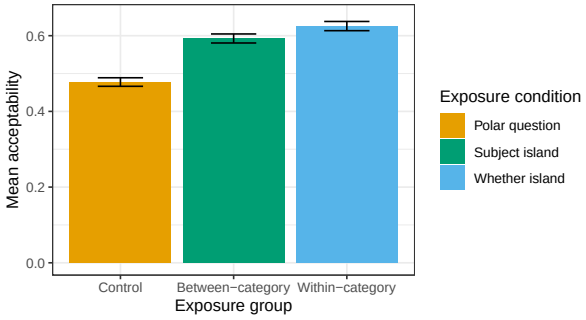

Figure 6: Test phase mean acceptability by exposure group (Exp. 2).

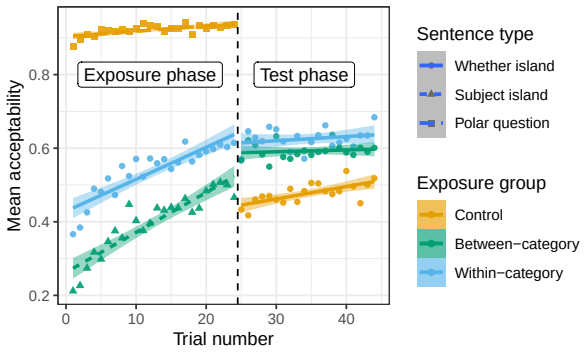

Figure 7: Mean acceptability ratings over time (Exp. 2).

A linear mixed-effects model predicted acceptability ratings from dummy-coded fixed effects of experimental phase (reference level: test), exposure group (reference level: between-category), and their interaction. The model included random by-participant and by-item intercepts, by-participant slopes for experimental phase, and by-item slopes for both fixed effects and their interaction.

There was a significant exposure group effect in the test phase: compared to the between-category group ratings, the control group ratings were significantly lower (β=-.10, SE=.021, t=-4.74, $p<.001$), and the within-category group ratings were significantly higher (β=.062, SE=.020, t=3.04, $p<.01$).

Once again, our assumption that the control group provides a negative control with minimal satiation is confirmed by comparing the control group's responses on the first six trials of the test phase with the within-category group's responses on the first six trials of the exposure phase. Using a linear mixed-effects model with the fixed effect of exposure group (control vs. within-category) predicting the acceptability ratings from the first six trials of both phases, there was no significant different between the two groups (β=-.025, SE=.021, t=-1.19, p=0.24).

Compared to the control group, the between-category group who were exposed to subject island sentences in the exposure phase rated the *whether*-island sentences in the test phase significantly higher. This suggests that the satiation to subject island sentences generalized to *whether*-island sentences, which supports the hypothesis that participants track and adapt to island type-general representations during satiation to island-violating sentences. Furthermore, the within-category group rated the test sentences significantly higher than the between-category group. This shows that the amount of between-category generalization is smaller than within-category satiation, suggesting that island type-specific representations are also tracked by participants.

Finally, to show that the acceptability increases in Exp. 2 are indeed satiation rather than the result of an equalization response strategy (Sprouse, 2009), we plotted the cumulative mean of acceptability ratings on each trial in Fig. 8. The cumulative mean crosses the midpoint of the scale (0.5, marked by the dashed line in Fig. 8) during the test phase, again ruling out the task artifact explanation of acceptability ratings increases.

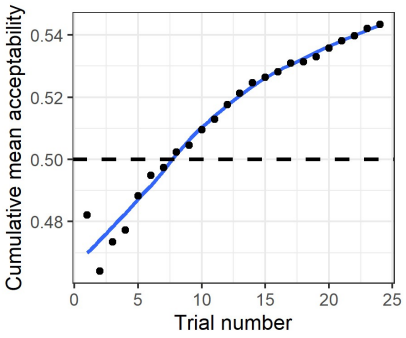

Figure 8: Cumulative average acceptability ratings during exposure phase (Exp. 2). Dashed line represents the midpoint of the scale (0.5).

General Discussion

In this study, two acceptability judgment experiments tested whether comprehenders track and adapt to island type-specific representations, island type-general representations, or both, during satiation of sentences containing subject island and *whether*-island violations. The results suggest that comprehenders track both island type-specific and type-general representations.

In both experiments, we saw that the between-category exposure group rated the test sentences as significantly more acceptable than the control group, and significantly less acceptable than the within-category group. Assuming that the control group established a negative baseline where the exposure phase leads to no satiation generalization on test phase sentences (as confirmed by the lack of significant difference between the beginning of the control group test phase and the beginning of the within-category group exposure phase in both experiments), the contrast between the between-category and the control groups suggests that exposure to one island type leads to satiation in another island type. The contrast between the within-category and the between-category groups suggests that satiation generalization is smaller in magnitude than maximal satiation achieved through consistent within-category exposure.

How do these results inform the question regarding the representational targets of satiation? Recall the adaptation account of satiation. Comprehenders track and update their beliefs about the probabilistic distributions of linguistic representations. The increased expectation for certain linguistic representations leads to increased acceptability of utterances that embody such representations, yielding the satiation effect. The observation that exposure to one island type leads to satiation in another island type suggests that comprehenders track linguistic representations that are shared across the two island types. However, these island type-general representations are not the only type of information tracked: if they were, there should be no difference between the magnitude of between-category generalization and within-category satiation, contrary to observation. Therefore, our results suggest that the representational targets of island satiation include both island type-specific representations and island type-general representations.

There are various possible island type-specific representations (e.g., embedded clause types, particular syntactic structures) and island type-general representations (e.g., filler-gap dependencies, the degraded acceptability status) that comprehenders could track. When subject island or *whether*-island sentences satiate, which particular island type-specific and island type-general representations do participants track and adapt to? Our current results cannot tease apart these more fine-grained possibilities. Future studies could use the same generalization paradigm introduced here to test the various possible representational targets of satiation in the hypothesis space.

References

Bock, K. (1989). Closed-class immanence in sentence production. *Cognition*, *31*(2), 163–186.

Bott, L., & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, *91*, 117–140.

Brown, J. M. M., Fanselow, G., Hall, R., & Kliegl, R. (2021). Middle ratings rise regardless of grammatical construction: testing syntactic variability in a new repeated exposure paradigm. *PLoS ONE*, *16*(5), e0251280.

Chaves, R. P., & Dery, J. E. (2019). Frequency effects in subject islands. *Journal of linguistics*, *55*(3), 475–521.

Crawford, J. (2012). Using syntactic satiation to investigate subject islands. In *Proceedings of the 29th west coast conference on formal linguistics* (pp. 38–45).

Do, M. L., & Kaiser, E. (2017). The relationship between syntactic satiation and syntactic priming: A first look. *Frontiers in psychology*, *8*, 18–51.

Fine, A., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS one*, *8*(10), e77661.

Francom, J. C. (2009). *Experimental syntax: Exploring the effect of repeated exposure to anomalous syntactic structure–evidence from rating and reading tasks*. Unpublished doctoral dissertation, U. of Arizona.

Goodall, G. (2011). Syntactic satiation and the inversion effect in english and spanish wh-questions. *Syntax*, *14*(1), 29–47.

Hiramatsu, K. (2001). *Accessing linguistic competence: Evidence from children's and adults' acceptability judgments.* Unpublished doctoral dissertation, University of Connecticut.

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, *86*(2), 366–415.

Kleinschmidt, D., & Jaeger, T. F. (2011). A bayesian belief updating model of phonetic recalibration and selective adaptation. In *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics* (pp. 10–19).

Lu, J., Lassiter, D., & Degen, J. (2021). Syntactic satiation is driven by speaker-specific adaptation. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Ross, J. R. (1967). *Constraints on variables in syntax*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge.

Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, *203*, 104285.

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

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *LI*, *31*(3), 575–582.

Snyder, W. (2021). Satiation. In *The cambridge handbook of experimental syntax* (pp. 154–180). Cambridge University Press.

Sprouse, J. (2009). Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*, *40*(2), 329–341.