languages

MDPI

Article

On the Nature of Syntactic Satiation

William Snyder

Department of Linguistics, University of Connecticut, Storrs, CT 06269-1145, USA; william.snyder@uconn.edu

Abstract: In syntactic satiation, a linguist initially judges a sentence type to be unacceptable but begins to accept it after judging multiple examples over time. When William Snyder first brought this phenomenon to the attention of linguists, he proposed satiation as a data source for linguistic theory and showed it can be induced experimentally. Here, three new studies indicate (i) satiation is restricted to a small, stable set of sentence types; (ii) after satiation on one sentence type (e.g., *wh*-movement across … *wonder whether* … or … *believe the claim* …), acceptability sometimes increases for distinct but syntactically related sentence types (… *wonder why* …; … *accept the idea* …); (iii) for sentence types susceptible to satiation, the difficulty of inducing it (e.g., number of exposures required) varies systematically; and (iv) much as satiation in linguists persists over time, experimentally induced satiation can persist for at least four weeks. These findings suggest a role for satiation in determining whether the perceived unacceptability of two sentence types has a common source.

Keywords: syntactic satiation; linguistic judgments; island effects; experimental syntax

check for updates

Citation: Snyder, William. 2022. On the Nature of Syntactic Satiation. *Languages* 7: 38. https://doi.org/10.3390/languages7010038

Academic Editors: Anne Mette Nyvad and Ken Ramshøj Christensen

Received: 9 August 2021
Accepted: 1 February 2022
Published: 17 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

1.1. Overview of the Project

In generative linguistics, information about a person's mental grammar comes primarily from that person's judgments of acceptability: certain combinations of form and meaning are fully acceptable, while others are not. The standard idealization is that any given native-speaker consultant who is asked, on different occasions, to judge the same <form, meaning> pair will provide the same judgment on each occasion.

A systematic exception is presented by "satiation" effects: for certain initially unacceptable sentence types, after a linguist has judged multiple examples over a period of time, the perceived acceptability increases. Satiation calls out for investigation, not only because linguistic theories need to take account of its possible effects on the data they use but also because it may provide new insights into the basic phenomena that linguistic theories are meant to explain.

This article performs some of the necessary groundwork for linguistic investigation of satiation by providing evidence for the following points:

(1) a. While satiation effects were first noticed informally among professional linguists, they can also be induced in non-linguists, under controlled conditions in the laboratory;
b. Satiation effects induced in the laboratory are replicable, in the sense that the set of sentence types that potentially satiate is consistent across studies (and for the majority of sentence types, satiation does not occur);
c. Satiation effects for different types of "satiable" grammatical violation have different signatures (e.g., in the number of exposures typically needed before satiation occurs and in the typical percentage of experimental participants whose judgment changes).

The objective will be to show that investigation of satiation can broaden the range of empirical phenomena (and, thus, sources of data) bearing on key linguistic issues, including

in particular a whole range of issues concerning the nature and status of acceptability judgments.

1.2. Overview of Satiation

Building on earlier unpublished work by Karin Stromswold, Snyder (2000) published a squib drawing linguists' attention to the phenomenon Stromswold had termed "syntactic satiation". For a highly circumscribed set of sentence types, linguists sometimes experience a "shift" in their native-speaker acceptability judgments. The paradigm case is (2).

(2) Who does John wonder whether Mary likes?
(*Answer: He wonders whether she likes Pat*.)

On first exposure to an example like (2), a linguist may have thought it sounded starkly unacceptable. Yet, by the time the linguist was teaching an introductory course on syntax and presented an example like (2) to the students, the perception of grammatical impossibility may have been weaker, or even absent altogether. If so, the linguist had experienced satiation on that sentence type.

The description in (Snyder 2000), together with anecdotal reports and personal experience, motivates the characterization in (3).

(3) Characteristics of syntactic satiation when experienced by linguists:
   a. Lexical Generality: Satiation operates at the level of a grammatical structure. The increased acceptability of the structure is general, extending beyond the specific sentences that caused satiation—at a minimum, to sentences with different open-class lexical items.
   b. Structural Specificity: Only a limited number of sentence structures (i.e., types of grammatical violation) are potentially affected by satiation.
   c. Between-speaker Consistency: At least across native speakers of English, the same sentence types (notably sentences involving *wh*-extraction of an argument from a *Wh*-Island, Complex NP, or Subject Island) are the ones that are, at least in principle, susceptible to satiation.
   d. Within-speaker Persistence: Once an individual has experienced satiation on a given sentence type, the increased acceptance persists for a considerable period of time, even in the absence of routine exposure to sentences of that type.

In judging whether an experimental effect qualifies as "satiation" in the relevant sense, the characteristics in (3) will serve as a guide.

In the sections that follow, three new experimental studies are presented and discussed in light of the following questions:

(4) a. Can satiation in fact be demonstrated in the laboratory, or were the findings in (Snyder 2000) simply an experimental artifact, as proposed in Sprouse's (2009) Response Equalization Hypothesis (REH)? (Sections 2 and 3)
   b. When satiation is induced in the laboratory, does it persist beyond the experimental session? (Section 4)
   c. Does the difficulty of inducing satiation vary between different sentence types that are susceptible to the effect? (Section 5)
   d. Does satiation on one sentence type ever carry over to judgments of related sentence types, for example, from *Whether*-Island violations to sentences violating another type of *wh*-island? (Section 6)
   e. How sensitive is the satiation phenomenon to details of experimental methodology? What aspects of the methodology appear to matter? (Section 7)

Section 7 includes a survey of the literature on satiation. Section 8 turns to larger questions: the nature of satiation and its relevance to the objectives of generative linguistics.

2. Review of the Original Findings

An important component of (Snyder 2000) was the evidence suggesting satiation can be induced in the laboratory, under controlled experimental conditions, and measured

objectively.[1] Yet, Sprouse (2009) raised an important concern: the findings might have been due to what he termed "response equalization", rather than syntactic satiation. To facilitate discussion of the issue, this section will describe Snyder's (2000) experiment in detail. Section 3 will present a new experiment based on that study but modified to preclude response equalization.

2.1. Overview of Methodology and Findings

The experimental task in (Snyder 2000) took the form of a lengthy, printed questionnaire. Native English speakers, with no prior exposure to linguistics, were required to provide acceptability judgments on sentence-meaning pairs. A certain number of (initially) unacceptable sentence structures were systematically repeated in the course of the questionnaire. Thus, the participant received a compressed version of the linguist's experience of judging structurally equivalent sentences on multiple, distinct occasions.

On each page there was a single item like (5) (Snyder 2000, p. 576).

(5) (Context: Maria believes the claim that Beth found a $50 bill.) Test Sentence: "What does Maria believe the claim that Beth found?" Judgment: ____ (Y/N)

Prior to starting, participants were told they would be asked for a series of 60 judgments. On each page there would be a declarative sentence (the "context") and then an interrogative sentence (the "test sentence"). Participants were instructed to provide a Yes/No judgment: Is the test sentence grammatically possible in English, given the meaning that fits the context? In other words, could the test sentence have the intended meaning and still be accepted as "English", in their personal opinion? Participants were advised that many items would be similar to one another, but they should not look back to previous pages or try to remember previous answers. Given that no two items would be identical, and given that the differences might be important, they should simply provide an independent judgment on each new test sentence and then move on.

Fifty of the items corresponded to a series of five experimental blocks (although this structure was invisible to participants). Each block contained items of the following types, in pseudo-random order: three fully grammatical items and seven items that would typically be perceived as anywhere from mildly to severely unacceptable, namely one item of each type in (6).[2]

(6) a. Adjunct-Island violation
(Context: Paula wrote two novels before meeting the great playwright.)
Test Sentence: "Who did Paula write two novels before meeting?"

b. Complex Noun Phrase Constraint (CNPC) violation
(Context: John believes the claim that Mary likes Tony.)
Test Sentence: "Who does John believe the claim that Mary likes?"

c. Left Branch Constraint (LBC) violation
(Context: Bill knows that Alice smoked two cigarettes.)
Test Sentence: "How many does Bill know that Alice smoked cigarettes?"

d. Subject-Island violation
(Context: Sally knows that a bottle of vinegar fell on the floor.)
Test Sentence: "What does Sally know that a bottle of fell on the floor?"

e. *That*-trace violation
(Context: Fred believes that Greta frightened Bob.)
Test Sentence: "Who does Fred believe that frightened Bob?"

f. *Want-for* violation
(Context: Bob wants Vanessa to buy a hammer.)
Test Sentence: "What does Bob want for Vanessa to buy?"

g. Whether-Island violation
(Context: Dmitri wonders whether John drinks coffee.)
Test Sentence: "What does Dmitri wonder whether John drinks?"

In addition to the 50 test items, the experimental materials included six practice items immediately prior to Block 1 and four post-test items immediately following Block 5. (The distinction between these items and the actual test items was invisible to participants.) No two test items were ever identical: even within a single sentence type, almost all the open-class lexical items differed across sentences. There were two exceptions: CNPC violations in the body of the experiment—but not the post-test—consistently used the phrase *believe the claim*, and *Whether*-Island violations in the body of the experiment—but not the post-test—consistently used the phrase *wonder whether*.

An informal poll of linguists (all of them native speakers of English) indicated to Snyder that the phenomenon of syntactic satiation was relevant (at least) to *wh*-extraction of an argument across a *Whether* Island (6g) and out of a complex noun phrase of the type in (6b). In contrast, there appeared to be no satiation on LBC violations (6c) or *That-trace* violations (6e).[3] Thus, Snyder reasoned that if syntactic satiation could indeed be induced by his task, there ought to be a systematic tendency for participants to become more accepting of *Whether*-Island violations and/or CNPC violations by the end of the experiment. There should not, however, be increased acceptance of LBC or *That-trace* violations. (For the other sentence types in (6), the possibility of satiation was treated as an open question.)

The findings were as follows: As predicted, for both *Whether* and CNPC items there was a significant increase in acceptance from the beginning (Blocks 1 and 2) to the end (Blocks 4 and 5) of the questionnaire (two-tailed $p < .05$ by Binomial Test).[4] In contrast, for LBC and *That-trace*, there was no appreciable change. Hence, the findings were broadly consistent with the possibility that the task was inducing the same kind of judgment change that linguists sometimes experience. (Of the other sentence types, only Subject Islands showed any appreciable increase, and it was only marginally significant; $p < .07$.)

The four post-test items following Block 5 were two fully grammatical fillers, plus the two items in (7).

(7) a. Complex Noun Phrase Constraint (CNPC) violation, with *accept the idea*
(Context: Madge accepted the idea that Bob would run for mayor.)
Test Sentence: "What did Madge accept the idea that Bob would do?"

b. Whether-Island violation, with *ask whether*
(Context: Mildred asked whether Ted had visited Stonehenge.)
Test Sentence: "What did Mildred ask whether Ted had visited?"

To check for a possible "carryover" effect from judging CNPC violations with *believe the claim* (as in Blocks 1–5; cf. 6b), to *accept the idea* (7a), Snyder focused on participants who had initially rejected both of the CNPC violations in Blocks 1 and 2. (The intention was to focus on individuals whose grammar had clearly excluded such sentences prior to the experiment.) These participants were first classified as satiating or not satiating on *believe the claim*, based on whether they accepted at least one of the two *'believe the claim'* items in Blocks 4 and 5. They were then cross-classified as accepting or rejecting the post-test item, (7a).

Both in (Snyder 2000) and in the new experiments reported below, a participant is classified as having "satiated" on a given sentence type if (and only if) one of the following three situations holds true: (i) the exemplars in the first two blocks of the study were both rejected and exactly one of the exemplars in the final two blocks was accepted, (ii) the exemplars in the first two blocks were both rejected and the exemplars in the final two blocks were both accepted, or (iii) exactly one of the exemplars in the first two blocks was accepted and both of the exemplars in the final two blocks were accepted.

The rate of acceptance of the post-test item among participants who had consistently rejected the CNPC violations, both in Blocks 1 and 2 and in Blocks 4 and 5, was calculated as a baseline. A binomial test was then used to assess the data from participants who had likewise rejected the CNPC violations in Blocks 1 and 2 but accepted at least one of the CNPC violations in Blocks 4 and 5 (i.e., had satiated), in order to answer the following question: What was the probability of obtaining, simply by chance, an acceptance rate for the post-test item that was as high as (or even higher than) the rate observed in these latter

participants? "Simply by chance" meant that the probability was calculated under the null hypothesis that, in general (among participants who rejected both items in Blocks 1 and 2), the participants who satiated (i.e., accepted at least one of the items in Blocks 4 and 5) had the same probability of accepting the post-test item as the participants who had not satiated.

In Snyder's data, all 22 participants had rejected the CNPC violations (i.e., with *believe the claim*) in Blocks 1 and 2. Of those 22, 17 also rejected the CNPC violations in Blocks 4 and 5. Only four of these 17 "non-satiators" accepted the post-test item with *accept the idea*. In contrast, among the five satiators, four accepted the post-test item. Under the null hypothesis that the general acceptance rate for satiators was the same as for non-satiators, namely 4/17 = 23.5%, the likelihood of seeing acceptance by at least four out of five satiators simply by chance is given by the binomial test: in the present case, two-tailed $p < .05$. Hence, there was significant carryover.

In the case of *Whether*-Island violations, 18 of the 22 participants rejected the items (i.e., with *wonder whether*) in Blocks 1 and 2. Of these 18, seven also rejected the *wonder-whether* items in Blocks 4 and 5. Only three of these non-satiators accepted the post-test item (with *ask whether*). This provided a baseline acceptance rate of 3/7 = 42.9%. Of the 11 satiators, however, 10 accepted the post-test item (binomial $p < .005$). Hence, there was also significant carryover for *Whether* Islands.[5]

In sum, Snyder (2000) obtained statistically reliable satiation on argument *wh*-extraction from both the complex-NP (*believe the claim*) environment and the w*onder-whether* environment, although far fewer participants showed the effect with complex NPs (five out of 22, as opposed to 11 of 22 for *whether*). Moreover, the satiation overwhelmingly "carried over" from *believe the claim* to *accept the idea* and from *wonder whether* to *ask whether*: four of the five satiators on CNPC violations exhibited carryover, as did 10 of the 11 satiators on *Whether* Islands.

2.2. Some Possible Concerns

A few further details of methodology are important for the present discussion. A major issue in any type of work with acceptability judgments is the fact that many different factors can influence them. These include not only the grammatical structure of the sentence being judged, but also the choices of open-class lexical items and the characteristics of the test item that was judged immediately prior. Therefore, alongside satiation, one of the possible reasons for participants to become more accepting of a given sentence type, as they work their way through an experiment, is that the specific examples presented later in the experiment are somehow intrinsically more acceptable, for reasons independent of their grammatical structure (e.g., due to the open-class lexical items that they happen to contain). Another possibility is that the specific test items positioned immediately prior to the sentences of interest made the earlier sentences seem less acceptable, and/or the later ones seem more acceptable, than they would ordinarily.

A simple way to minimize these possibilities is to counterbalance, across participants, the order of presentation: half the participants receive the items in forward order, and the other half receive the same items but in reverse order. Snyder (2000) therefore gave half of his participants a questionnaire containing the 50 test items in the order "..., Item 1, Item 2, Item 3, ..." and gave the other half the same items but in the order "..., Item 50, Item 49, Item 48, ...". Any items that were intrinsically more acceptable than others of the same type would yield an increase in acceptability for half the participants but an equally strong decrease for the other half. Similarly, if judging a certain test item had a special effect on the participant's next judgment, then this effect would apply to different "next judgments" in the different orders of presentation. Crucially, if the experiment induced actual satiation on sentences of a given grammatical type, then it should yield increasing acceptance not only overall but also both in the subset of participants who received a "forward" order of presentation and in the subset who judged the same items but in reverse order.[6]

2.3. Response Equalization?

Let's now consider Sprouse's (2009) Response Equalization Hypothesis (REH). One type of task effect that is not addressed simply by counterbalancing the order of presentation, and that must be addressed separately, is the following. Suppose that participants come to any Yes/No task, such as the one in (Snyder 2000), with an expectation that exactly half the test items will have an expected answer of "Yes". A problem, then, is that, for 50 of the 60 items in Snyder's experiment (i.e., the five blocks of ten mentioned above), there was a ratio of seven items with a grammatical violation for every three items that were fully grammatical. Now, blocking of the items was invisible to the participants, and exactly half of the other 10 items (i.e., practice and post-test items) were fully grammatical. Therefore, participants saw a single series of 60 items, and 40 of them (66.7%) contained a grammatical violation.

Assuming participants noticed the discrepancy between the expected frequency of "Yes" items (50.0%) and the actual frequency (presumably 33.3%, for an unchanging native-speaker grammar of English), the REH says participants should have become more willing to say "Yes" as the experiment progressed. To make sure that Snyder's (2000) findings were not simply due to response equalization, the best approach is to rerun the experiment with exactly one change: add enough fully grammatical items for a 1:1 balance. This will be the first of three new experiments reported below.[7]

3. Experiment I: A Direct Test of the Response Equalization Hypothesis

3.1. Materials

Experiment I was identical to the experiment in (Snyder 2000) except that 20 new, fully grammatical test items were added to the questionnaire, so as to create a perfect balance: 40 items that were fully grammatical and 40 that violated a grammatical constraint. For each of the experimental blocks in Version A, four of the new items were randomly selected and inserted among the original 10 items, as follows: 1 2_3 4_5 6_7 8_9 10. Following these additions, each of the five blocks contained seven fully grammatical items and seven grammatical violations (one item for each of the seven types in (6), above), and there were never more than two expected "NO" items in a row. A new Version B was created from Version A by reversing the order of the resulting 70 test items. Together with the six practice items and four post-test items, this yielded 80 items per participant.

In keeping with the original materials of (Snyder 2000), the new grammatical items were designed to be comparable in their structural complexity to the ungrammatical items. Some representative examples are provided in (8b,d,f).

(8) a. CNPC violation, with *believe the claim*:
(Context: Maria believes the claim that Beth found a $50 bill.)
Test Sentence: "What does Maria believe the claim that Beth found?"

b. Grammatical item, with *claim to believe*:
(Context: John claims to believe that Mary likes Tony.)
Test Sentence: "Who does John claim to believe that Mary likes?"

c. Whether-Island Violation, with *wonder whether*:
(Context: Henry wonders whether George discovered the answer.)
Test Sentence: "What does Henry wonder whether George discovered?"

d. Grammatical item, with *wonder what*:
(Context: Gina wonders whether Einstein discovered relativity.)
Test Sentence: "Who wonders what Einstein discovered?"

e. LBC violation, with *how many ... books*:
(Context: Edwin thinks Margaret read three books.)
Test Sentence: "How many does Edwin think Margaret read books?"

f. Grammatical item, with *how many books*:
(Context: Edward thinks that Anne read ten books.)
Test Sentence: "How many books did Edward think that Anne had read?"

Under the REH account of Snyder's (2000) findings, the prediction for Experiment I (where participants now see the same number of expected "YES" and expected "NO" items) is that there will be no systematic tendency for any sentence type to be accepted more often at the end than at the beginning of the experiment. In contrast, what we might call the "Satiation Hypothesis" predicts an increased likelihood of "Yes" responses at later points in the experiment for *Whether*-Island and/or CNPC violations but no systematic tendency toward increasing acceptance of *That*-trace or LBC violations.

3.2. Plan for Data Analysis

In a yes–no task, the responses cannot be expected to obey a normal (Gaussian) distribution. Snyder (2000) therefore relied primarily on binomial tests and Fisher Exact Tests, which are both "non-distributional" in the sense of not assuming a normal distribution. Here, the approach to data analysis will again rely on non-distributional methods of two main types. First, for each of the initially unacceptable sentence types, a Wilcoxon Signed-Rank test will be used to assess whether "Yes" responses were significantly more frequent at the end of the experiment (in the final two blocks) than at the beginning (in the initial two blocks).

Second, whenever possible, mixed-effect (ME) logistic regression will be used as a follow-up test. The logistic extension to linear regression is in many ways ideal for the analysis of yes–no judgment data, but sometimes, there are difficulties in achieving convergence (i.e., in fitting a model to the dataset), especially if the number of participants is relatively small. Given that convergence is not always possible, the role of ME Logistic Regression will be secondary: in the event that convergence cannot be achieved, the results of the Wilcoxon Tests will have to suffice. (In practice, such a situation will arise during the analysis of data from Experiment II, below.)

When applying ME logistic regression, the search for a model fit will always begin with a "maximally" specified model (cf. Barr et al. 2013), which will then be simplified if necessary in order to achieve convergence. A limit on simplification, however, will be that the Random Effects (RE) portion of the model must always include "random intercepts" for individual participants and for individual test items and must include by-participant "random slopes" for the effect of each of the major factors in the experiment. (For the present purposes, the major factors are the sentence type being judged and the block of the experiment in which the judgment is made.) This ensures that the model is appropriately adjusted for (i) variation in the overall willingness of a participant to say "yes" to test items in general (i.e., the by-participant random intercept), (ii) the participant's general willingness to say "yes" to each of the different types of sentence (i.e., the by-participant random slopes for Type), and (iii) the participant's general willingness to say "yes" in each successive Block of the experiment (i.e., the by-participant random slope for Block). It also ensures that the model adjusts for variation across the different sentences (i.e., "ItemCodes") that exemplify a particular sentence type (i.e., within any single experimental treatment).

A small change from (Snyder 2000) is that the blocks of Experiment I (as well as Experiments II and III below) will be numbered from 0 to 4, rather than 1–5. This has the desirable consequence that, for each of the (initially) unacceptable sentence types, the block number can be interpreted as the participant's number of previous exposures to that sentence type during the experiment.

One special strength of ME Logistic Regression is its ability to evaluate a given participant's response to a test sentence relative to that same participant's responses to control sentences. The control sentences (in all of Experiments I–III) will be grammatically well-formed sentences that are similar to the test sentences in their structural complexity and that are judged in the same block as the corresponding test item. If participants experience genuine satiation on sentences of type T, then we expect ME logistic regression to reveal a significant interaction between block number and sentence type, for sentence type T.

More precisely, ME logistic regression will be conducted with one level of each factor specified as a baseline for use in "treatment contrasts" (i.e., pairwise comparisons) with

each of the other levels of that factor. For Type, the baseline level will be "Good" (i.e., within each block, the results for the seven fully grammatical items). A treatment contrast will then be calculated for each of the seven other (i.e., deviant) sentence types. Crucially, for each of these non-baseline levels, the analysis will check for an interaction effect: did the effect of "changing" from the grammatical items (the baseline) to an item of that type differ significantly, as a function of the experimental block in which the judgments were made?

Finally, evidence of increased acceptance at the end of the experiment (in the form of a significant Wilcoxon test and, when ME logistic regression converges, a significant interaction effect) is necessary, but not sufficient, for a claim of satiation on T. If participants exhibit genuine satiation of the kind characterized earlier in (3), then we expect some additional findings, and we need to confirm their presence. Specifically, the increased acceptability of a given sentence type following satiation should be evident regardless of the order in which sentences were presented. Hence, the next step will be to examine the data from Versions A and B separately. If genuine satiation occurred, we expect each version to show a statistically significant increase, from the beginning to the end of the experiment, in the frequency of acceptance.

3.3. Experimental Participants and Procedure

The participants in Experiment I were 22 undergraduate students, all native speakers of English, who were recruited by means of printed flyers posted on campus. Compensation was provided in the form of a $5 gift card, redeemable at the university bookstore. Participants were brought into an individual testing room and told the instructions (which were also provided in printed form). Participants then received the materials in the form of a printed booklet, exactly as in (Snyder 2000). Completion of the task took about 15 min.

3.4. Checking for Outliers

Prior to running inferential statistics, the data were checked for participants more than two standard deviations from the group average on either expected "YES" or expected "NO" items, because any such participants may not have understood the instructions. Indeed, two participants were more than two standard deviations below the group mean on acceptance of expected "YES" items and were excluded from further analysis, leaving $N = 20$.[8]

3.5. Primary Analysis: Wilcoxon Tests

Wilcoxon Signed-Rank tests were used to assess statistical reliability of changes in acceptance rate, for each sentence type, between the first two blocks (0 and 1) and the final two blocks (3 and 4). The main results were as follows. Acceptance in Blocks 3 and 4 was significantly greater for *whether* items ($W = -93$, $n_{s/r} = 14$, $Z = -2.9$, $p < .005$), but there was no significant change for any other sentence type (all $p > .10$). The data are shown graphically in Figures 1–4.[9]

For *whether* items, when the 20 participants are viewed individually, some 14 showed a change between the initial two blocks and the final two, and in 13 cases, it was an increase. (Nine increased from 0/2 to 1/2, three increased from 0/2 to 2/2, and one increased from 1/2 to 2/2. The individual showing a decrease changed from 2/2 to 1/2.) Among the six participants whose level of acceptance was unchanged, three consistently rejected the sentences, and three consistently accepted them.

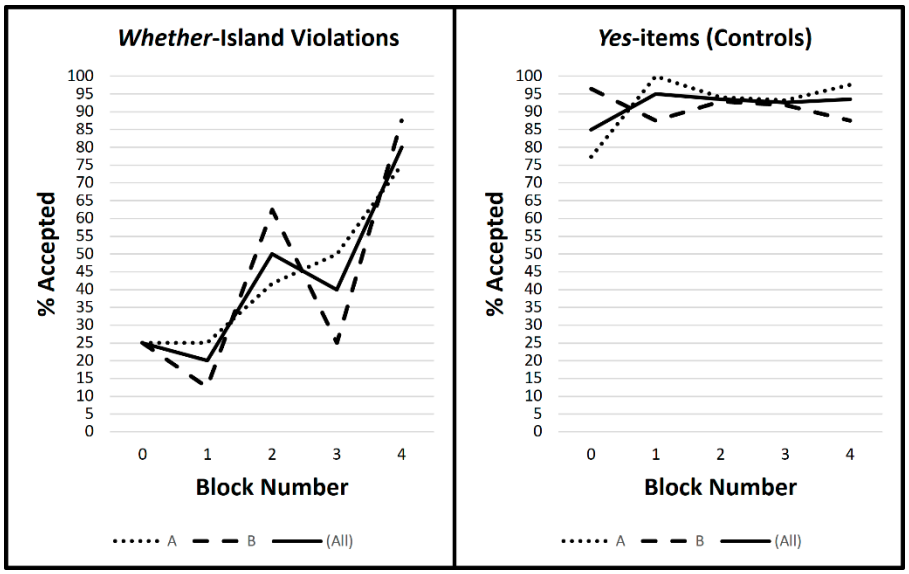

Figure 1. (**Left**) Percentage of participants in each block of Experiment I, who accepted the *Whether*-Island violation; Version A (N = 12) used forward presentation; Version B (N = 8) used reverse order; "All" indicates the total (N = 20). (**Right**) Mean percentage of the "Yes" items that were accepted; each participant judged seven items per block.

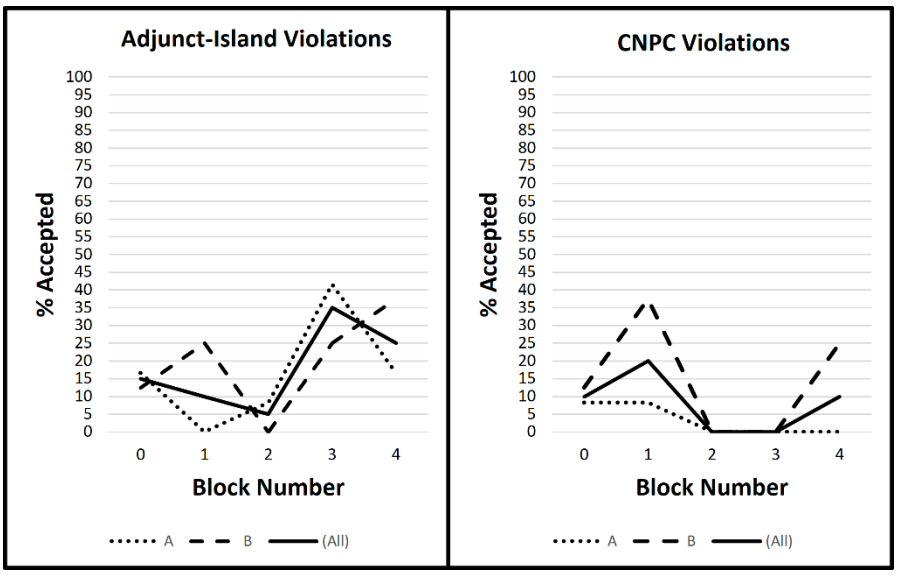

Figure 2. Experiment I, Adjunct-Island and CNPC violations.

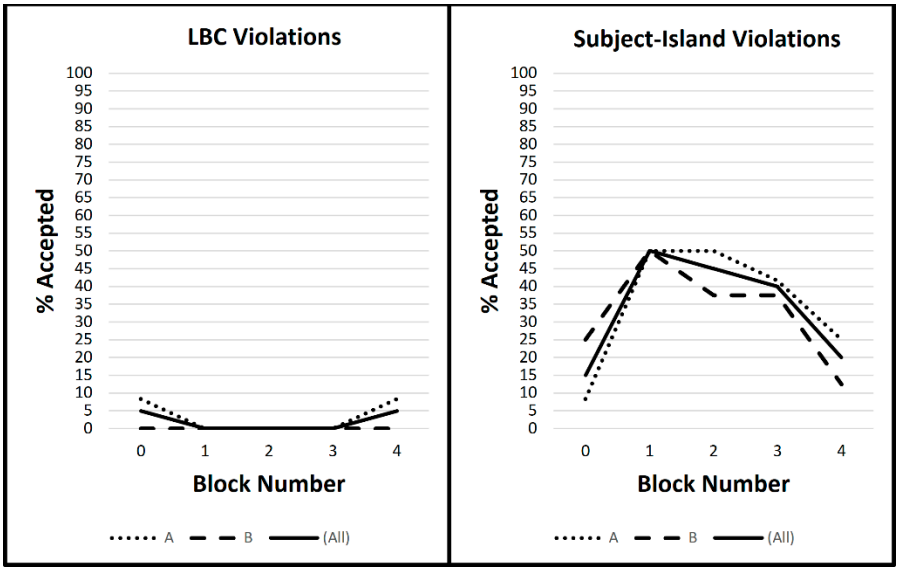

Figure 3. Experiment I, LBC and Subject-Island violations.

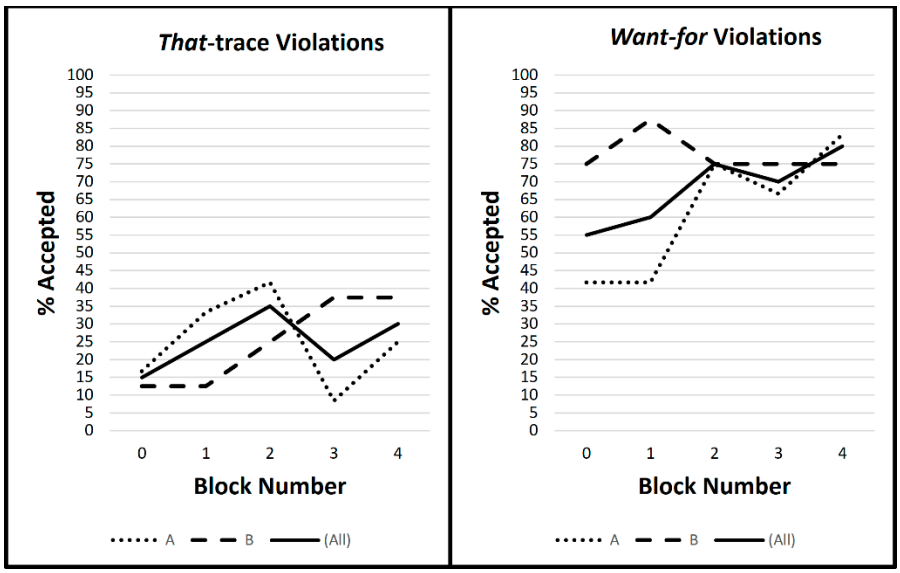

Figure 4. Experiment I, *That*-trace and *Want-for* violations.

3.6. Cross-Checking: ME Logistic Regression

Linear modeling was performed using R (version 3.2.3; R Core Team 2015) and the lme4 software package (version 1.1–11; Bates et al. 2015). An ME logistic model was constructed using lme4's "glmer" function. In the notation of the lme4 package, the model was specified as in (9).

(9) Response~Block*Type + (1 + Block + Type | Participant) + (1 | ItemCode) + (1 | Version)

Thus, the software searched for an optimally specified logistic-regression model with which to "predict" each participant's yes/no response to each test item, based on (i) the (integer) number of the block (0–4) in which the test item appeared, (ii) the grammatical type 'T' of the test item, and (possibly) (iii) an interaction effect between Block and Type for each (non-baseline) value of Type. As noted above, a significant interaction is precisely what we expect to see if participants experience satiation on a given sentence type.

The initial attempt to fit a model with structure (9) to the data from Experiment I was unsuccessful: the glmer program failed to converge. Inspection of the program's best attempt revealed two issues. First, in the RE structure, the random intercepts by Version

explained none of the variation in the dataset.[10] Second, in the fixed-effects structure for the program's "best attempt" at a fit, the main effect of Block had an estimated coefficient (0.16) that was an order of magnitude smaller than the coefficients for the main effects of the different levels of Type (which ranged from 1.73 to 8.21). A difference in scale of one or more orders of magnitude can prevent convergence. Hence, two changes were made. First, Version was removed from the RE structure in (9). Second, the factor Block was re-scaled: the values of Block in the dataset were simply divided by 10 (so that Block ranged from 0.0 to 0.4).

Following these changes, the program converged on the model summarized in Table 1.[11] As expected, pairwise comparisons showed that each of the ungrammatical levels of Type differed significantly from the grammatical (baseline) items. There was no main effect of Block ($p > .10$), and there was exactly one significant interaction of Block with Type, namely for Type = *Whether*; acceptance of *Whether* items increased significantly, as the participants progressed from Block 0 to Block 4.

Table 1. Table of fixed effects for Experiment I.

Predictor	Estimate	SE	*Z*	*p*
(Intercept)	2.90	0.44	6.64	<.001
Block	1.41	1.49	0.95	(>.10)
TypeAdjunct	−6.28	1.05	−5.98	<.001
TypeCNPC	−5.83	1.08	−5.41	<.001
TypeLBV	−10.64	3.40	−3.13	<.01
TypeSubject	−3.90	0.83	−4.68	<.001
TypeThat	−5.01	0.93	−5.38	<.001
TypeWant	−2.17	1.06	−2.04	<.05
TypeWhether	−5.68	1.03	−5.50	<.001
Block:TypeAdjunct	3.58	2.74	1.31	(>.10)
Block:TypeCNPC	−3.46	3.31	−1.05	(>.10)
Block:TypeLBV	0.18	5.99	0.03	(>.10)
Block:TypeSubject	−1.62	2.38	−0.68	(>.10)
Block:TypeThat	0.65	2.48	0.26	(>.10)
Block:TypeWant	3.11	2.74	1.14	(>.10)
Block:TypeWhether	9.58	3.00	3.20	<.01

Thus, the results of ME logistic regression are entirely consistent with the results from Wilcoxon tests: in Experiment I there was possible satiation on items with *wonder whether*, but (in contrast to Snyder 2000) there was no satiation on the complex-NP items with *believe the claim* ($p > .10$). Consistent with Snyder 2000, there was no satiation on any of the other sentence types tested.

3.7. Follow-Up Testing

The next question is whether the apparent satiation on *wonder whether* meets the additional criterion discussed above: Did acceptance increase, from the beginning to the end of the questionnaire, in both versions? Indeed, Versions A and B each showed the same general pattern as the full study. Overall (as noted above), fourteen participants showed a change, and in 13 cases, it was an increase. On Version A, seven participants changed, and in all cases, it was an increase. On Version B, seven participants changed, and in six cases, it was an increase. Hence, the findings conform very closely to what is expected in satiation.

Experiment I shows that satiation can indeed be obtained under laboratory conditions, at least for *wonder whether* items, even if participants judge a perfect balance of fully acceptable, versus initially unacceptable, sentences. The main difference from Snyder 2000 is the absence of a change for CNPC items. In Section 5, we will see evidence that the final sample size ($N = 20$) in Experiment I was far too low for reliable detection of satiation on CNPC sentences, but regardless, the specific sentence type (*wonder whether*) that showed satiation was also one of the types showing it in (Snyder 2000). Hence, the findings from Experiment I are fully in-line with Between-speaker Consistency (3c) (as well as Generality and Structural Specificity). Next, we check for Within-speaker Persistence (3d).

4. Experiment II: Persistence

Did the increase in acceptance of *Whether*-Island violations observed in Experiment I persist beyond the time of the experiment? To find out, each participant was invited to return for testing one month later. Of the 20 participants whose data were included in the analyses for Experiment I, 15 agreed to return.

Each of these participants was tested again, 4 to 5 weeks later, in much the same way as the first time. In almost all cases, if a participant (for example) received Version A at Time 1, then Version B was given at Time 2. One participant was accidentally given B at both Time 1 and Time 2. Among the other 14, eight received version A and six version B at Time 1; hence, six (of these 14) received A and eight received B at Time 2.

The predictions were as follows. If the satiation on *Whether* items in Experiment I quickly faded, then there should be no significant difference between participants' judgments at the beginning of Experiment I (Blocks 0 and 1), and the same participants' judgments one month later at the beginning of Experiment II (Blocks 0 and 1). In contrast, if the satiation that was detected on *Whether* items showed Within-speaker Persistence, then, at least for *Whether* items, the frequency of acceptance at the beginning of Experiment II should be significantly higher.

Moreover, if the satiation on *Whether* items persisted, this should be evident when we examine the data by participant. For example, someone who accepted neither of the examples in Blocks 0 and 1 of Experiment I but accepted one of the examples in Blocks 3 and 4 of Experiment I would be expected to accept at least one of the two examples in Blocks 0 and 1 of Experiment II.

4.1. Primary Analysis: Wilcoxon Tests

Wilcoxon Signed-Rank Tests were performed to check for increased acceptance at Time 2. For each sentence type, each participant's responses in Blocks 0 and 1 of Experiment I were compared against Blocks 0 and 1 of Experiment II. As predicted by Within-speaker Persistence, there was a significant increase for *Whether* items from the beginning of Experiment I ("Time 1") to the beginning of Experiment II ("Time 2"; $W = -45$, $n_{s/r} = 9$, $p < .01$). No other sentence type showed a significant increase. On *Whether*, when the participants are viewed individually, six were consistent across Times 1 and 2, and nine showed a change. In all cases, if there was a change, it was an increase: for four participants, from 0/2 "yes" responses at Time 1 to 1/2 "yes" at Time 2 and, for five participants, from 0/2 at Time 1 to 2/2 at Time 2. The full results are shown graphically in Figure 5.

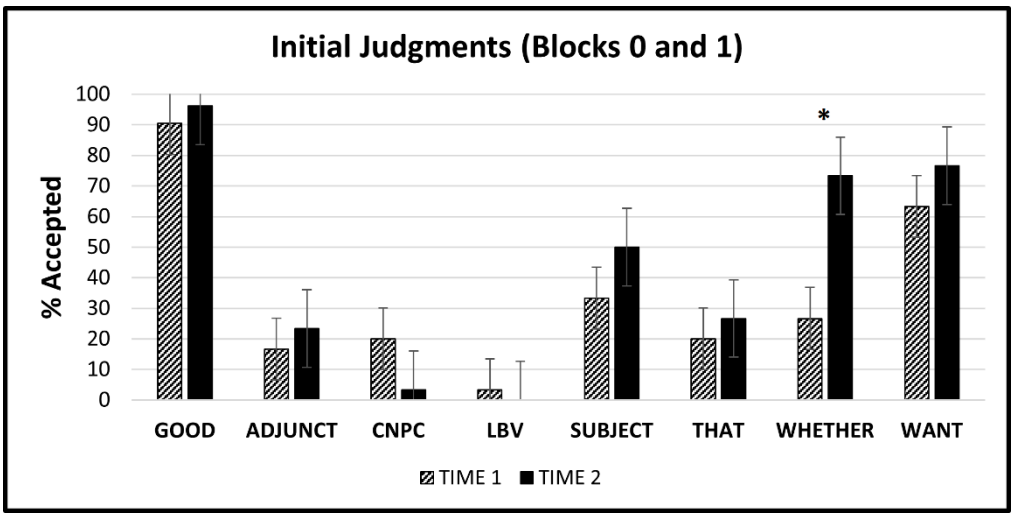

Figure 5. Experiment II: Acceptance of initial sentences (Blocks 0 and 1) at Times 1 and 2; Error bars show standard error; "*" indicates significance ($p < .05$) by Wilcoxon Signed-Rank Test.

The by-participant results are shown in Figure 6. Of the 15 individuals who participated in Experiment II, 11 rejected both of the *Whether*-Island violations at the beginning (Blocks 0 and 1) of Experiment I, and four accepted both of them. Of the 11 who rejected them, one participant continued to reject them at the end (Blocks 3 and 4) of Experiment I, while the other ten accepted at least one (i.e., they had satiated). As can be seen in the table, eight of the ten satiators accepted at least one of the two exemplars of a *Whether*-Island violation in Blocks 0 and 1 of Experiment II; the remaining two satiators both accepted one fewer than they had in Blocks 3 and 4 of Experiment I. (The four participants who accepted the exemplars in Blocks 0 and 1 of Experiment I continued to accept (in all but one case) the exemplars at the end of Experiment I and beginning of Experiment II.)

	Initially 0											*Initially 2*			
Participant ID:	**16**	**4**	**7**	**22**	**3**	**6**	**11**	**1**	**5**	**8**	**10**	**12**	**9**	**15**	**21**
T1,Beginning:	0	0	0	0	0	0	0	0	0	0	0	2	2	2	2
T1,End:	0	1	1	1	1	1	1	2	2	1	2	1	2	2	2
T2,Beginning:	0	1	1	1	2	2	2	2	2	0	1	2	2	2	2

T1 Satiation

T2 Maintenance of Satiation

Figure 6. By-participant findings in Experiment II, showing the number (0–2) of Whether-Island violations accepted at three time points (Beginning and End of Experiment I and Beginning of Experiment II).

4.2. Cross-Checking: ME Logistic Regression

The complete Time 1 and Time 2 data for Blocks 0 and 1 were submitted to GLMER, with the following model specification.

(10) Response~Time * Type + (1 + Time + Type | Participant) + (1 + Time | ItemCode) + (1 | Version)

Unfortunately, GLMER failed to converge on a fit (even when the RE component for Version had been removed, due to a theta parameter of zero). Given that the primary point of interest concerned the *Whether* items in relation to the grammatically well-formed (Good) items (i.e., because *Whether* was the only sentence type for which the Wilcoxon tests had indicated a significant effect), the dataset was next trimmed to include only the *Whether* and Good sentence types. At that point, using the model specification in (10), GLMER succeeded. The resulting parameters for the fixed effects are shown in Table 2. The results are fully consistent with those from the Wilcoxon tests: the effect of "changing" from the control items to the *Whether* items was a large reduction in acceptance at Time 1 but a significantly smaller reduction at Time 2.

Table 2. Table of fixed effects for Experiment II.

Predictor	Estimate	SE	*Z*	*p*
(Intercept)	3.53	0.82	4.31	<.001
TimeT2	2.21	1.69	1.31	(>.10)
TypeWhether	−7.11	2.76	−2.57	<.05
TypeWhether:TimeT2	7.90	4.01	1.97	<.05

4.3. Follow-Up Testing

As indicated in Section 4.1, when we compare the initial responses (i.e., during the first two blocks of the stimuli) for Time 1 and Time 2, nine of the 15 participants showed an increase in their acceptance of *whether* items, and none showed a decrease. The results were very similar for each version. Of the eight participants who saw Version A at Time 1 and Version B at Time 2, five (i.e., about half) showed an increase from Time 1 to Time 2, and none showed a decrease. Of the six who saw Version B at Time 1 and A at Time 2, three

increased, and none decreased. The one participant who saw Version B at both Time 1 and Time 2 also showed an increase.

In sum, Experiment II indicates that the characteristic of Within-speaker Persistence (3d), reported anecdotally by linguists, also holds (at least in the case of *Whether*-Island violations) for experimentally induced satiation in non-linguists. The increase in "yes" responses on *Whether* items that was statistically significant by the end of Time 1 testing, was still statistically significant four weeks later. Indeed, when we examined performance of the 15 participants individually (Section 4.1), we found that 10 had changed at Time 1 from zero "yes" responses in Blocks 0 and 1 to at least one "yes" response in Blocks 3 and 4 (i.e., they had satiated). In Blocks 0 and 1 of the Time 2 testing, nine of these 10 satiators still said "yes" to at least one *Whether* item. Indeed, given the extremely low likelihood of encountering any *Whether*-Island violations between testing sessions, and given that participants had judged only six examples at Time 1 (five with *wonder whether*, plus a post-test item with *ask whether*), the persistence of the satiation effect is remarkable. This degree of persistence suggests that the satiation the participants had experienced on *whether* items was a "learning" effect rather than a short-term priming effect.

5. Experiment III: Variation in Effect Size

5.1. Overview

A possible concern about Experiments I and II is that they are based on a sample of only 15–20 individuals. This is an especially important consideration given that satiation on CNPCs was weak in (Snyder 2000) (i.e., detected in only five of 20 participants) and not detected at all in Experiments I and II. Increasing the sample size will potentially allow us to reproduce, and better characterize, whatever satiation effect is present for CNPCs.

In Experiment III, the sample size was increased to 151 individuals. The participants were undergraduates taking a large introductory course on the philosophy of language. (None of them had participated in Experiment I or II.) The stimuli were nearly identical to those in Experiments I and II, but they were presented online (one item at a time, so as to control the order in which the judgments were made), and the two items shown in (11) were added to the post-test.

(11) a. *Wh*-Island violation with *wonder why*
(Context: Olga wonders why Sally likes Fred.)
Test Sentence: "Who does Olga wonder why Sally likes?"

b. *Wh*-Island violation with *know how*
(Context: Sue knows how Bill fixed the motorcycle.)
Test Sentence: "What does Sue know how Bill fixed?"

Two fully grammatical items were also added to the post-test. Hence, a participant judged 84 items in total.

Participants began by answering questions about their language background and then were randomly assigned to Version A or B. The initial sample included 194 individuals, but the data were discarded from 29 participants who reported (in answer to the initial questions) that English was not the first language they had acquired. Data were also discarded if a participant's rate of "Yes" responses to fully grammatical items was more than two standard deviations below the group's average or if the rate of "Yes" responses to "deviant" items was more than two standard deviations above the average. Fourteen additional individuals were excluded by these criteria for a final sample of 151.

5.2. Primary Analysis: Wilcoxon Tests

Wilcoxon Signed-Rank Tests indicated possible satiation on four sentence types, namely the sentences violating *Whether*-Island, CNPC, *That*-trace, and Subject-Island constraints. On *Whether* Islands, 70 of 151 participants showed a change between the initial two and the final two blocks, and for 56, it was an increase ($W = -1{,}645$, $n_{s/r} = 70$, $Z = -4.81$, $p < .0001$). For CNPC violations, 36 showed a change, and for 28, it was an increase ($W = -394$, $n_{s/r} = 36$, $Z = -3.09$, $p < .005$). For *That*-trace violations, 72 showed a change,

and for 49, it was an increase ($W = -940$, $n_{s/r} = 72$, $Z = -2.64$, $p < .01$), and for Subject Islands, 60 showed a change, and for 40, it was an increase ($W = -650$, $n_{s/r} = 60$, $Z = -2.39$, $p < .05$). None of the other sentence types showed a significant change.

5.3. Cross-Checking: ME Logistic Regression

Findings were cross-checked using ME logistic regression with the same model structure (9) that was tried initially on the data from Experiment I (i.e., with random intercepts for Version and without any re-scaling of the Block number). The software converged on a model fit, as shown in Table 3, and indicated possible satiation on extraction from *Whether* Islands, Complex NPs, Subject Islands, and *That-trace* environments. Hence, the results were fully consistent with the results from the Wilcoxon tests.

Table 3. Table of fixed effects for Experiment III.

Predictor	Estimate	SE	*Z*	*p*
(Intercept)	3.10	0.22	13.89	<.001
Block	−0.01	0.04	−0.13	(>.10)
TypeAdjunct	−5.28	0.54	−9.75	<.001
TypeCNPC	−7.71	0.70	−10.95	<.001
TypeLBC	−10.72	1.58	−6.80	<.001
TypeSubject	−5.14	0.55	−9.43	<.001
TypeThat	−4.73	0.53	−8.84	<.001
TypeWant	−1.78	0.53	−3.37	<.001
TypeWhether	−4.38	0.52	−8.38	<.001
Block:TypeAdjunct	0.11	0.09	1.29	(>.10)
Block:TypeCNPC	0.53	0.14	3.81	<.001
Block:TypeLBC	−0.28	0.24	−1.19	(>.10)
Block:TypeSubject	0.24	0.10	2.56	<.05
Block:TypeThat	0.23	0.08	2.79	<.01
Block:TypeWant	0.12	0.09	1.45	(>.10)
Block:TypeWhether	0.40	0.08	4.85	<.001

5.4. Follow-Up Testing

The next step was to check whether these cases met the additional criterion discussed above: Did acceptance increase significantly in both versions? For *wonder whether*, findings were fully consistent between versions. Recall that with the two versions combined, 70 participants showed a change in acceptance, and for 56, it was an increase. In Version A, 34 showed a change, with 28 increasing (Wilcoxo*n* Signed-Rank Test, $W = -399$, $n_{s/r} = 34$, $Z = -3.41$, $p < .001$), and in Version B, 36 showed a change, with 28 increasing ($W = -434$, $n_{s/r} = 36$, $Z = -3.41$, $p < .001$). This qualifies as reliable evidence of a satiation effect on *Whether* Islands.

For CNPC violations as well, the findings were fully consistent across versions. Recall that, with the two versions combined, 36 participants showed a change, with 28 increasing. In Version A, 18 showed a change, with 15 increasing ($W = -114$, $n_{s/r} = 18$, $Z = -2.47$, $p < .05$), and in B, 18 showed a change, with 13 increasing ($W = -91$, $n_{s/r} = 18$, $Z = -1.97$, $p < .05$). Note that, in their initial acceptance rate, the CNPC items were quite similar to LBC items. Figure 7 provides a side-by-side comparison of LBC, where Block 0 acceptance was approximately 5% and no satiation was evident, versus CNPC, where Block 0 acceptance was just under 5% and satiation clearly occurred.

In the case of *That*-trace, recall that 72 participants showed a change, and for 49, it was an increase. Yet, this increase was overwhelmingly driven by Version B, where 39 showed a change, and for 30 (i.e., 77%), it was an increase ($W = -444$, $n_{s/r} = 39$, $Z = -3.09$, $p < .005$). On A, however, where 33 showed a change, this was an increase in only 19 (58%) of the cases ($W = -47$, $n_{s/r} = 33$, $Z = -0.42$, $p > .10$ NS). The lack of a significant change in Version A means the findings do not qualify as reliable evidence of satiation. Instead, they were quite possibly an artifact of the particular order in Version B. (For a side-by-side comparison

of *That*-trace with a Block 0 acceptance of approximately 25% and *Whether*-Island violations with a Block 0 acceptance just above 30%, see Figure 8.)

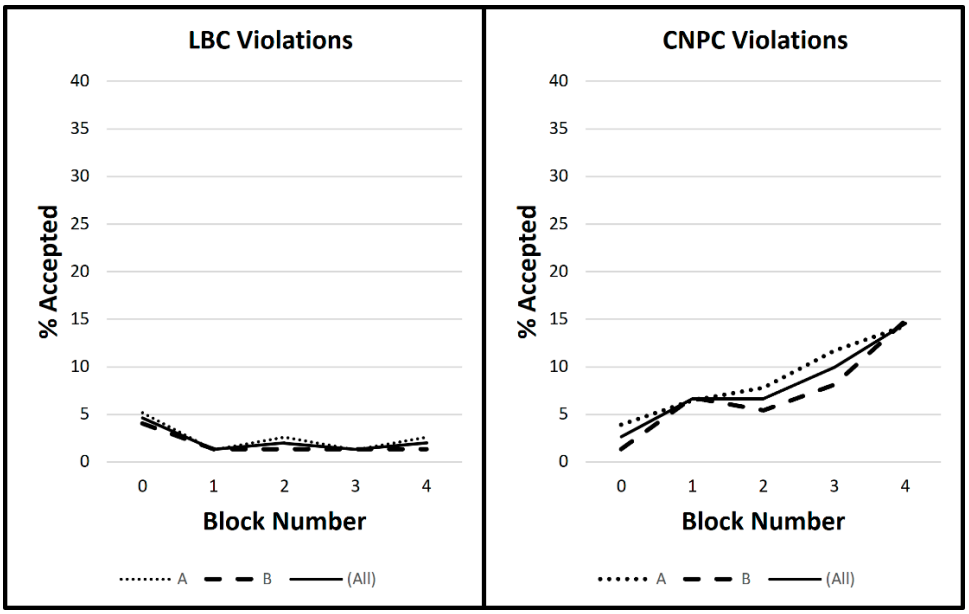

Figure 7. Comparison of LBC violations and CNPC violations in Experiment III.

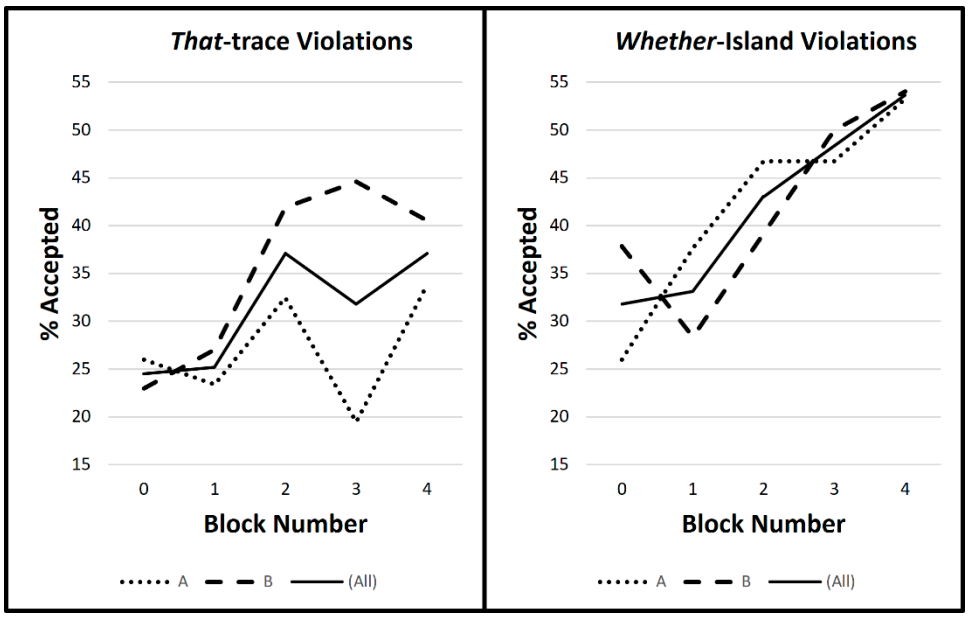

Figure 8. Comparison of *That*-trace and *Whether*-Island violations in Experiment III.

Figures 9 and 10 show the findings for the remaining sentence types. In the case of Subject Islands, recall that 60 participants showed a change, and for 40, it was an increase. When the versions are viewed separately, there is still a significant change in version A ($W = -213$, $n_{s/r} = 30$, $Z = -2.19$, $p < .05$), but the change observed in version B does not reach significance ($W = -116$, $n_{s/r} = 30$, $Z = -1.19$, $p > .10$ NS). The absence of a significant change in Version B means the findings from Experiment III do not qualify as reliable evidence of satiation, but this could well change in a follow-up study (as will be discussed momentarily).

In sum, Experiment III provides clear evidence of satiation on *Whether* Islands and complex NPs but not on the other sentence types examined. The results are entirely

consistent with (Snyder 2000) and largely consistent (i.e., in all respects, except for CNPC) with Experiments I and II. Once again, there was no possibility of response equalization (in the sense of the REH), and yet, a familiar pattern emerged: clear-cut satiation on *Whether* Islands (and, this time, also on CNPC violations, as in Snyder 2000) but not on LBC violations and not on *That*-trace violations. There was also no reliable evidence of satiation on Subject Islands, Adjunct Islands, LBC violations, or *Want-for* environments. Naturally, this does not preclude the possibility that one or more of these latter sentence types will show clear evidence of satiation in another study, especially if the experimental conditions are different. Indeed, in the literature review in Section 7, we will see that satiation is sometimes found for Subject-Island violations but chiefly in studies where participants judged a greater number of examples than they did in Experiments I–III.

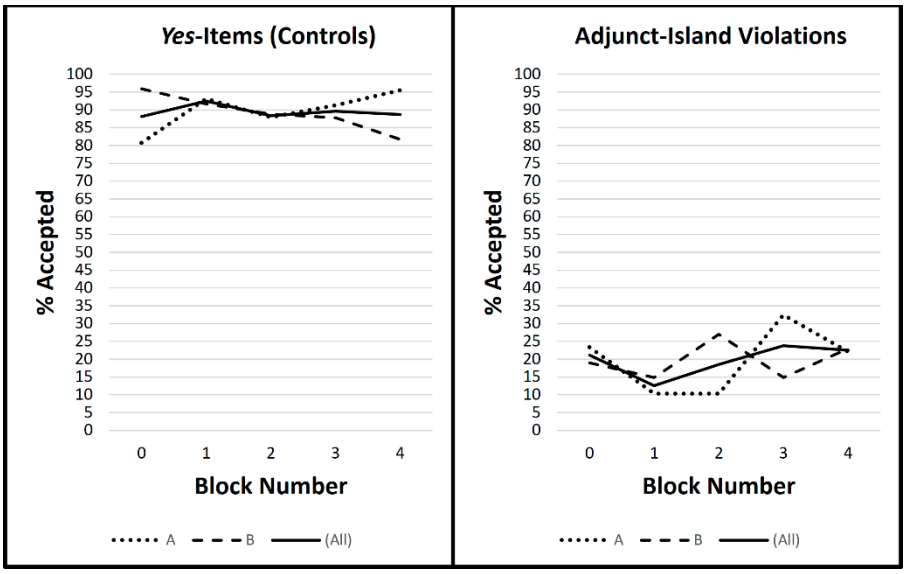

Figure 9. Experiment III, *Yes*-items and Adjunct-Island violations.

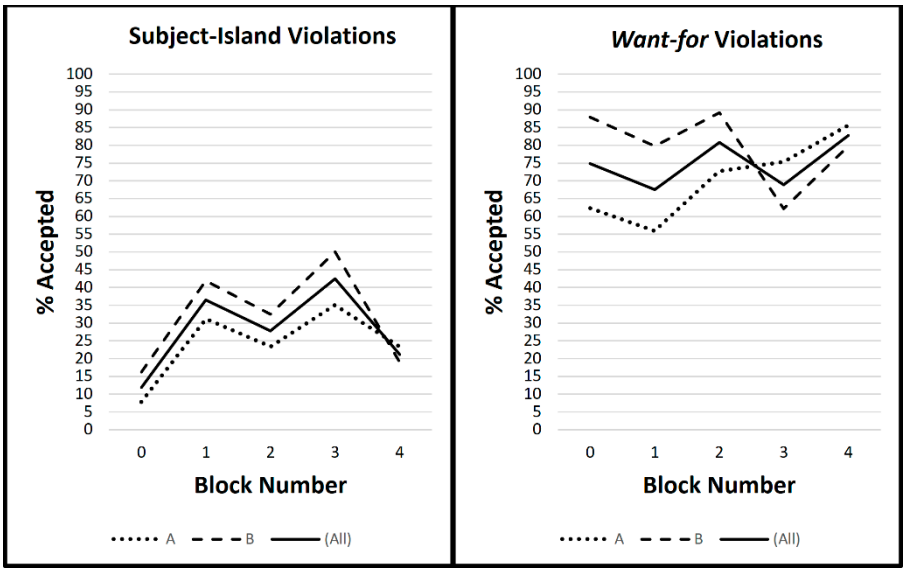

Figure 10. Experiment III, Subject-Island and *Want-for* violations.

5.5. Assessing Effect Size

Finally, a key difference from Experiment I is the clear satiation on CNPC items. The findings of Experiment III actually serve to clarify why this difference exists: in absolute terms, the effect size for CNPC violations was extremely small. With our sample size of

more than 150 participants, we can characterize the effect fairly precisely, and it turns out to have been unrealistic to expect reliable detection of satiation on CNPC items with only 20 participants, as in Experiment I.

In Experiment III, the average acceptance rate for CNPC violations increased from 5% of participants in Blocks 0 and 1 to 12% in Blocks 3 and 4. For present purposes, let's make the (generous) assumption that these rates are a good estimate for the larger population of English-speaking college students. In that case, a simple probability calculation indicates the following. Any participant drawn from this general population and presented with the same materials should have a .013 probability of accepting exactly two more CNPC items in Blocks 3 and 4 than in Blocks 0 and 1. This is because p("No" in Block 0) $\times$ p("No" in Block 1) $\times$ p("Yes" in Block 3) $\times$ p("Yes" in Block 4) = (1 − .05) (1 − .05) (.12) (.12) = .013. Likewise, there should be a .192 probability of accepting one more, a .074 probability of accepting one fewer, and a .00194 probability of accepting two fewer. By power analysis, it follows that N = 76 is the absolute smallest sample size for which the expected frequencies of participants in these four categories (i.e., an expected (76) (.013) = one participant who increases by two, 15 who increase by one, six who decrease by one, and zero who decrease by two) will result in a significant change by Wilcoxon test (at $p < .05$).

The moral is that, even within the set of sentence types that are susceptible to satiation, the strength of the effect may vary as a function of the specific linguistic constraint that is violated. Of the 151 participants in Experiment III, there were 125 who rejected at least one of the two *wonder whether* items in Blocks 0 and 1 and therefore had the possibility of showing increased acceptance (satiation) in Blocks 3 and 4. In fact, 56 of the 125 (45%) satiated. In contrast, 149 of the 151 participants rejected at least one of the CNPC items in Blocks 0 and 1, and only 28 (19%) showed satiation. This raises several questions. For a start, we might ask why, even on *whether* items (where we saw quite a strong satiation effect), fewer than half the participants showed any detectable change. Is this purely a matter of chance, or does susceptibility to the effect relate systematically to some other aspect of an individual's cognitive profile? This would be an interesting question for future research.

At the same time, we might ask whether this dimension of the satiation phenomenon increases its value as a diagnostic tool: perhaps both the susceptibility of a sentence type to satiation in the first place and the strength of the satiation effect observed can provide useful information about the source of the initial unacceptability. This will be taken up again in Section 8.

6. Carryover Effects of Satiation (Experiments I and III)

Recall that (Snyder 2000) found evidence of carryover effects. Restricting attention to those participants who rejected both *wonder whether* items in the first two blocks, the ones who satiated on *wonder whether* (i.e., who accepted at least one of the two exemplars in the final two blocks) were significantly more likely than the others to accept a post-test item involving *ask whether*. Likewise, among those participants who rejected both CNPC items (with *believe the claim*) in the initial two blocks, the ones who accepted at least one CNPC item in the final two blocks were significantly more likely than the others to accept a post-test item involving *accept the idea*. This section presents the corresponding findings from Experiments I and III.[12]

In Experiment I, the participants showed clear satiation on *Whether* Islands (although not on CNPC items), and they judged the same post-test items used in (Snyder 2000). Following the procedure of (Snyder 2000) (described in Section 2), we will restrict our attention to the 15 participants who had rejected both of the *wonder-whether* items in Blocks 0 and 1. Of these 15, three also rejected the *wonder-whether* items in Blocks 3 and 4. Among these three non-satiators, none accepted the post-test item with *ask whether.* In contrast, of the 12 satiators (all of whom accepted at least one of the two *wonder whether* items in Blocks 3 and 4), four (=33%) accepted the post-test item. Hence, the data are consistent with the presence of carryover to the post-test item, although the small numbers make it

difficult to assess the finding statistically. (In particular, the base rate of zero acceptance in the non-satiators makes the use of a binomial test inappropriate.)

When we turn to the data from the larger sample in Experiment III, we find evidence of satiation carryover with both CNPC items and *Whether* Islands. For CNPCs, 139 participants rejected both items in Blocks 0 and 1. Of these 139, 112 were non-satiators (i.e., they also rejected the CNPC items in Blocks 3 and 4), and of these 112, 28 (=25%) accepted the post-test item. In contrast, 15 (=56%) of the 27 satiators accepted it. Hence, there was significant carryover (binomial $p < .005$).

As noted in Section 5, the post-test in Experiment III included an example of *wh*-extraction across *ask whether* (as was also the case in Experiment I), plus one example each for *wonder why* and *know how*. Possible carryover from satiation on *wonder whether* was checked for all three of these post-test items. For *wonder whether*, 79 participants rejected both examples in Blocks 0 and 1. Of these 79, 38 were non-satiators: three (8%) accepted *wonder why*, eight (21%) accepted *know how*, and six (16%) accepted *ask whether*. Of the 41 satiators, 11 (27%) accepted *wonder why*, 10 (24%) accepted *know how*, and nine (22%) accepted *ask whether*. Application of the binomial tests yields $p < .001$ for *wonder why* but $p > .10$ for both *know how* and *ask whether*. The lack of a significant carryover effect for *ask whether* is a departure from (Snyder 2000).

To sum up, in Experiment III (as in Snyder 2000), there was significant carryover from CNPC items involving *believe the claim* to items with *accept the idea*. Yet, the findings for *wh*-islands were more complex. On the one hand, there was significant carryover from *wonder whether* to *wonder why*, which is interesting insofar as it suggests the satiation on *Whether* Islands may be independent of the many special characteristics of the English *wh*-complementizer *whether*. Yet, in Experiment III, there was no significant carryover from *wonder whether* to *ask whether* (as there had been in Snyder 2000 and possibly also in Experiment I), nor was there significant carryover to *know how*. Hence, there is clearly a need for additional research.

Before concluding this section, one final point should be examined. Given that Experiment III yielded evidence of satiation on both *wonder-whether* and CNPC items, we can ask about the relationship between the two within individual participants. Did participants who satiated on one necessarily also satiate on the other? In particular, did individuals in the smaller set of participants who satiated on CNPCs necessarily also accept *wonder-whether* items by the end of the experiment?

The answer is "no". Overall, there were 55 individuals who satiated on *wonder whether* and 29 who satiated on CNPCs. There were only 14 individuals in the intersection. In other words, there were 15 individuals who satiated on CNPCs but showed no increase in their acceptance of *wonder whether*, and there were 41 individuals who satiated on *wonder whether* but showed no increase in their acceptance of CNPCs. Moreover, five of the individuals who satiated on CNPCs actually rejected *wonder whether* entirely (i.e., in both of Blocks 3 and 4). Hence, the satiation effects for these two sentence types appear to be independent.

7. Comparison of Findings across Studies

The work in (Snyder 2000) has given rise to a substantial, growing literature on satiation, which includes new findings from experimental studies (e.g., Hiramatsu 2000; Braze 2002; Francom 2009; Goodall 2011; Crawford 2012; Maia 2013; Christensen et al. 2013; Hofmeister et al. 2013; Chaves and Dery 2014; Do and Kaiser 2017), as well as some efforts to apply these findings to issues in theoretical syntax (e.g., Boeckx 2003, Chapter 3; Stepanov 2007).[13] The present section reviews this literature to assess the consistency of findings across studies. Attention will be focused on studies examining one or more of the same English sentence types studied in (Snyder 2000) (and, hence, Experiments I–III), in order to identify the possible effects of methodological differences across studies.[14] (For other recent surveys of the satiation literature, see Sprouse and Villalta 2021; Snyder 2021.)

7.1. Response Equalization?

First, while Experiment I was the most direct test to date of Sprouse's (2009) REH critique of (Snyder 2000) (as discussed in Section 2, note 7), the conclusions were certainly anticipated by others in the literature. For example, the following four researchers had all reported satiation effects in experiments with a perfect balance of "expected YES" and "expected NO" items: Hiramatsu (2000, Experiment II - henceforth "E2"), Braze (2002), Francom (2009, E2), and Crawford (2012).[15]

7.2. Consistency of Findings and Points of Variation

For the English sentence types examined in Experiments I–III, the results are very much in-line with other studies in the literature. (A synopsis is provided in Tables 4 and 5.) A clear majority detect satiation on argument extraction from *Whether* Islands.[16,17] Sprouse's (2009) A3 and B1 and 2 are exceptions and will be discussed in Section 7.3.1.[18]

Table 4. Sentence types on which satiation has been induced experimentally. ("ME" = magnitude estimation; "Y/N" = yes/no task.).

Sentence Type	Satiation?	Experiment	Context Sentence?	N	Exposures	Task
Whether Island	Yes	(Braze 2002)	Yes	35	9	Y/N
	Yes	(Crawford 2012)	Yes	22	7	Scale
	Yes	(Francom 2009: E1)	No	205	5	Y/N
	Yes	(Hiramatsu 2000: E1)	Yes	33	7	Y/N
	Yes	(Hiramatsu 2000: E2)	Yes	11	7	Y/N
	Yes	(Snyder 2000)	Yes	22	5	Y/N
	Yes	*(Experiment I)*	Yes	20	5	Y/N
	Yes	*(Experiment III)*	Yes	151	5	Y/N
	No	(Sprouse 2009: A3)	No	20	10	ME
	No	(Sprouse 2009: B1)	No	25	5	Y/N
	No	(Sprouse 2009: B2)	No	19	5	Y/N
CNPC	Yes	(Goodall 2011)	Yes	45	5	Y/N
	Yes	(Snyder 2000)	Yes	22	5	Y/N
	Yes	*(Experiment III)*	Yes	151	5	Y/N
	No	(Hiramatsu 2000: E1)	Yes	33	7	Y/N
	No	(Sprouse 2009: A4)	No	17	10	ME
	No	(Sprouse 2009: A5)	Yes	20	10	ME
	No	(Sprouse 2009: B1)	No	25	5	Y/N
	No	(Sprouse 2009: B2)	No	19	5	Y/N
	No	*(Experiment I)*	Yes	20	5	Y/N
Subject Island	Yes	(Chaves and Dery 2014: E1)	No	60	20	Scale
	Yes	(Chaves and Dery 2014: E2)	No	55	14	Scale
	Yes	(Francom 2009: E1)	No	205	*5*	Y/N
	Yes	(Francom 2009: E2)	No	22	8	Y/N
	Yes	(Hiramatsu 2000: E1)	Yes	33	7	Y/N
	Yes	(Hiramatsu 2000: E2)	Yes	11	7	Y/N
	No	(Crawford 2012)	Yes	22	7	Scale
	No	(Goodall 2011)	Yes	45	5	Y/N
	No	(Snyder 2000)	Yes	22	5	Y/N
	No	(Sprouse 2009: A1)	No	20	14	ME
	No	*(Experiment I)*	Yes	20	5	Y/N
	No	*(Experiment III)*	Yes	151	5	Y/N

Table 5. Sentence types on which studies have consistently failed to induce satiation. ("ME" = magnitude estimation; "Y/N" = yes/no task.).

Sentence Type	Satiation?	Experiment	Context Sentence?	N	Exposures	Task
Adjunct Island	No	(Braze 2002)	Yes	16	9	Y/N
	No	(Crawford 2012)	Yes	22	7	Scale
	No	(Francom 2009: E1)	No	205	5	Y/N
	No	(Francom 2009: E2)	No	22	8	Y/N
	No	(Goodall 2011)	Yes	45	5	Y/N
	No	(Hiramatsu 2000: E1)	Yes	33	7	Y/N
	No	(Hiramatsu 2000: E2)	Yes	11	7	Y/N
	No	(Snyder 2000)	Yes	22	5	Y/N
	No	(Sprouse 2009: A2)	No	24	14	ME
	No	(Sprouse 2009: B1)	No	25	5	Y/N
	No	(Sprouse 2009: B2)	No	19	5	Y/N
	No	*(Experiment I)*	Yes	20	5	Y/N
	No	*(Experiment III)*	Yes	151	5	Y/N
Left Branch	No	(Francom 2009: E1)	No	205	5	Y/N
	No	(Francom 2009: E2)	No	22	8	Y/N
	No	(Goodall 2011)	Yes	45	5	Y/N
	No	(Hiramatsu 2000: E1)	Yes	33	7	Y/N
	No	(Hiramatsu 2000: E2)	Yes	11	7	Y/N
	No	(Snyder 2000)	Yes	22	5	Y/N
	No	(Sprouse 2009: B1)	No	25	5	Y/N
	No	*(Experiment I)*	Yes	20	5	Y/N
	No	*(Experiment III)*	Yes	151	5	Y/N
That-trace	*(See text)*	(Hiramatsu 2000: E1)	Yes	33	7	Y/N
	No	(Francom 2009: E1)	No	205	5	Y/N
	No	(Francom 2009: E2)	No	22	8	Y/N
	No	(Goodall 2011)	Yes	45	5	Y/N
	No	(Snyder 2000)	Yes	22	5	Y/N
	No	*(Experiment I)*	Yes	20	5	Y/N
	No	*(Experiment III)*	Yes	151	5	Y/N
Want-for	*(See text)*	(Hiramatsu 2000: E1)	Yes	33	7	Y/N
	(See text)	(Francom 2009: E1)	No	205	5	Y/N
	No	(Snyder 2000)	Yes	22	5	Y/N
	No	*(Experiment I)*	Yes	20	5	Y/N
	No	*(Experiment III)*	Yes	151	5	Y/N

Two other sentence types have sometimes, but not always, shown satiation: CNPCs and Subject Islands. Studies testing CNPCs include Goodall 2011, which found clear satiation, as well as several others that did not.[19] As discussed in Section 5, Experiment III sheds considerable light on this variability; it appears the effect size for CNPCs is much smaller than for *Whether* Islands. Without a sizable number of participants (a bare minimum of $N = 76$, it seems, for the specific design and materials used in Experiment III), there is a high probability of failing to detect satiation on CNPCs (i.e., even if some degree of satiation is occurring). As seen in Table 4, two of the three experiments finding significant satiation on CNPCs (including Experiment III above) had at least 40 participants, while those not finding it all had fewer than 40.

Note that (Sprouse 2009) included four experiments trying to induce satiation on CNPCs, each with 25 or fewer participants. Individually, these experiments probably had little chance of succeeding, but overall (with 81 participants in total), the chances of detecting it (at least once) were perhaps not so bad. The larger issue may have been that the stimuli in all but one of these experiments (A5) omitted the context sentence, with the result that there was no clear indication of the intended meaning. This looks like it might have been a critically important change, because, in Table 4, the experiments that succeeded all provided context sentences. If so, the fact that only A5 included context sentences, together with the fact that A5 had only 20 participants, may explain the absence of satiation in Sprouse's experiments.

Turning to Subject Islands, ten of the other studies in Table 4 tested argument extraction from DPs in the subject position (especially DPs that were underlyingly direct objects, as

with passives and unaccusatives). Six found satiation, and four did not. One problem in some of the latter studies may have been an insufficient number of exposures. Almost all the studies finding satiation (five out of six) increased the number of exposures beyond the original five in (Snyder 2000).[20] In fact, Hiramatsu (2000, E1), who employed seven exposures, noted that the satiation evident in Block 7 was not yet detectable in Block 5.[21]

For other sentence types examined in Experiments I–III above, the absence of detectable satiation is also largely consistent with the literature (see Table 5). For the LBC, seven other studies tested for satiation, and none found it. Adjunct Islands were checked in eleven other studies, and again, none found satiation.

That-trace violations have not in general shown satiation, but more needs to be said. Sprouse (2009, p. 331, Table 2) characterized (Hiramatsu 2000) as having found satiation on *That*-trace, but the situation was unclear. Hiramatsu (p. 107) expressed concerns about the quality of her data for *That-trace* and *Want-for* (which were tested only in E1). She reported that multiple participants had eventually begun crossing out the word *that* or *for* and then marking "Yes". Moreover, on p. 111, she seems to disavow the data for these sentences altogether: "As we saw in the previous section, we do not have a clear picture of the results for [...] *That-trace* and *Want-for* sentences." Hence, the cautious approach would be to set those findings aside, and the other studies of *That-trace* in Table 5 found no satiation.

In the case of *want-for*, once we set aside (Hiramatsu 2000), the main data in the literature (aside from Snyder 2000, which found no satiation, and Experiments I–III above, which likewise found no satiation) come from Francom (E1), who does report satiation. As it happens, Francom employed Snyder's (2000) method of counterbalancing the order of presentation. He did not originally provide information about the consistency of responses across the two orders, but he very kindly shared his data. This made it possible to check whether the change in acceptance was comparable across the two versions.

As it turned out, the evidence for satiation on *want-for* did not satisfy this criterion. Collapsing across the two versions, acceptance increased from 75% in Blocks 1 and 2 to 83% in Blocks 4 and 5 ($W = -856$, $n_{s/r} = 61$, $Z = -3.08$, $p < .005$). Yet, this change was driven almost entirely by participants receiving Version B.[22] The acceptance on Version B shifted from 71% to 83% ($W = -390$, $n_{s/r} = 36$, $Z = -3.06$, $p < .005$), but the acceptance on Version A went only from 79% to 83% ($W = -86$, $n_{s/r} = 25$, $Z = -1.15$, $p > .10$ NS). Hence, the increased acceptance of *want for* at the end of Francom's experiment was probably due to an accidental property of the presentation order in Version B. By the criteria used in Experiments I–III (specifically, the requirement for the effect to be present in both orders of presentation), the findings do not qualify as reliable evidence of satiation.

In sum, across the studies reviewed here, the sentence types showing a satiation effect have consistently been some combination of *Whether*-Island, CNPC, and Subject-Island violations. At least by the criteria employed here, no study has yielded reliable evidence of satiation on Adjunct-Island, Left-Branch, *That*-trace, or *Want-for* violations.

7.3. Points of Variation in Method

7.3.1. Experimental Set-Up

Studies attempting to induce satiation have varied somewhat in their experimental set-up. As can be seen in Tables 4 and 5, one potentially important variable is whether a context sentence was provided. The studies that included a context sentence have mostly succeeded in inducing satiation, at least for *Whether* Islands, but the results have been less consistent when it was omitted.[23]

Note that providing a context sentence is one way of conveying the intended meaning of a sentence. Arguably, judgments of linguistic acceptability are always (at least implicitly) relative to an interpretation. For example, the acceptability of the English sentence *John likes him* depends critically on whether *him* is taken to mean the same person as *John*; hence, referential indexing is provided in the literature on binding theory. In other cases, one does find linguists simply placing an asterisk on a sentence without specifying an intended meaning, but in practice, this appears to mean one of two things. Either the sentence is

unacceptable on what the linguist takes to be the "obvious" interpretation or the linguist believes the sentence is unacceptable no matter what the intended meaning is. Thus, in an experimental study of linguistic acceptability, one possible effect of including a context sentence is simply facilitation of the judgment task by making it easier for the participant to identify an intended meaning when making the judgment.

Yet, as suggested by an anonymous reviewer, the inclusion of a context sentence (and thus, clarification of the intended meaning) might have an additional, quite important role that would be specific to an experiment on satiation effects. This is because it helps the participant identify one particular way of parsing the test sentence. As will be discussed in Section 8, there could be a number of relevant consequences. For one, having this information could lead a probabilistic parser to increase the expected probability of an uncommon parse (e.g., in the context of *wh*-extraction from a subject island, the probability of a parse positing a gap inside the subject of an embedded clause). Another effect could be helping the participant recognize that adopting a "marked" syntactic option will render the sentence grammatically possible (e.g., in the context of *wh*-extraction from a CP inside an NP, adopting the option—which is potentially a marked option—of treating the CP as a complement to N rather than simply an N-bar adjunct). Thus, there are good reasons to expect that the inclusion of a context sentence might facilitate satiation and, moreover, that the facilitation might apply to certain sentence types more than others.

Another salient point of variation across different satiation studies is the nature of the judgment task: Does the participant provide a Yes/No judgment, a rating on a numerical scale, or an estimate of magnitude? Most studies that successfully induced satiation employed a Yes/No task, although Crawford 2012; Chaves and Dery 2014 (E1–2) employed a numerical scale. Sprouse 2009 (A1–5) differed in choosing magnitude estimation (ME). At present, it is unclear whether the choice of task affects the findings for satiation—and, if so, why this would be the case. (For a recent discussion of the task characteristics of ME, see Featherston 2021 and the references therein.) What is needed is a side-by-side comparison of these methods within a single satiation study.

As already noted, two other variables appear to be critically important: the number of exposures to each sentence type and the number of participants in the study. The information in Table 4 suggests that satiation on Subject Islands is difficult to obtain, unless the number of exposures is at least seven, and that satiation on CNPCs is likewise difficult to obtain, unless the number of participants is substantial (at bare minimum 76 for the specific materials and design in Experiment III). These points will be taken up again in Section 8.[24]

7.3.2. Variation in the Stimuli

Another salient point of variation concerns the detailed syntax of the test sentences. For example, Hiramatsu 2000 contrasted two types of Subject-Island violations, involving extraction from a subject DP that was either the underlying object or the underlying subject of a transitive verb. Interestingly, she found satiation only with underlying objects. In a similar vein, she contrasted the extraction of arguments versus adjuncts from a *Whether* Island and found satiation only for arguments.

7.3.3. Variation in Data Analysis

Studies have varied in their statistical methods, but the differences seem to be immaterial. Francom (2009, pp. 32–35) applied sign tests, paired *t*-tests, ANOVA, and logistic regression, with identical results. Likewise, the datasets from Experiments I–III in this paper were analyzed both with ME logistic regression and with a more traditional method (the Wilcoxon Signed-Rank Test), and the results were effectively identical.

In contrast, what is clearly of great importance is confirming that one's data are internally consistent: Do we see the consistency across orders of presentation that we should for an effect at the level of grammatical structure? As illustrated in Experiment III, this common-sense check can have a critically important influence on the conclusions

drawn. Furthermore, in Section 7.2, they helped eliminate an apparent conflict across studies in the findings for *want-for*.

8. Directions for Future Research

8.1. Satiation as a Diagnostic Test

We now return to the question of how investigating satiation could benefit generative linguistics. First, a number of potentially valuable ways to apply our current knowledge of satiation might follow from a proposal made by Goodall (2011, p. 35):

> [I]f one unacceptable sentence type is satiation-inducing and another is not, it is unlikely that their unacceptability is attributable to the same underlying principle. This suggests, for instance, that violations of *whether* islands, which are susceptible to satiation, and *that*–trace violations, which are not, must be due to different underlying principles, in accord with the general consensus in the literature about these two phenomena.

Following this line of reasoning, and incorporating the findings discussed in this article, one can see a number of immediate applications. Whenever a linguistic theory (be it a theory in syntax, semantics, or morphophonology) posits a single source for the unacceptability of two different sentence types X and Y, testable predictions immediately follow.

For example, one possibility will be to run a pair of studies modeled on Experiment III. In one of the studies, we add a single example of sentence type X to each block. In the second study, we use examples of Y in place of X. Upon completion, we check to see if X and Y are alike (or disparate) in whether their initial unacceptability satiates. If satiation is present for both, we can also check whether the number of exposures required for satiation is comparable across X and Y, and we can check whether the percentage of participants who show a change in their judgment is comparable across X and Y.

If the satiation findings for X and Y are either highly similar, or highly dissimilar, the interpretation will be straightforward. More complex (and, no doubt, more interesting) will be the intermediate cases, where some of the diagnostics come out as expected under the hypothesis of a single source of unacceptability and others do not. This sort of mixed case might, for example, indicate that X and Y overlap only partially in the factors rendering them (initially) unacceptable.

Yet, there are a number of ways for the ideas just sketched to be too simplistic. In particular, there is a tacit assumption that the underlying source of unacceptability is the thing undergoing change. Suppose, for example, that a specific UG constraint on syntactic movement is what is rendering both X and Y unacceptable. If this constraint is somehow weakened by satiation, then both X and Y should become more acceptable. However, suppose that the UG constraint is immune to satiation and that something else is changing. For example, perhaps a speaker can learn to reanalyze structure X as a superficially similar but syntactically distinct structure X-prime, to which the UG constraint does not apply. If the reanalysis operation depends on surface characteristics that are present on X but absent from Y, only X will be able to satiate, even though the cause of the initial unacceptability of X and Y is exactly the same.

8.2. Explaining Satiation

Before we try to use satiability as a diagnostic, we will naturally want to know as much as we can about what exactly satiation is. A logical starting point is to ask whether satiation is a unitary phenomenon. Is there essentially the same process at work in every example of a sentence type that satiates (according to the operational definition of satiation in 3)? Alternatively, are there different mechanisms at work in different sentence types?

The findings in this article can at least help narrow down the possible answers. Consider the following "strongly unitary" scenario:

> ***Scenario 1.*** Suppose that a kind of "mental alarm" goes off whenever a person's language-processing mechanisms are forced to postulate a grammatically deviant structure for a linguistic expression. Let's assume that the alarm system is highly

> similar from one speaker to another; the strength of the alarm varies along a single, smoothly continuous dimension, and violations of different grammatical constraints all trigger the same alarm, although the strength of the resulting alarm signal may vary with the type of violation. If so, satiation could perhaps be a kind of habituation effect: perhaps repeatedly experiencing a certain level of alarm, over a certain period of time, can make one tolerant.

Under Scenario 1, no matter which sentence types undergo satiation, the mechanism is exactly the same: habituation to alarm signals of a certain magnitude. Grammatical constraints associated with a weak signal should always satiate prior to constraints with a stronger signal. Indeed, satiation on a constraint with a strong signal should yield satiation not only on sentences violating that particular constraint but also on sentence types yielding weaker signals, even if those sentence types violate different constraints and even if those sentence types have never actually been encountered.

The evidence presented in this article speaks against an account along these lines. Specifically, the fact that satiation on CNPC violations in Experiment III was found in a much smaller percentage of participants than satiation on *wonder whether* more or less forces us to conclude, under Scenario 1, that CNPC violations elicit a "louder" alarm signal than *wonder whether* violations. Hence, there is a strong prediction that every single individual who satiated on CNPC violations by the end of Experiment III must have ended up accepting *wonder whether* items as well. At the end of Section 6, however, it was noted that five of the 29 participants who satiated on CNPC violations actually rejected both of the *wonder whether* items in Blocks 3 and 4.

In place of a strongly unitary account, one might consider a "weakly unitary" account along the following lines:

> ***Scenario 2.*** Suppose the language processor has a number of distinct alarm signals, each of which indicates the violation of a different grammatical constraint. In this case we might once again imagine that satiation results from habituating to an alarm signal (and, hence, that satiation is unitary in a certain sense), but now, satiation will proceed independently for different grammatical constraints (i.e., as a separate process of habituation for each of several different alarms). Satiation on a given constraint will require exposure to sentences violating that particular constraint.

Note that, under Scenario 2, the number of exposures required before full habituation occurs might still vary as a function of the constraint in question if (for example) some constraints have "louder" alarms than others.

Is Scenario 2 compatible with the evidence from Experiment III? This depends on our assumptions. If we assume—as seems fair—that, prior to the experiment, the participants had no exposure to either CNPC violations or *wonder whether* violations, and if we assume that each exposure during the experiment is equally effective at promoting habituation, then the same prediction that defeated Scenario 1 will probably exist for Scenario 2. Specifically, by the end of the experiment, every participant will have encountered the same number of CNPC violations and *wonder whether* violations; if that number (of CNPC violations) is sufficient to create habituation on the alarm signal for CNPCs (again assuming that these are the more difficult sentence type to satiate), then the same number (of *wonder whether* violations) should be sufficient to produce habituation on the (distinct, but weaker) alarm signal for *wonder whether*.

Yet, the prediction will change if we assume, for example, that habituation to a given alarm signal requires not only some number of encounters with relevant examples but also some particular internal state in the experimental participant (perhaps something like introspective awareness) that fluctuates from moment to moment. In this case, it might be possible, simply by chance, for a participant to have "genuinely" experienced a smaller number of alarm signals for *wonder whether* violations by the end of the experiment than alarm signals for CNPC violations.

In any case, a complete alternative to Scenarios 1 and 2 would be Scenario 3, which sketches a strongly non-unitary model.

> ***Scenario 3.*** Suppose that different satiable constraints may owe their satiability to different mechanisms. Perhaps, in some cases, satiation results from habituation to a particular constraint's alarm signal, but in other cases, it results from, say, discovering an alternative syntactic analysis of a particular sentence type. For example, perhaps CNPC violations involving *wh*-extraction across *...believe the claim that...* are usually assigned an "unmarked" structure in which the CP is treated as an appositive (i.e., an N-bar adjunct), but UG also permits another, more marked analysis (at least for epistemic nominals, like *claim* and *idea*) in which the CP is a complement selected by N. In terms of Chomsky's (1986) *Barriers* system, the appositive analysis forces the *wh*-phrase to cross two barriers (the lower CP, which is not L-marked, as well as the NP above it, which is a barrier by inheritance). In contrast, no barrier will be crossed if the lower CP is selected by the N. Hence, in this case, satiation is not habituation but, rather, the discovery of a new, UG-compatible (but "marked") parse, which (by hypothesis) was not being exploited before.

Under Scenario 3, it is perhaps less surprising (than under Scenarios 1 and 2) to find participants who have satiated on CNPC violations but who still firmly reject *wonder whether* violations. If satiation on CNPC violations results from a sudden (tacit) insight into UG-compatible structures but satiation on *wonder whether* violations results (say) from the gradual accumulation of a particular volume of experience over time, then an individual can easily satiate on one and not the other.[25]

At this point, it is interesting to note that Chaves and Dery (2018) proposed an explicit model of satiation on Subject Island (SI) violations, and their model seems to be far more compatible with Scenario 3 than Scenarios 1 and 2. This is because their work does not address satiation on sentence types other than SIs, and the proposed mechanism of satiation appears to be specific to SIs. In brief, Chaves & Dery argued that SI violations are not ungrammatical but merely difficult to parse. They assumed the parsing difficulty results from "the fact that subject-embedded gaps are pragmatically unusual—as the informational focus does not usually correspond to a dependent of the subject phrase—and are therefore highly contrary to comprehenders' expectations about the distribution of filler gap dependencies" (Chaves and Dery 2018, p. 1). In their view, comprehenders' expectations can change fairly rapidly with exposure to clear examples of subject-embedded gaps.

Thus, the Chaves–Dery mechanism seems like a plausible candidate for a source of satiation that is specific to SI violations. Let's suppose this proposal is correct for SI violations. Then, as suggested above, perhaps satiation on CNPC violations will turn out to involve discovering a new, UG-compatible (but ordinarily nonpreferred) parse for a CP following an epistemic nominal. Perhaps satiation on extraction from certain *wh*-islands will turn out to involve habituating to a mental "alarm" triggered by a certain type of grammatical violation. This type of non-unitary scenario leads to distinctive predictions, such as the strict absence of satiation carryover effects between sentences of these three types. Experimental tests of such predictions would be a reasonable next step for research into the nature of satiation.

In conclusion, Experiments I–III have provided evidence (i) that experimentally induced satiation, like the satiation that sometimes affects linguists, is restricted to a small, stable set of sentence types; (ii) that, after satiation on one sentence type (e.g., *wh*-movement across *...wonder whether...* or *...believe the claim...*), acceptability sometimes increases for distinct but syntactically related sentence types (such as *...wonder why...* or *...accept the idea...*); (iii) that, for sentence types susceptible to satiation, the difficulty of inducing it (e.g., number of exposures required) varies systematically; and (iv) that, much as satiation in linguists persists over time, experimentally induced satiation (at least in the case of *wonder whether*) can persist for at least four weeks. These findings may suggest an eventual role for satiation in determining whether the perceived unacceptability of two distinct

sentence types has a common source, but more immediately, they suggest that satiation may be a powerful tool for examining the tacit mental operations that are responsible for our judgments of linguistic (un)acceptability.

Funding: This work was supported by the National Science Foundation under NSF IGERT DGE-1144399 and grant DGE-1747486.

Institutional Review Board Statement: This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of the University of Connecticut.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Anonymized data are available from the author.

Acknowledgments: The author is grateful for helpful discussions with (among others) Dave Braze, Rui Chaves, Jean Crawford, Jerrod Francom, Grant Goodall, Kazuko Hiramatsu, and Whit Tabor. The author also thanks the anonymous reviewers for their numerous astute suggestions.

Conflicts of Interest: The author declares no conflict of interest. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Notes

1 The experimental findings published in (Snyder 2000) were first presented as part of (Snyder 1994).

2 The constraints operating in (6a–d,g) were first characterized in (Ross 1967). Early discussions of *That*-trace and *Want-for* effects (6e,f) can be found in (Perlmutter 1968) and (Rosenbaum 1967), respectively.

3 One might be surprised by the the reported lack of satiation on *that*-trace violations, given Sobin's (1987, et seq.) claim that certain varieties of English lack this constraint altogether. Yet, one needs to proceed with caution here, because more recent work has called Sobin's claim into question (see, in particular, Chacón 2015; Cowart and McDaniel 2021).

4 Note that Snyder (2000) measured satiation on a given sentence type by examining (i) the number of participants who accepted more tokens at the end (i.e., final two blocks) of the experiment than at the beginning (initial two blocks) and (ii) the number who accepted more tokens at the beginning than the end. Participants who accepted equal numbers of tokens at the beginning and end were set aside. The assumption was that random variability in judgments is equally likely to create an increase or a decrease. Satiation was detected, for a given sentence type, when there was a statistically significant preponderance (among those whose rate of acceptance changed) of participants who accepted more tokens at the end. Statistical significance was assessed by means of a two-tailed binomial test based on the null hypothesis that increases and decreases each have a 0.5 probability.

5 In (Snyder 2000, p. 579) some of the numbers reported for the carryover effects in *Whether* Islands and CNPC violations were transposed. In the calculations reported here, these errors have been corrected, and the impact is minimal: the two carryover effects that were reported as statistically significant in (Snyder 2000) remain significant after the corrections.

6 These considerations will play a critical role in the approach to data analysis in Experiments I–III below.

7 To foreshadow the findings, the REH will receive no support in Experiment 1. Indeed, as an anonymous reviewer pointed out, there has never been any evidence directly supporting the REH (even in Sprouse 2009). The experiments presented in (Sprouse 2009) had made the change (from Snyder 2000) of perfectly balancing the number of expected yes versus expected no answers, and the participants did not show satiation. The REH was proposed as an explanation for this difference in results, but surprisingly, Sprouse did not report the natural follow-up study of increasing the number of expected "no" answers and showing that the (apparent) satiation of (Snyder 2000) resurfaced. Moreover, as will be discussed in Section 7, Sprouse made additional changes to Snyder'smethodology that could plausibly account for the replication failure. Thus, Experiment 1 is the first direct test of the REH as an explanation for the findings of (Snyder 2000); all aspects of the experiment are exactly the same as in (Snyder 2000) except for a perfect balance of expected "yes"and expected "no" items.

8 Participants, on average, said "yes" to 90% of the grammatical items (standard deviation 8.2). The excluded participants had each answered "yes" to approximately 70%.

9 An anonymous reviewer noted that the acceptance rates for "yes" items, shown in the right panel of Figure 1, are not at the ceiling but, rather, vary somewhat across the different blocks of the experiment. Importantly, this variability will be controlled for statistically in the logistic regression model described in Section 3.6. As noted in Section 3.2, ME logistic regression will be conducted with one level of each factor specified as a baseline for use in "treatment contrasts" (i.e., pairwise comparisons) with each of the other levels of that factor. For Type, the baseline level will be "Good" (i.e., within each block, the results for the seven fully grammatical "yes" items). Hence, whatever rate of acceptance the "yes" items might receive in a given block, that same rate will be the comparison point for each of the other sentence types in that block.

[10] Specifically, values of the theta parameters for the RE structure were obtained by executing getME(Data.model,"theta"), and this revealed a value of zero for Version.(Intercept).

[11] Note: throughout this article, *p*-values are two-tailed.

[12] Experiment II is being set aside here, because the participants had already seen the same "post-test" items earlier in Experiment I.

[13] Much, but not all, of this work focuses on English. Goodall (2011) performed a cross-linguistic comparison of satiation in English versus Spanish (although the findings reviewed here will be limited to his English data). Christensen et al. (2013) detected possible satiation on *wh*-islands in Danish. Maia (2013) discussed a study that he conducted with Wendy Barile finding satiation on *wh*-in-situ (within islands) in Brazilian Portuguese. Interestingly, this final study compared judgments of undergraduates who had recently completed a syntax course covering island effects versus students who had never studied linguistics.

[14] Due to the decision to focus on satiation studies looking at the same sentence types examined in (Snyder 2000), some work on another sentence types, namely superiority violations, will receive only the following remarks. Briefly, Hofmeister et al. (2013) for English and Brown et al. (2021) for both German and English reported increased acceptance of superiority violations after multiple judgments. Yet, there are some reasons to be skeptical that their findings resulted from "syntactic satiation" in the sense intended here—in other words, from the type of satiation reported anecdotally among professional linguists. First, Brown et al. reported that the change in acceptance occurred very rapidly (appearing on the second exposure); second, for both studies, the increase in acceptance was slight, in absolute terms; and third, Brown et al. reported that the same slight increase was found across a range of anomalous sentences that varied in their grammatical structure but that were similar in initially having an intermediate level of acceptability. For the benefit of future investigations, one other point bears mentioning: in both studies, the researchers apparently decided to omit the context sentences of (Snyder 2000). As will be discussed below, the same change may have been responsible for the absence of an expected satiation effect in some of the studies that are reviewed here.

[15] Note that Braze (2002) argued for the existence of a counterpart to satiation in sentence-processing based on an eye-tracker study that he ran in conjunction with an off-line judgment study (with different participants). The findings cited in this section come from the offline study. Yet, if Braze is correct about the sentence-processing counterpart, it both speaks against an account in terms of response equalization (since no judgment of acceptability was elicited) and also has important implications for the project of explaining satiation. (Some related topics will be taken up in Section 8.)

[16] In the case of (Hiramatsu 2000), the discussion here concerns only the results from participants who met her stated inclusionary criterion, namely answering at least 90% of the filler and control items, as expected.

[17] Francom (2009) included a total of five experimental studies, but E1 and 2 seemed to be the most directly comparable with the other studies in this section and, hence, will be the focus.

[18] To disambiguate, Experiments 1–5 from Section 3 of (Sprouse 2009) are prefixed with an "A", and Experiments 1 and 2 from Section 4 are prefixed with a "B".

[19] Francom (2009) (E1 and 2) also reported results for "CNPC violations", but he included a wide range of sentence types under this label, as can be seen from the stimulus lists that he provides in the appendices (pp. 103, 105, 108). Given that the relevance is unclear for "CNPC violations" in the sense intended here, those studies are omitted from the CNPC section of Table 4.

[20] Note that Sprouse, in A1, increased Subject-Island exposures from 5 to 14 but omitted the context sentences. Either this lack of context sentences or his use of a Magnitude Estimation task could (in principle) be responsible for the absence of a satiation effect.

[21] Aside from Sprouse's A1, most studies providing seven or more exposures found satiation on Subject Islands, at least when the subject DPs were underlying objects. The exception is (Crawford 2012), where the 22 participants received seven blocks of exposure (together with context sentences), but no satiation was detected, even for extraction from the subjects of unaccusatives.

[22] To maintain consistency with the earlier discussion of (Snyder 2000) and Experiments I–III, Francom's versions 1 and 2 will be referred to here as A and B, respectively.

[23] One example of success without the use of context sentences is the satiation on Subject Islands in (Chaves and Dery 2014, E1 and 2). Some features of their study that could (in principle) be relevant include the large number of exposures (14–20), the lack of variety in the stimuli (all of the initially unacceptable items had *wh*-extraction from a Subject Island), and the fact that half of the test items employed the D-linked form *which*.

[24] Note that, in principle, there might, or might not, be a simple trade-off between the number of participants in a study and the number of exposures to a given sentence type. If the effect of increasing the number of exposures is a linear increase in the percentage of participants who experience satiation, then perhaps even a small number of exposures will yield a detectable satiation effect if the number of participants is sufficiently large. Alternatively, a given sentence type might turn out to require a minimum number of exposures before any change occurs, no matter how many people participate. The findings from Experiment 3 perhaps favor the latter scenario, since, even with 151 participants, there was no sign of satiation on Subject Islands within the space of five exposures.

[25] As brought to my attention by an anonymous reviewer, there are predictions about the length of time for which satiation will persist that follow directly from different proposed mechanisms. For example, we might reasonably expect indefinite persistence when satiation is due to a "learning" effect (as in my suggestion about satiation on CNPC violations, sketched under Scenario 3). The appropriate predictions could well be different, however, for a mechanism akin to sensory habituation (as sketched under

Scenarios 1 and 2), and they might be different still for the mechanism that Chaves and Dery proposed for Subject Islands (in terms of changing the probability associated with a given parse in a probabilistic parser). In fact, as noted at the end of Section 4.3, the persistence of satiation on *wonder whether* in Experiment II is strongly suggestive of a learning effect rather than the sort of habituation to an alarm signal (i.e., for subjacency violations or the like) suggested in Section 8.2.

References

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68: 255–78. [CrossRef] [PubMed]

Bates, Douglas, Martin Maechler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–148. [CrossRef]

Boeckx, Cedric. 2003. *Islands and Chains: Resumption as Stranding*. Amsterdam: John Benjamins.

Braze, Forrest David. 2002. Grammaticality, Acceptability, and Sentence Processing: A Psycholinguistic Study. Doctoral dissertation, University of Connecticut, Storrs, CT, USA.

Brown, J. M. M., Gisbert Fanselow, Rebeccah Hall, and Reinhold Kliegl. 2021. Middle ratings rise regardless of grammatical construction: Testing syntactic variability in a repeated exposure paradigm. *PLoS ONE* 16: e0251280. [CrossRef] [PubMed]

Chacón, Dustin A. 2015. Comparative Psychosyntax. Doctoral dissertation, University of Maryland, College Park, MD, USA.

Chaves, Rui P., and Jeruen E. Dery. 2014. Which subject islands will the acceptability of improve with repeated exposure? Paper presented at the 31st West Coast Conference on Formal Linguistics, Tempe, Arizona, USA, 8 February 2013; Edited by Robert E. Santana-LaBarge. Somerville: Cascadilla Proceedings Project, pp. 96–106.

Chaves, Rui P., and Jeruen E. Dery. 2018. Frequency effects in Subject Islands. *Journal of Linguistics* 55: 475–521. [CrossRef]

Chomsky, Noam. 1986. *Barriers*. Cambridge: MIT Press.

Christensen, Ken Ramshøj, Johannes Kizach, and Anne Mette Nyvad. 2013. Escape from the island: Grammaticality and (reduced) acceptability of wh-island violations in Danish. *Journal of Psycholinguistic Research* 42: 51–70. [CrossRef] [PubMed]

Cowart, Wayne, and Dana McDaniel. 2021. The That-trace Effect. In *The Cambridge Handbook of Experimental Syntax (Cambridge Handbooks in Language and Linguistics)*. Edited by Grant Goodall. Cambridge: The Cambridge University Press, pp. 258–77.

Crawford, Jean. 2012. Using syntactic satiation effects to investigate subject islands. Paper presented at the 29th West Coast Conference on Formal Linguistics, Tucson, AZ, USA, April 24; Edited by Jaehoon Choi, E. Alan Hogue, Jeffrey Punske, Deniz Tat, Jessamyn Schertz and Alex Trueman. Somerville: Cascadilla Proceedings Project, pp. 38–45.

Do, Monica L., and Elsi Kaiser. 2017. The relationship between syntactic satiation and syntactic priming: A first look. *Frontiers in Psychology* 8: 1851. [CrossRef] [PubMed]

Featherston, Sam. 2021. Response methods in acceptability experiments. In *The Cambridge Handbook of Experimental Syntax (Cambridge Handbooks in Language and Linguistics)*. Edited by Grant Goodall. Cambridge: The Cambridge University Press, pp. 39–61.

Francom, Jerid Cole. 2009. Experimental Syntax: Exploring the Effect of Repeated Exposure to Anomalous Syntactic Structure–Evidence from Rating and Reading Tasks. Doctoral dissertation, University of Arizona, Tucson, AZ, USA.

Goodall, Grant. 2011. Syntactic satiation and the inversion effect in English and Spanish *wh*-questions. *Syntax* 14: 29–47. [CrossRef]

Hiramatsu, Kazuko. 2000. Accessing Linguistic Competence: Evidence from Children's and Adults' Acceptability Judgments. Doctoral dissertation, University of Connecticut, Storrs, CT, USA.

Hofmeister, Philip, T. Florian Jaegger, Inbal Arnon, Ivan A. Sag, and Neal Snider. 2013. The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes* 28: 48–87. [CrossRef] [PubMed]

Maia, Marcus. 2013. Linguística experimental: Aferindo o curso temporal e a profundidade do processamento. *Revista de Estudos da Linguagem* 21: 9–42. [CrossRef]

Perlmutter, David M. 1968. Deep and Surface Structure Constraints in Syntax. Doctoral dissertation, MIT, Cambridge, MA, USA.

Rosenbaum, Peter S. 1967. *The Grammar of English Predicate Complement Constructions*. Cambridge: The MIT Press.

Ross, John Robert. 1967. Constraints on Variables in Syntax. Doctoral dissertation, MIT, Cambridge, MA, USA.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Snyder, William. 1994. A psycholinguistic investigation of weak crossover, islands, and syntactic satiation effects: Implications for distinguishing competence from performance. Paper presented at the CUNY Human Sentence Processing Conference, CUNY Graduate Center, New York, NY, USA, March 17.

Snyder, William. 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry* 31: 575–82. [CrossRef]

Snyder, William. 2021. Satiation. In *The Cambridge Handbook of Experimental Syntax (Cambridge Handbooks in Language and Linguistics)*. Edited by Grant Goodall. Cambridge: The Cambridge University Press, pp. 154–80.

Sobin, Nicholas. 1987. The variable status of Comp-trace phenomena. *Natural Language & Linguistic Theory* 5: 33–60.

Sprouse, Jon. 2009. Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry* 40: 329–41. [CrossRef]

Sprouse, Jon, and Sandra Villalta. 2021. Island effects. In *The Cambridge Handbook of Experimental Syntax (Cambridge Handbooks in Language and Linguistics)*. Edited by Grant Goodall. Cambridge: The Cambridge University Press, pp. 227–57.

Stepanov, Arthur. 2007. The end of CED? Minimalism and extraction domains. *Syntax* 10: 80–126. [CrossRef]