EXPERIMENTAL SYNTAX: EXPLORING THE EFFECT OF REPEATED EXPOSURE TO ANOMALOUS SYNTACTIC STRUCTURE –EVIDENCE FROM RATING AND READING TASKS

by

Jerid Cole Francom

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF LINGUISTICS

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2 0 0 9

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by Jerid Cole Francom
entitled Experimental Syntax: exploring the effect of repeated exposure to anomalous syntactic structure –evidence from rating and reading tasks
and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Simin Karimi — Date: 14 May 2009

Andrew Barss — Date: 14 May 2009

Kenneth Forster — Date: 14 May 2009

Michael Hammond — Date: 14 May 2009

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director: Janet Nicol — Date: 14 May 2009

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Jerid Cole Francom

ACKNOWLEDGEMENTS

Although this dissertation was written by me, it was a collaborative effort and would not have come to fruition without the help of many over a number of years. I will not be able to recognize nor fully thank everyone who has supported me in this effort in particular, and in my studies more generally, but I hope to name a few people who made special and direct contributions to the work documented in this thesis and my growth as an academic and as a person.

First and foremost I would like to express my gratitude to my chair Janet Nicol. Our weekly meetings, email exchanges and walks on the mall were invaluable to the development of the ideas in this work. I will always be grateful for her intellectual insights and keen ability to instill calm in an otherwise turbulent process.

I want to thank the other members of my advising committee; Simin Karimi, Ken Forster, Mike Hammond and Andy Barss. I feel very fortunate to have been able to assemble such a capable and caring group of people to collaborate on this project. I will never forget our roundtable discussions –extremely engaging and productive– it was an honor to work with such a talented group, thank you.

I have to thank two individuals that share the blame for leading me on my current academic path. I would not be working in this area if it were not for the intervention of Eliud Chuffe during a summer course I took for a language requirement. His encouragement that I study Spanish at the graduate level led to my contact with language study and modern linguistic theory. However, without Antxon Olarrea the seed planted might not have taken root. Between A-bar and Zappa, Antxon showed me that there was a place for my rebel spirit and intellectual curiosity in academia. His keen insight into scientific inquiry, dedication to students and love for life helped me realize this thesis and will continue to inspire me throughout my days.

I am fortunate to have had shared my graduate career with great and talented students along the way. I would like to give special thanks to members of my weekend support group: Mercedes Tubino-Blanco and Mans Hulden. Our discussions over spirits, peppered with comments on language, linguistics, politics and life, tied things together for me and were a source of relief and a harbor of sanity. Cheers.

I cannot conclude without expressing my indebtedness to my family. As a youngster I did not see the value of Saturday morning yard work and weekly chores, I do now. Mom and Dad, the work ethic you instilled in me over the years made this degree possible –academia is cake compared to the hot Arizona sun, cactus thorns and the occasional rattlesnake! Finally, I am completely indebted to my wife Claudia and daughter Inés for being there, good times and bad. Thank you for making our house a home, a source of stability and love. Without you none of this is possible.

TABLE OF CONTENTS

TABLE OF CONTENTS – *Continued*

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This thesis explores the nature of linguistic introspection through the phenomenon known in the literature as the Syntactic Satiation Effect, where the perceived unacceptability of some syntactic structures is attenuated on repeated exposure. Recent findings suggest that rating change in experimental settings may not reveal the underlying grammatical status of syntactic objects by mitigating performance factors related to memory limitations, as initially proposed, but rather arise as a response bias conditioned by characteristics of some experimental designs, in effect introducing task-based performance factors. Findings from rating and reading times suggest that there is evidence supporting both accounts of rating change in experimental designs and highlights areas of development for the *Experimental Syntax* program.

Exploring anecdotal reports, Snyder (2000) found that in as few as five exposures, participants found some types of *wh*-extraction anomaly ('weak Islands') significantly more acceptable at the end of the session compared to the beginning whereas others ('strong Islands') did not experience any rating improvement. Varied success in replicating initial results casts doubts on the proposal that rating data, experimentally elicited, can tease apart grammatical from performance sources of unacceptability. Sprouse (2009) suggests an alternative –Satiation arises as an artifact of a disproportionate number of ungrammatical to grammatical sentences in the testing session. This approach provides an explanation for the apparent mismatch in findings, but also highlights issues regarding the advances of experimental syntax: do experimental methods provide better data or do aspects of some designs systematically introduce extraneous influences themselves?

Evidence from three rating and two self-paced reading tasks suggests that although robust evidence supporting the memory-based claim is not found, evidence that Satiation is strictly task-based is not substantiated either; sentences that sa-

tiate *are* similar across experiments. A novel observation is made that satiating sentences are also more readily interpretable than non-satiating sentences – providing some explanation for the apparent mismatch between Satiation studies, and also points to another source of variability associated with experimental approaches to linguistic intuition. In sum, evidence here underlines the composite nature of introspection, points areas of refinement for experimental techniques and advocates for the adoption of cross-methodological procedures to enhance syntactic investigation.

CHAPTER 1

Introduction

This thesis explores the nature of linguistic introspection and the psychological factors that conspire to affect acceptability judgments. Linguistic intuition has served as the primary source for data for syntactic theory since the inception of Generative Grammar. This is due in part to the speed and efficiency with which introspective data is obtained, but also because it provides access to negative linguistic evidence and avoids many performance limitations found in natural language use. However, acceptability judgments unavoidably reflect more than the contents of a native speakers tacit knowledge of language. A host of cognitive processes interact with language knowledge diminishing the degree to which data collected from the intuition process is a direct reflection of linguistic structure.

For the most part, theoretical linguists have accepted the composite nature of linguistic intuition as an unavoidable occupational hazard. Strategies to limit contaminating influences are typically learned, and employed by syntacticians informally. However a growing research program in *Experimental Syntax* suggests that some extraneous performance influences, inseparable from informal elicitation techniques, can be controlled for and mitigated through the adoption and implementation of experimental methodologies. Proponents of experimental approaches to linguistic intuition claim group testing procedures, typical in psychology, can reveal differences between syntactic structures that go otherwise undetected in investigations that employ informal methods.

Experimental syntax has primarily aimed to gauge the relative rating differences between syntactic structures for an entire experimental session. But rating data collected in experimental procedures also provides acceptability data for participants at various points over the course of the experimental session, opening the possibility to investigate the stability of the linguistic intuition. Exploring anecdotal evidence

from syntacticians that some types of *wh*-Island violations appear to become more acceptable with increased familiarity and repeated evaluation, Snyder (1994, 2000) conducted a series of experiments tapping into the relative changes of acceptability scores for a number of well-known *wh*-Island constructions as a function of exposure. Initial experimental results found that in as few as five exposures, participants find some types of syntactic anomaly significantly more acceptable whereas others undergo no rating improvement.

Rating improvement or 'Satiation' was taken as evidence that the surface unacceptability of some anomalous sentences is not due to syntactic constraint, but rather performance limitations related to working memory limitations. On this account, initial assessments for some syntactic constructions are colored by undo burdens on memory imposed by these structures and not on underlying syntactic well-formedness alone. If this hypothesis is correct, data from rating change provides syntacticians another angle from which to assess the grammatical status of structures. Moreover it points to a further advantage of employing experimental techniques for syntactic investigation.

The question to be raised, and the focus of this dissertation, is, to what extent does rating change as a function of exposure provide new data on the underlying grammatical status of syntactic structure?

1.1 Linguistic intuition and syntactic theory

The goal of modern linguistic theory is to ascertain the contents of speakers' unconscious grammatical knowledge. This tacit knowledge, or grammar, will contain rules and principles that generate all of the licit linguistic sequences of a language and none of the illicit ones (Chomsky, 1957). In a break from earlier approaches to data and data collection influenced by the Behaviorist tradition [1] which focused on 'observable' language as a primary data source, Chomsky (1955, 1957, 1959) argued that data from a corpus of naturally occurring speech contains instances of language

[1] Associated with B.F. Skinner in regards to language.

use that should not be considered as generated by the grammar (disfluencies, slips of the tongue – due to memory limitations, fatigue and other) and also inevitably lacks a range of structures that can be generated by the grammar.

As a consequence, Generative Grammarians turned to linguistic intuition as a key data source on the contents of tacit linguistic knowledge. The proposal contends that despite the limitations inherent in linguistic performance, speakers of a language do have a tacit internalized knowledge of the general principles that govern their language, can consult this knowledge through introspective reflection and can report distinctions in acceptability such as those in (1) and (2).

(1) Well-formed

   a. Agatha Christie has written many books.

   b. I don't like detective stories.

(2) Ill-formed

   a. *Agatha Christie many books written has.

   b. *I detective stories like.

(Haegeman, 1994)

Employing data from linguistic intuition signaled a return to introspective approaches in psychology (Wundt, 1900) and opened the possibility of abstracting away from limitations and errors common in speech in order to access a more direct reflection of the contents of grammatical knowledge. However a series of conditions were stipulated in order to secure the validity of the elicited introspective data:

> Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance. (Chomsky, 1965, p. 3)

Furthermore, a native speaker can only report the acceptability of an utterance, not its grammaticality. Grammaticality is an internalized status which characterizes a linguistic sequence generated by the grammar, the sum of the rules and principles that form a speaker's grammatical 'competence'. However, acceptability is subject is only partially derived from grammaticality, a host of other non-syntactic 'performance' factors potentially enter into the introspective process (to more or less degree as in naturally occurring language use.). Herein lies the key distinction made between *Competence* and *Performance* (Chomsky, 1964, 1965).

Although the advantages to introspective data are apparent, a paradox arises: no data from competence can be obtained without engaging performance to some degree. This inextricable link between competence and performance, and the composite nature of linguistic intuition has been accepted, by and large, by linguists as part and parcel of "doing syntax". However a growing research program in Experimental Syntax suggests that while typical methods employed by linguists are adequate to deal with clear-cut cases as in (1) and (2) this type of approach is less equipped to evaluate more subtle cases of ungrammaticality as in (3).

(3) a. ??Which car did John ask how Mary fixed?

b. ?Who did John ask which car fixed?

Proponents of experimental approaches to syntactic inquiry suggest that by adopting more standard "group testing" practices from psychology, variability introduced by small sample sizes and unstandardized collection practices can be mitigated (Schütze, 1996; Cowart, 1997; Tremblay, 2005; Sprouse, 2007c, inter alia). Moreover, experimental syntax introduces the prospect of revealing contrasts between structures that cannot be readily detected with standard techniques.

Evidence that experimental approaches can reveal contrasts for syntactic structures that have gone undetected through informal techniques come from both typically accepted grammatical and ungrammatical structures. In the case of grammatical structures, it is widely accepted that the *That-trace* filter, active languages such as English, is not operative in Spanish as seen in the contrast in (4b) and (5b).

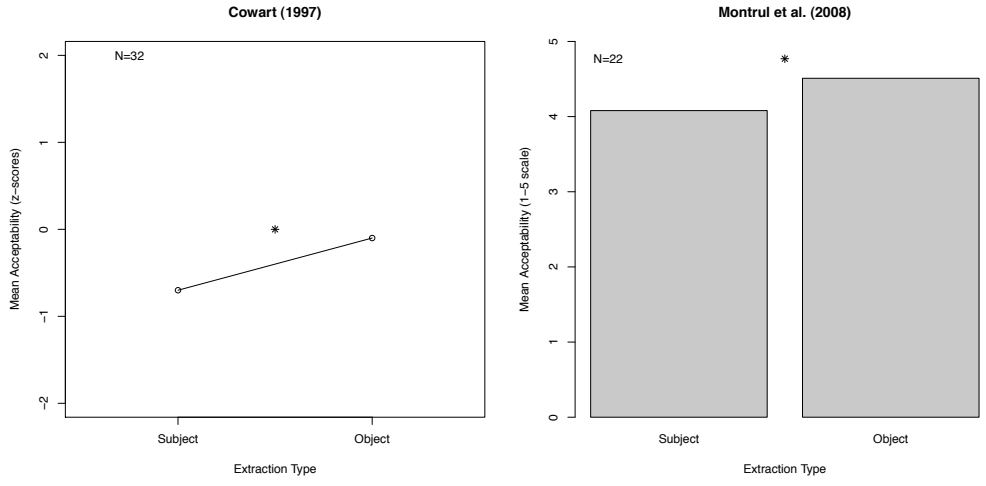
(4) a. What$_i$ did John think that Mary gave away t$_i$?

b. * Who$_i$ did Ana say that t$_i$ broke the door?

(5) a. Qué$_i$ pensó José que María regaló t$_i$?
what thought José that María gave away
What did John think that Mary gave away?

b. Quién$_i$ dijo Ana que t$_i$ rompió la puerta?
who said Ana that broke the door
Who did Ana say that broke the door?

Investigating the acquisition of subject and object extraction preferences for L2 learners of Spanish Montrul et al. (2008) found unexpectedly that the control group of native Spanish speakers showed reliable differences between subject (5b) and object (5a) extractions from subordinate clauses. It is important to note that both extraction types are highly acceptable, but the contrast between object and subject extraction appears to be qualitatively similar to patterns found for English by Cowart (1997), seen graphically in Figure 1.1.[2]



(a) English (b) Spanish

Figure 1.1: Object versus subject extraction in English and Spanish.

[2]Asterisks will be used in graphs to represent statistical significance.

Experimental findings for object/subject extractions in Spanish subordinate clauses show subtle differences for otherwise 'grammatical' sentence configurations but contrasts are also detected for structures that typically show robust unacceptability. Taking another example from Spanish, preverbal subjects in interrogatives, in (6), are found unacceptable for speakers of standard varieties (Torrego, 1984; Olarrea, 1996; Goodall, 2005).

(6) a. * Qué$_i$ **tú** leíste en la biblioteca t$_i$?
what you read in the library
What did you read in the library?

b. * Qué$_i$ **el niño** leyó en la biblioteca t$_i$?
what the boy read in the library
What did the boy read in the library?

However for speakers of Caribbean Spanish pronouns in preverbal subject position, as in (6a), are found fully acceptable (Ordóñez and Olarrea, 2001, 2006). Contrary to typical accounts, Goodall (2005) found that ratings scores elicited experimentally from Spanish speakers of standard varieties pattern with the Caribbean speakers, suggesting again the asymmetries between dialects (or languages) may be a matter of degree not qualitative difference.

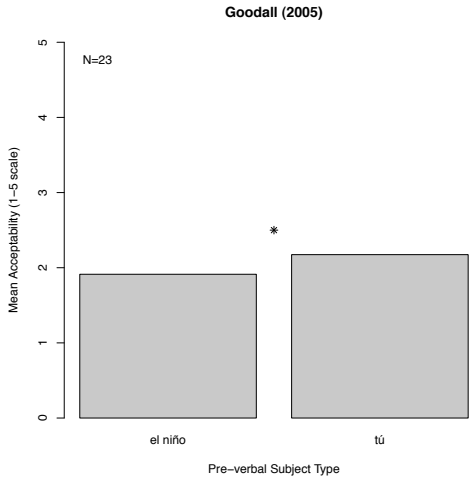
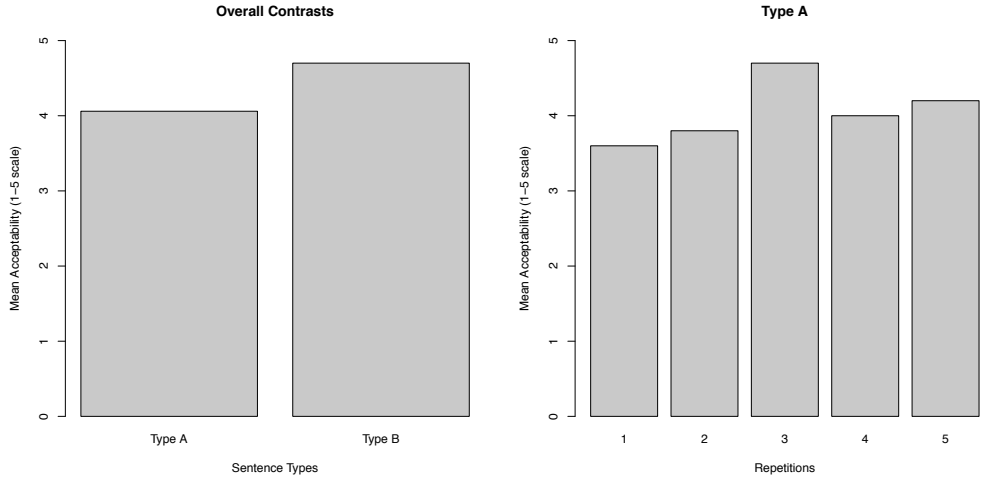


Figure 1.2: Contrasts for type of pre-verbal subject in Spanish.

The cases cited here from Goodall (2004) and Montrul et al. (2008) are examples of a growing trend to apply experimental techniques to syntactic investigation (Nagata, 2003; Featherston, 2005; Sorace and Keller, 2005, inter alia). Most typically experimental techniques have focused on the mean sentence acceptability scores for an entire session, illustrated in Figure 1.3a. Yet data collection methods implemented in experimental designs allow researchers another type of view into the introspective process – change in mean acceptability scores over the course of an experiment, Figure 1.3b.



(a) Contrasts between sentence types. (b) Contrasts per repetition for a single type.

Figure 1.3: Overall and repetition mean acceptability.

Snyder (2000) explored the possibility that rating change in an experimental session could provide corroborating evidence for anecdotal accounts from syntacticians that initial perception of unacceptability for some syntactic anomaly appears to be attenuated with more exposure and familiarity. Known informally as 'judgment fatigue', ' the linguist disease', etc., experimental findings for seven distinct *wh*-Islands show that participants' responses corroborate informal evidence; some types of anomaly are rated more favorably with increased exposure and repeated evaluation, whereas other are not. Rating improvement or 'Satiation' for classically 'weak' Islands (Subjacency violations) was taken as evidence that the surface

unacceptability of these structures is due to processing constraints related to working memory limitations and not to syntactic constraint. The basic idea being that memory resources are inherently more capable to adapt to changing demands than grammatical constraint, which by hypothesis dichotomous and immutable. [3]

Snyder's basic proposal suggests that experimental approaches to syntax can provide a novel tool for syntacticians to evaluate the underlying grammatical status of syntactic structure – effectively creating a diagnostic test able to tease apart grammatical from performance sources of unacceptability. However, subsequent studies following up this line of inquiry report varied success in replicating Snyder's results, casting doubts on the initial account for the Satiation effect. Sprouse (2007c) suggests that increased ratings in judgment tasks arise as artifacts of particular design conditions employed in Satiation studies, namely a disproportionate number of ungrammatical to grammatical sentences in the testing session. If on the right track, this approach provides some explanation for the apparent mismatch in results for Satiation effects, but also highlights larger issues regarding the advances of experimental syntax in general: do experimental methods provide more revealing data or do particular design aspects applied in some experimental designs systematically introduce extraneous influences themselves?

1.2 Research questions and overview

This thesis has three main goals:

1. to provide a detailed investigation into the potential sources of rating change in Satiation studies.

2. to evaluate the extent to which experimental approaches to acceptability judgment tasks provide a clearer view into the structure of tacit linguistic knowl-

[3] Another approach to the nature of grammar suggests that grammatical constraints are better understood as gradient and violable, and potentially subject to change over time. See Bard et al. (1996); Keller (1996); Sorace and Keller (2005).

edge compared to traditional methods, i.e. better equipped to mitigate extraneous performance factors associated with linguistic intuition.

3. to explore the nature of linguistic introspection and the relationship between linguistic knowledge, linguistic behavior and linguistic theory.

The investigation begins the with a closer look at evidence for the Syntactic Satiation effect. Two main proposals are explored. First, Snyder's initial hypothesis that satiating sentence types are not ungrammatical, but rather difficult to process, is reviewed. Mixed findings in replication studies following up this line of inquiry feed a second proposal: Satiation effects are not linguistically-based but rather essentially task-based and arise as artifacts under particular experimental designs typically employed in Satiation studies. Evidence from Experiment 1 suggests that although robust evidence supporting the sentence processing claims for Satiation is not found, evidence that the effect is strictly non-linguistic is not substantiated either; the types of structures that satiate *are* similar across experiments. A novel approach is introduced based on the observation that satiating sentences tend to be more readily 'interpretable' when compared to non-satiating sentences. This account puts forward the hypothesis that participants confuse interpretability with grammaticality in task-bias conditions, again pointing to experimental confounds, not asymmetries in processing effects, as the primary source of Satiation effects.

Chapter 3 attempts to address the disparity between replication studies by looking at the parallels between Satiation effects and a more well-studied phenomenon on repetition effects in language, Syntactic Priming. A review of the literature suggests, that, whereas priming for syntactic form is robustly found in production, and to some extent in comprehension, it is unknown whether there are priming effects for anomalous syntactic form: 1) given no priming studies have investigated decidedly anomalous syntactic structures and 2) rating improvement in Satiation studies have only been found in ungrammatically-biased judgment tasks. Taken together a prediction is made: if Satiation is an instance of Syntactic Priming, rating improvement should be a function of mere exposure, not task type nor task balance.

Evidence from Experiment 2 and 3 suggest that satiation *does* occur in balanced task designs but may in fact be contingent on the judgment process itself –mixed finding in support of the notion that Syntactic Satiation is an instance of Syntactic Priming.

Chapter 4 uses both rating and reading time measures to explore the extent to which Satiation effects arise due to facilitated processing of syntactic structure or rather as an artifact of the judgment process itself. Despite claims that repeated exposure to sentence anomaly in a rating task facilitates processing, it has not been corroborated in tasks more typically employed in psycholinguistics, such as self-paced reading. Given the composite nature of the judgment process, and null results for word-by-word reading exposure in Experiment 3, the prediction is that Satiation effects stem from judgment process itself and not from facilitated syntactic processing. Results from a two-part experiment in which participants were exposed to anomalous sentences in a rating task and subsequently participated in a reading task suggest that Satiation effects do appear to have a processing component. Although the findings here make an important contribution to the connection between syntactic processing and linguistic intuition for unacceptable sentence types, there is also evidence that participants deploy non-linguistic strategies that pose obstacles to clear conclusions about the nature of Syntactic Satiation effects.

The results from the five experiments presented here suggest that Satiation effects are replicable in rating tasks. However it is less clear what underlies these effects. Evidence gathered cannot exclude the possibility that the memory limitations, Syntactic Priming or test-taking strategies play some role in Satiation effects. But any one of these factors alone can not account for the variety of sentence types that undergo rating improvement and exclude those that do not given the evidence.

Thus, results here suggest that adoption of group testing practices, typically employed in experimental approaches to syntactic inquiry, can inadvertently introduce systematic non-syntactic differences interpreted as syntactically relevant. These findings point to potential shortcomings of the Experimental Syntax program as currently practiced and highlight areas for future investigation.

CHAPTER 2

The Syntactic Satiation Effect

The current chapter takes a closer look at the phenomenon known in the literature as the Syntactic Satiation Effect and its potential sources. A phenomenon first noted anecdotally and then effectively reproduced in the lab, Satiation refers to the increased acceptability ratings in a judgment task for sentence types initially assessed as unacceptable. Snyder (1994, 2000) interpreted results showing differential response patterns between 'weak' and 'strong' Island violations as evidence in favor of two distinct etiologies for otherwise unacceptable sentence types; processing burden versus syntactic violation. Other studies following up this line of inquiry report varied degrees of difficulty in replicating the initial results. A counter proposal from Sprouse (2009) suggest that, rather than a processing phenomenon, Satiation is an artifact of the particular design conditions typically employed in Satiation studies. Findings from a replication experiment here successfully reproduce increased-rating responses. Sentence types that show Satiation effects in the current replication experiment highlight the complications of categorizing Satiation effects in the terms of attenuated-memory burdens or experimental design alone.

2.1 Initial evidence for the Satiation effect

For years linguists have used linguistic intuition judgments as data for their own research, research for their colleagues and in the classroom. A common observation about introspective reactions to syntactic structure is that not all structures appear to be equally acceptable – or unacceptable. Further, it has been observed that some types of otherwise anomalous sentence structures appear to become more acceptable with repeated exposure to them. This phenomenon has been labeled 'judgment fatigue', the 'linguist disease' and more recently 'Syntactic Satiation'. Spencer (1973)

introduced the term 'Satiation' into the literature to describe changes in judgments with exposure to particular syntactic structures. The term makes reference to Semantic Satiation a phenomenon first documented by Severance and Washburn (1907) in which repeated exposure to a word produces a sense of detachment, or 'foreignness' to the form of the word and its meaning.[1] Syntactic Satiation, on the other hand, appears from anecdotal evidence to produce quite the opposite effect – those that experience the effect have the feeling that they understand the utterance better.[2] Consider a sentence such as in (1). Linguists' ability to clearly respond 'Jill' (or some other name) to this question underscores this key difference between semantic and syntactic satiation effects.

(1) Who does Mary believe the claim that John saw?

To some degree all syntacticians experience this sensation for some types of syntactic structure, often forgoing intuitions for commonly assumed ungrammatical sentence types and relying on memorized judgments as a source of stability. This phenomenon has been codified into theory as the observational distinction between 'weak' (2), marginally unacceptable and 'strong', robustly unacceptable (3 and 4) syntactic Island constraints (Ross, 1967). This distinction has been theoretically subsumed by Subjacency violations (2) on the one hand and violations of the Condition on Extraction Domains (CED) (3) or the Empty Category Principle (ECP) (4) on the other (Lasnik, 1984).

(2) What did John ask who wrote? (Subjacency)

[1]Spencer, however, did not imply that the satiation process itself lead to changes in introspective judgments. In order to explain why linguists and naïve participants have contrasting judgments, he suggests three separate processes bias those with linguistic training towards particular analyses of syntactic structure. First, satiation applies on repeated exposure, detaching initial form-meaning relations. Second, a reorganization process attempts to categorize the utterance. At this point in the process, a lack of clear linguistic reorganizing principles is filled by theoretical expectations and assessments of syntactic form 'contextually stabilizes' the response in a particular way, based on particular knowledge of proposed rules acquired in linguistic training.

[2]The confusion in terms is problematic. The term 'training effects' introduced for other reasons may be less problematic.

(3) Who did you eat pizza after calling? (CED)

(4) Who did John ask what wrote? (ECP)

2.1.1 Snyder 2000

An attempt to investigate the Satiation effect in the lab was conducted by William Snyder at MIT. This experiment aimed at gauging the robustness of the effect for a range of anomalous Wh-extraction types. Twenty-two undergraduate participants were recruited to provide yes/no acceptability judgments for 50 sentences. Five instances for each of seven different anomalous interrogative constructions were created for presentation, four of which were classic Island violations and three other known *wh*-extraction violations. These types were selected as a representative sample of both 'satiating' and non-satiating violation types according to anecdotal reports. Examples of these sentence types appear in Table 2.1.

Want for	Who does John want for Mary to meet?
Whether Island	Who does John wonder whether Mary likes?
That-trace	Who does Mary think that likes John?
Subject Island	What does John know that a bottle of fell on the floor?
Complex NP	Who does Mary believe the claim that John likes?
Adjunct Island	Who did John talk with Mary after seeing?
Left Branch	How many did John buy books?

Table 2.1: Examples of sentence types used in Snyder (2000).

These sentences were grouped into five ten-item blocks each containing one of the seven violations and three grammatical distractors. Blocks were balanced for presentation across subjects to avoid any influence of sentence ordering.

A context sentence accompanied the *wh*-question for each of the 50 sentences to ensure participants were judging the sentences based on the same context. Participants were instructed not treat this as a memory test and to provide independent evaluations for each structure. Each sentence and context pair was provided on a separate page. Participants provided responses with pen and paper.

```
Example I.
[Context:  Mary put a box on the table.]
Test Sentence: "What did Mary put on the table?"
Judgement: _____ (Y/N)
```

A subject was considered to have satiated for a given sentence type if the number of 'yes' responses reported in the last two blocks was statistically greater than in the first two blocks.

Findings showed rating improvement for 'weak' Whether and Complex NP Islands. Apart from a reported marginal effect for Subject Islands, no other effects were found for Want for, *That*-trace, Adjunct Islands nor Left Branch Condition violations. These results were taken as evidence that repeated exposure to syntactic structure in judgment tasks can reveal distinct etiologies for anomalous constructions. In other words, the perceived ungrammaticality produced by Whether, Complex NP and to some extent Subject Island violations is not due to syntactic constraint, but rather to processing difficulties related to deficiencies in working memory in on-line processing.

2.1.2 A memory bottleneck

Snyder's proposal connects to literature on the influence of working memory constraints on syntactic processing. It has been long known (Chomsky and Miller, 1963; Bever, 1970) that some structures, such as center-embedded configurations (5), can be structurally well-formed but be perceived as unacceptable.

(5) The woman$_i$ the man$_j$ the host$_k$ knew$_k$ brought$_j$ left$_i$ early.

Nevertheless the perceived ungrammaticality can be attenuated by manipulating the lexical content of the utterance. Replacing full NPs with pronouns and indefinites improves the acceptability of (5) by making the utterance less 'confusing', seen in (6).

(6) The woman$_i$ someone$_j$ I$_k$ knew$_k$ brought$_j$ left$_i$ early.

Investigations into the comprehension of long-distance dependencies cite two dynamics related to working memory constraints that lead to graded performance: 1) structural distance between the filler and gap and 2) the referential status of the DPs that intervene between filler and gap (Fodor, 1978; Gibson, 1998; Kluender, 1998; Gibson and Warren, 2004; Kluender, 2005). Evidence for the influence of structural distance comes from relatives, subject relatives (7) are read quicker and more accurately than object relatives (8) (King and Just, 1991).

(7) The reporter$_i$ that t$_i$ attacked the senator admitted the error.

(8) The reporter$_i$ that the senator attacked t$_i$ admitted the error.

The notion is that a filler ('the reporter', in this case) must be held in working memory until a gap (the structural position where the filler is to be interpreted) can be found. Working memory allocations decay over time, and the longer the distance between filler and gap, the more difficult the relationship becomes to establish.

The referential status of DP fillers and DPs that intervene between the filler and gap also have an effect on the sentence parser's ability to parse dependency relations. In general terms, the more referential (Discourse Linked) a DP is, the more memory resources it requires, and higher activation levels it achieves. For interveners, the more referential, the more disruptive to a successful parse. Hence, the referential intervener 'the reviewer' makes (9) more difficult to process than (10) where a non-referential DP 'someone' appears.

(9) That's the article$_i$ that we need to find **the reviewer** who understands t$_i$.

(10) That's the article$_i$ that we need to find **someone** who understands t$_i$.

(Kluender, 1998)

On the other hand, more referential fillers (11) are more resilient to intervening material, and thus are more comprehensible than non-referential fillers (12).

(11) **That's the article**$_i$ that we need to find someone who understands t$_i$.

(12) **What**$_i$ do we need to find someone who understands t$_i$.

(Kluender, 1998)

Along these lines the apparent perception of ungrammaticality for Subjacency violations has been attributed to a processing 'bottleneck' in which memory demands exceed resource availability (Berwick and Weinberg, 1984; Kluender and Kutas, 1993; Kluender, 2005; Sag et al., 2007, 2008). This approach suggests that the underlying syntactic structure of Subjacency constructions is well-formed, but that a combination of filler-gap distance and excessive memory demands posed by intervening lexical material result in perceived unacceptability in (13a), attenuated in (13b).

(13) Complex NPs

a. * **What**$_i$ does John believe **the claim** that Mary likes t$_i$?

b. ? **Which restaurant**$_i$ does John believe **claims** that Mary likes t$_i$?

The Memory Bottleneck account of Satiation suggests that 1) repeated exposure attenuates processing burdens for some anomalous sentences revealing their underlying grammatical status and 2) judgment tasks are sensitive enough to detect some types of sentence processing effects. This view holds that experimental syntax provides the precision necessary in order to evaluate and tease apart this grammatical contrast that is otherwise obscured in more informal methods for eliciting judgment data.

2.2 Other findings for exposure-based rating change

Attempts to effectively reproduce the effects found in Snyder's work have resulted in mixed support for the Memory Bottleneck hypothesis. On the one hand, Braze (2002) and Goodall (2004) found evidence supporting Snyder's main conclusions.

Braze specifically tested the contrast between Whether and Adjunct Islands and found that Whether Islands showed robust Satiation effects, but Adjunct Islands did not. In a somewhat more comparable sample of the original seven extraction types, Goodall looked at Subject island, Complex NP, That-trace, Adjunct Island and Left Branch Condition sentence types.[3] He also found that Complex NP violations show Satiation effects. However, with regards to Subject Islands, which Snyder found a trend towards Satiation, Goodall did not.

Hiramatsu (2000) performed a series of experiments exploring the original materials and methods provided by Snyder. Her results show some overlap, but also a number of particular unexpected mismatches. Of the three predicted satiating constructions, she found effects for Whether and Subject Islands – but not Complex NP violations. In addition, Want for and That-trace sentences improved as a function of exposure. There are no clear commonalities between Want for and That-trace effects, on the one hand and Subjacency violations on the other, such that a strictly working memory-based processing could account for both.

2.2.1 An equalization strategy

The apparent mismatch in findings across Satiation studies is of some concern and raises doubts that rating improvement exclusively indexes attenuated memory demands. Sprouse (2007b, 2009) takes the perspective that Satiation effects are an artifact of the particular experimental design typically employed in Satiation studies, specifically the preponderance of ungrammatical items compared to grammatical items presented to subjects. He suggests that the uneven ratio of ungrammatical to

[3]This experiment also included 'No inversion' sentence types seen in (i) and (ii).

(i) no inversion case Spanish (with gloss)

(ii) no inversion case English

In a split population experiment contrasting Satiation effects for native English and native Spanish speaker's variable acceptability, Goodall found Spanish speakers satiate for No inversion cases, while English speakers do not.

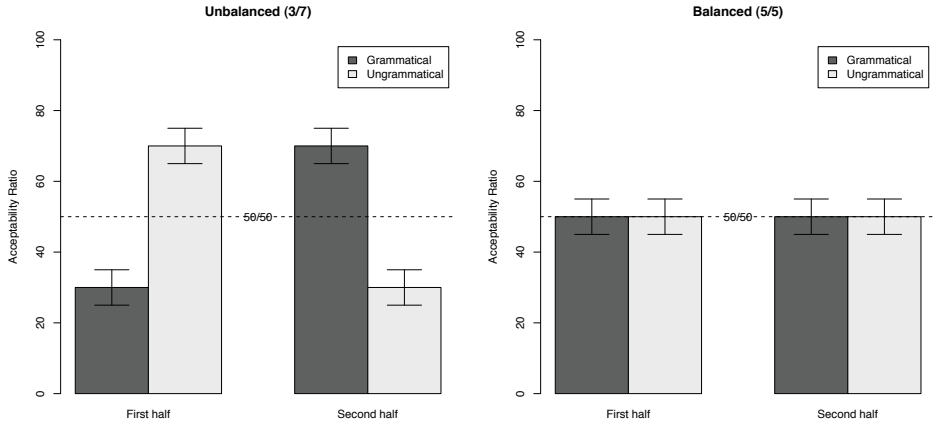
Sentence Type	Snyder(1999)	Hiramatsu(2000)	Braze(2002)	Goodall(2005)
Complex NP	+	-		+
Subject Island	(+)	+		-
That-trace	-	+		-
Want for	-	+		
Whether Island	+	+	+	
Adjunct Island	-	-	-	
Left Branch	-		-	

Table 2.2: Findings Summary for Satiation Effects.

grammatical sentences presented in the experimental session engages participants in an extra-linguistic strategy – the Equalization Strategy – that induces more acceptable ratings in the second half of the experimental session compared to the first (Figure 2.1).

Sprouse notes that the majority of studies demonstrate this particular design characteristic. As an example, Snyder's original design included seven ungrammatical items for every ten. Participants are instructed to make a forced-choice between 'yes' or 'no' for 50 sentences. Under these expectations it would be reasonable, all else being equal, for participants to expect a 50/50 split between grammatical and ungrammatical items. However, mid-way through the session participants that respond 'correctly' will find themselves at a 30/70 split. In response to this mismatch participants may respond more favorably in the second portion of the session to effectively 'equalize' the distribution of yes/no scores. In effect, this hypothesis predicts that participants 'keep track' of their responses and check them against some expected outcome for typical testing situations with dichotomous, forced-choice scales.

The Equalization Strategy underlines the potential contextual influences on the decision-making process that can arise due to the particular conditions of an experimental session. Work by Tversky and Kahneman (1974) on 'Anchoring' and 'Framing Effects' suggest that, in fact, decision-making can be quite influenced by factors outside the indicated calculation space directed in a task. For example, participants asked to provide an estimation of the product of a calculation such as (14a)



(a) Satiation effects (b) No Satiation effects

Figure 2.1: Equalization: the proposed effect of task balance on acceptability rating change.

consistently find the result many times larger than in (14b) despite the fact that the product of both calculations is equal.[4]

(14) a. $[8 * 7 * 6] * 5 * 4 * 3 * 2 * 1 = 40320$

b. $[1 * 2 * 3] * 4 * 5 * 6 * 7 * 8 = 40320$

The rationale is that participants 'anchor' their responses to the initial context of the task (indicated with brackets) which in turn biases their expectation for the total calculation. In relation to the Satiation work, the ratio bias could serve as a response bias and itself may drive increased rating scores.[5] Given the non-linguistic nature of the strategy, spurious results found across the replication studies are a

[4]Moreover, participants responses typical underestimate the product of these calculations in both cases.

[5]It is of note that the argument made by Sprouse is that participants are aware of their previous responses and the equalization strategy underscores a holistic approach to the experimental session. In the 'Anchoring' case cited here, it is the fact that participants cannot keep track of the distribution that feeds the differences between predicted products for each of the two equations in (14a-14b). In any event, the key point to be made is that the design of the experiment can itself influence participant responses above and beyond the treatment variables under investigation.

predicted outcome in these types of experimental setups.

2.3 Summary and predictions

The review of the primary literature provided in this chapter highlights two potential sources for the Satiation Effect: a Memory Bottleneck and an Equalization Strategy. Each approach makes distinct predictions about outcomes for a replication of Snyder's original experiment. First, the Memory Bottleneck hypothesis predicts that syntactic structures that are hypothesized to be perceived as anomalous due to demands on working memory in on-line comprehension (Subjacency effects) should show increased ratings as a function of exposure. The apparent difficulty found by the cited authors in consistently replicating these particular effects raises questions as to whether memory limitations are at the root and/or the sole source of these effects. On the other hand, the Equalization strategy claims that the apparent mismatch in results is evidence in support of the notion that the underlying root of Satiation Effects is non-linguistic, and based on more general decision-making heuristics. This approach predicts inherent variability in Satiation studies that employ unbalanced experimental designs. However, the results of previous work have not been as inconsistent as expected on a response bias account; findings are robust for some satiating and non-satiating types across experiments. The following experiment aims to add another attempt to replicate Snyder's initial results in order to assess the degree of variability associated with experimental approaches to rating change.

2.4 Experiment 1

2.4.1 Methods

Participants

205 undergraduate students from six introductory linguistics courses at the University of Arizona gave informed consent to participate in a short in-class experiment

during regularly scheduled class time. Participants completed a brief questionnaire soliciting information on age, sex, place of birth, native language, bilingualism, and linguistic training.

13 participants were excluded as non-native speakers of English. The remaining 192 were born in a variety of states across the United States [6], however almost half were born in Arizona (82). 61 claimed linguistic proficiency in a language other than English (37 of which included Spanish). 13 were considered trained in linguistic theory. Participants ranged in age from 18 to 33 years of age (mean 20.8).

Materials and procedure

Each participant rated a total of 50 interrogative sentences. These sentences were the same sentences and sentence types used in Snyder (2000).[7] These included 35 ungrammatical target sentences and 15 grammatical distractors. The target ungrammatical items consisted of seven typical w*h*-extraction violations seen in Table 2.3.

Want For	Who does John want for Mary to meet?
Whether	Who does John wonder whether Mary likes?
That-trace	Who does Mary think that likes John?
Subject Island	What does John know that a bottle of fell on the floor?
Complex NP	Who does Mary believe the claim that John likes?
Adjunct Island	Who did John talk with Mary after seeing?
Left Branch	How many did John buy books?

Table 2.3: Target ungrammatical items for Experiment 1.

The experimental items were organized into five blocks of ten items. Each violation appeared once in each block along with three grammatical distractors. Two lists were created in which block ordering reversed for the first and last two blocks to avoid the relative contribution of a particular sentence to the overall rating pattern.

[6] States represented include: AZ, CA, CN, CO, DL, GA, IL, IW, KS, LA, MA, MI, MN, MO, MS, NC, NJ, NM, NV, NY, OH, OK, PA, SC, SD, TX, VA, WA

[7] A special thank you to William Snyder for sharing these materials.

In addition, items within each block were randomized for each list to avoid within block ordering effects.

Participants were presented one of the two lists using a projector and the presentation software Keynote. Each sentence was projected individually and displayed for 8 seconds before self-advancing to the next item. [8] Each sentence appeared on one line, centered in new courier font. [9]

Participants were instructed to read each sentence to silently and provide a "Yes" rating for acceptable sentences and "No" for unacceptable sentences. Two examples were provided in the instructions discouraging participants from using prescriptive grammatical forms as a basis for acceptability. In addition, an effort was made to highlight that the session was not designed as a memory test and participants were encouraged to make independent assessments of the items presented.

Participants recorded their responses within a table found on the opposite side of the questionnaire containing two columns corresponding to Yes and No and 50 rows for each sentence displayed. The session including consent, questionnaire and rating lasted from 20 to 25 minutes.

The questionnaire and rating responses were entered by hand into a database for organization and preparation for data analysis.

2.4.2 Results

Following the analysis provided in Snyder (2000), the first two (Set 1) and last two (Set 2) response scores were grouped for each of the seven violation types in each list. These sets served as the basis for relative rating change over the experimental session and were used as the factor Set(First,Last) in the following analyses.

[8] The timing for slide display was tested on an independent group and found to be adequate for the task.

[9] In contrast to Snyder's original design, context sentences were not provided. Evidence from Sprouse (2007a) suggest that contextualizing interrogative sentences does not condition acceptability judgments in a rating task. For this reason, and for matters of presentational consistency with subsequent experiments context sentences were not given in the experiments that appear in this thesis.

First set		Last set		Sign Value
N	Y	Y	Y	Increase
Y	N	Y	Y	
N	N	N	Y	
N	N	Y	N	
Y	N	N	N	Decrease
N	Y	N	N	
Y	Y	Y	N	
Y	Y	N	Y	

Table 2.4: Sign test calculation formula.

First, sign scores were calculated based on the difference between the number of 'Yes' to 'No' responses for Set 1 and 2, illustrated in Table 2.4. Each subject, then, is considered to have shown satiation for a given sentence type when the number of 'yes' responses is greater than the 'no' responses by Sign Test. The results are shown in Figure 2.2. By this measure Whether Island ($p < 0.001$), Want for ($p < 0.001$) and Subject Island ($p < 0.05$) violations showed Satiation patterns. No increase nor decrease-rating trends were significant for the other four anomalous sentence types (That-trace ($p = 0.22$), Complex NP ($p = 0.33$), Adjunct Island ($p = 1.0$) and Left branch ($p = 0.80$) violations). Grammatical fillers also showed no changes from first to last set exposures ($p = 0.70$).

One issue with the Sign Test that has been addressed in the literature is that it does not take into account the mean nor standard deviation of the ratings.[10] Converting a rating to a difference score (Sign value) abstracts away from those participants who do not show an increase nor a decrease, that is, those participants that rate a given sentence type the same over the four measured repetitions or those participants' ratings that are split in both the first two and last two judgments. In order to assess the effect of repetition exposure in a more comprehensive way, Snyder (2000) conducted a series of Paired t-tests and ANOVA analyses on the same data.

[10]It is also of note, and less apparent in the literature, that this measure can only be reliably performed on data sets containing 25 or more pairs (Johnson, 2008).

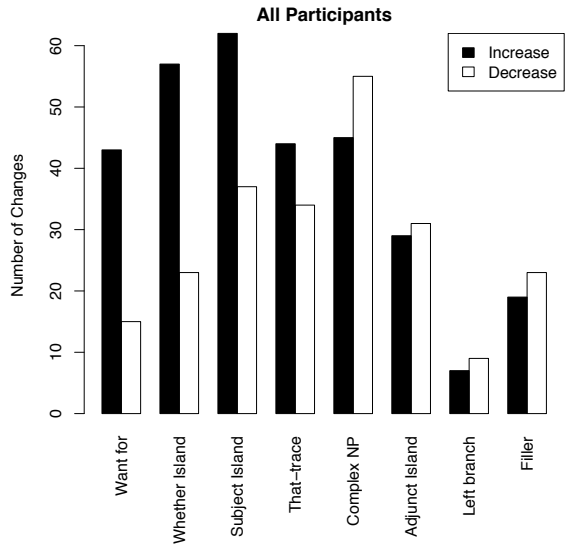


Figure 2.2: Number of participants with response changes of type 'Increase' (more 'no' to 'yes' responses) or 'Decrease' (more 'yes' to 'no' responses) by sentence type in Experiment 1.

Results matched almost entirely with the Sign Test analysis.[11]

Here too ANOVAs and pairwise comparisons were conducted. A 2 x 7 (x 2) ANOVA with Set(First, Last) and Type(Want for, Whether Island, Subject Island, That-trace, Complex NP, Adjunct Island, Left branch) as a within-subjects factors was conducted. List(1,2) was also included as a between-subjects blocking factor. The main effects for Set $[F(1, 382) = 11.8, p < 0.001]$ and Type $[F(7, 1337) = 1119.8, p < 0.001]$ were significant as well as the interaction between them $[F(7, 1337) = 5.61, p < 0.001]$. Simple effects for Set for each level of Type were conducted. Again, Want for $[F(1, 382) = 10.1, p < 0.01]$, Whether Island $[F1(1, 382) = 14.2, p < 0.001]$ and Subject Island $[F(1, 382) = 4.8, p < 0.05]$ violation reached significance. No other sentence types reached statistical significance.

However, parametric measures such as ANOVA are not ideal for categorial

[11]Paired t-tests showed significant effects for Whether Islands and Complex NP violations and marginal effects for Subject Islands ($t(21) = 1.79, p = 0.88$). ANOVA tests resulted in a main effect of sentence type, and a pairwise contrast for Whether Islands and each of the four non-satiating structures from the Sign Test.

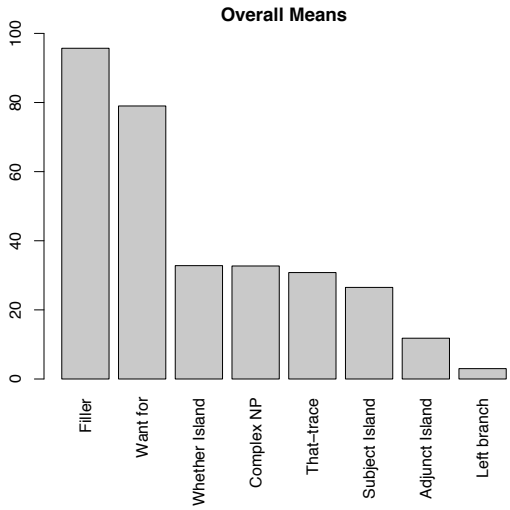


Figure 2.3: Overall acceptability for each sentence type in Experiment 1. (Grammatical distractors are labeled as 'Filler'.)

data.[12] For this reason, a more appropriate test, *logistic regression*, was performed given the categorial distribution of the rating scores collected here. A G^2 score was calculated for the difference between Set(First, Last) each sentence type.[13] Results from these tests match the findings from the Sign Test, t-test and ANOVA analyses (Whether Island ($G^2 = 13.8, p < 0.001$), Want for ($G^2 = 10.1, p < 0.01$) and Subject Island ($G^2 = 4.8, p < 0.05$) sentence types and no other effects).

2.5 Discussion

The current findings underline both the consistencies and inconsistencies across the previous replication experiments. On the one hand, Whether Islands and Subject Islands appear to be particularly sensitive to repeated exposure while Adjunct Islands and Left Branch Condition sentence types show marked resistance to rating change. These findings alone, corroborated in other studies, support the Memory

[12] Although these tests are often used in the literature.

[13] G^2 is is a binomial logit model that performs similar to X^2, however provides the added benefit of being able to model more complex effects (i.e. factorial analyses).

Type	First set	Last set
Filler	96.6	96.4
Want for	75.4	83.4
Complex NP	38.3	34.9
That-trace	28.3	33.2
Whether Island	25.6	37.6
Subject Island	22.9	29.3
Adjunct Island	12.2	11.7
Left Branch	2.9	2.4

Table 2.5: Percent acceptable by repetition for each sentence type ordered by overall acceptability in Experiment 1.

Bottleneck hypothesis that Subjacency constraints are constraints on processing and not syntactic constraints.

While this evidence shows some degree of support for the memory-based account, other anomalous types that show more variable effects across studies, Want for, That-trace and Complex NPs weakens Memory Bottleneck account for Satiation effects. The current experiment found effects for Want for constructions. This corresponds with evidence from Hiramatsu (2000) who also found effects for Want for, in addition to That-trace effects (which were not found here). As noted in her discussion, Comp-trace constructions (Want for and That-trace) have been noted in the literature to be acceptable in some varieties of English (Lasnik, 1984; Sobin, 1987, 2002). This approach aims to explain the disparate findings from Snyder's results that found no Satiation effects for either Want for nor That-trace effects as a sampling issue (differences in the diversity of the populations tested in Hiramatsu's study (U Connecticut) versus Snyder's (MIT)).

However two issues complicate the notion that dialect differences explain contrasting Comp-trace findings. First, Want for constructions employed in Snyder, in (15a), and subsequently adopted in Hiramatsu and the current experiment, are not Comp-trace effects such as in (15b) and (15b) –the true Islands violations.

(15) a. Who$_i$ does John want for Mary to meet t$_i$?

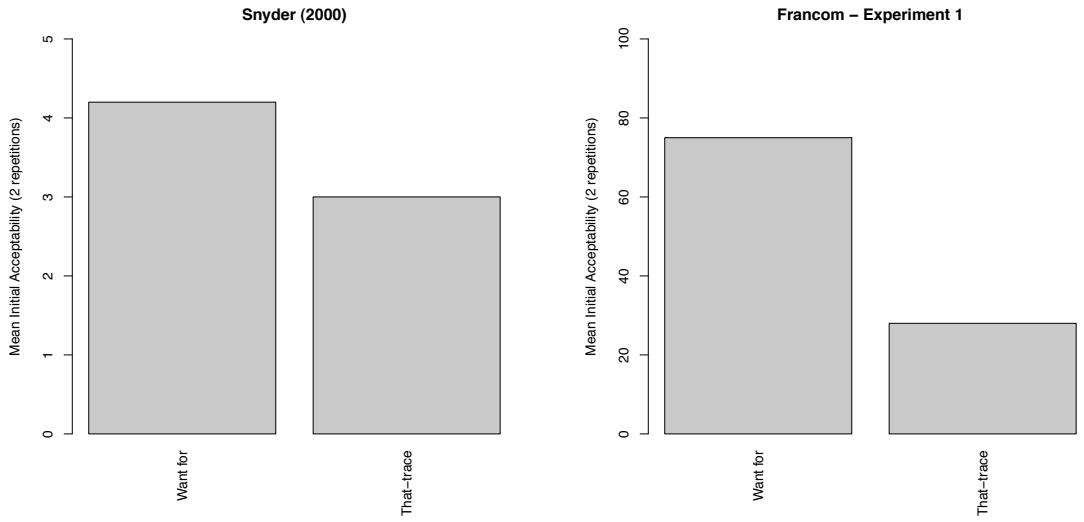
b. * Who$_i$ does John want for t$_i$ to meet Mary?

a. Who$_i$ does Mary think that John likes t$_i$?

b. * Who$_i$ does Mary think that t$_i$ likes John?

Therefore the (un)acceptability of Want for and That-trace cannot be tied to the same assumptions about dialectal variation. Furthermore evidence for the qualitative differences between Want for/That-t sentences' initial acceptability ratings argues against a uniform account. Want for constructions as in (15a) are found quite acceptable, whereas That-trace constructions (15b) show more distinctly graded acceptability, seen in Figure 2.4. The similar distribution of acceptability contrasts with the findings for Satiation: Snyder found no satiation for either construction whereas the current experiment found effects for Want for, and not That-trace. It may be the case that no effects for Want for in Snyder's experiment, and effects in the current experiment, is due in part to ceiling effects. In other words, the scale, not psychological effects prevent Satiation effects for Want for sentences in Snyder's experiment. Want for sentences in the current experiment were found slightly less acceptable, thus allowing room for improvement. Note however, that if this approach is on the right track, Want for constructions may in fact show more robust effects than are attested. These findings underscore problems for the Memory Bottleneck account as the only explanation for Satiation effects.

A further complication for memory-based accounts of Satiation pertains to the mixed results for Complex NP violations. Snyder and Goodall produced robust effects for these putative 'weak' constraints, even more so than for Subject Islands. In addition Complex NP violations form the cornerstone of many arguments for the influence of processing constraints on global acceptability (Kluender and Kutas, 1993; Sag et al., 2007). Nevertheless, Hiramatsu and the current investigation did not find evidence of increased rating.



(a) (b)

Figure 2.4: Initial acceptability differences for Want for and That-trace sentences.

2.5.1 Methodological issues

Methodological differences do exist that may contribute to the divergent findings. First, following a prediction that the intermediate results for Subject Islands in Snyder's original findings was due to insufficient exposure (i.e. number of repetitions), Hiramatsu incorporated two more repetitions and did obtain a reliable effect for Subject Islands. In this respect, some differences found across the studies may reflect the sensitivity of a particular construction to exposure effects. However, the current findings found effects for Subject Islands in five repetitions, as well as effects for Want for constructions that were not found in five-repetition conditions in the other studies reviewed suggesting that the number of repetitions is not key aspect in divergent results.

A second procedural difference in the current experiment is that no context sentence was given to accompany sentences during the judgment task. This approach was taken in order to maintain consistency and facilitate parity between upcoming experiments in this thesis in which context presentation could not be achieved given

the format of the task.[14] However the current experiment is not the only study that did not find effects for Complex NP violations, neither did Hiramatsu. Hiramatsu did include context sentences in the same fashion as in Snyder's original design which suggests that context alone is not the source of Satiation effects. Moreover, if context were in fact found to be a key factor behind Satiation effects for Complex NPs this evidence would independently argue against the Memory Bottleneck account which suggests that attenuated memory limitations as a function of exposure produce increased ratings.

Finally, judgments in both Snyder and Goodall's experiments were not under time constraints whatsoever whereas both Hiramatsu's study and the current investigation did potentially include some added pressure for participants to respond quicker than they may otherwise have naturally. For her part, Hiramatsu underlined in the instructions that participants were to proceed quickly and accurately through the session. In the current experiment the presentation of the material via slide projector meant that responses for some individuals, for some types of violations may have been hurried.

An aspect of the data here that potentially gives this argument some traction comes from the overall 'stability' for each sentence type throughout the experimental session. The Sign Test as performed by Snyder and here, takes into account those participants that show increased or decreased acceptability patterns for a given type, and those scores alone. The other two logical possibilities that are left out of the data are ratings that are 'stable', participants rate a given sentence type the same without change, or 'variable', those participants whose rating patterns are not stable – do not slope towards increased nor decreased acceptability.

Under this light, Complex NP violations show the most variability overall as seen in Figure 2.5.

More generally rating instability was also shown by Sprouse (2009) to co-vary with the item ratio balance of the experimental session. He conducted a pair of

[14]This is in reference to two reading tasks in which context sentences are not easily accommodated.

First set		Last set		Stability
Y	Y	Y	Y	Stable
N	N	N	N	
Y	N	Y	N	Variable
N	Y	N	Y	

Table 2.6: Stable versus unstable responses.

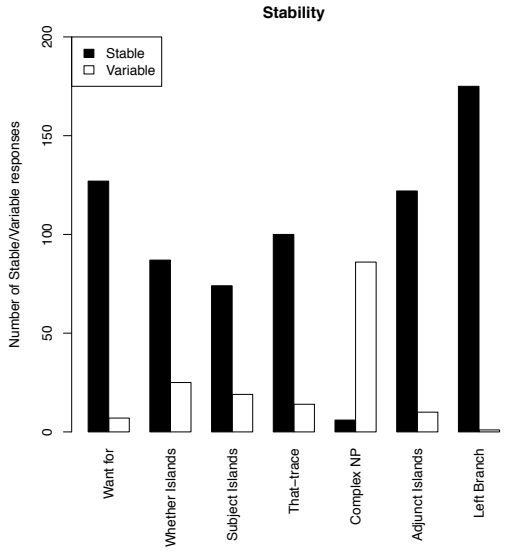


Figure 2.5: Rating stability: number of stable and unstable responses in Experiment 1.

experiments exploring the contribution of (un)balanced designs on the Satiation effect. The first included eight violation types for each ten item block (Ungrammatical Bias) and the second only four anomalous types per ten item block (Grammatical Bias). Although no repetition effects were found for any of the eight distinct sentence types tested – for either experiment, follow-up tests investigating the stability of judgments for each biased context suggested that ungrammatically biased contexts produce more variability overall compared to grammatically biased contexts.

2.5.2 Assessing the influence of response bias

Let's now reconsider the predictions made by the Equalization Strategy for rating change. First, a strong case that the source of the Satiation effect is not linguistic, and therefore subject to inherent variability cannot be made from the currently available findings. For the most part there is a clear pattern of satiating and non-satiating types across a number of independent studies. In particular, all five experiments that tested Whether Islands have found effects, whereas these same experiments found no evidence for Adjunct Islands undergoing effects. Under a strict interpretation of the Equalization strategy this level of consistency is not predicted. However, it cannot be ruled out that a task-based effect, such as the Equalization Strategy, may in fact be operative but grounded to some degree in a particular *linguistic* aspect of the target sentences – accounting for the differences robust asymmetry between robustly satiating and non-satiating sentences.

So the question is, what property could potentially group those types of sentences that show satiation effects and differentiate them from non-satiating types. One option to explore is 'interpretability'. Notice that Whether Islands (16) which consistently show effects in the literature are much more reliably interpretable than their Adjunct Island (17) counterparts which never show effects. This contrast appears salient not only for linguists with extended exposure to these sentences but also quite clearly for non-experts alike.

(16) Who does John wonder whether Mary likes? (Whether island)

(17) Who did John talk with Mary after seeing? (Adjunct island)

This approach suggests that interpretability is a key characteristic that distinguishes satiating and non-satiating sentence types, and participants cue in on this distinction over the course of the experimental session as a secondary criterion to grammaticality. Along these lines, there is supporting evidence that participants naturally search out the communicative value of language even when instructed not to do so. Maclay and Sleator (1960) found participants instructed to rate sentences according to their 'grammaticality' gave the same overall rating for "ungrammatical

but meaningful" sentences, in Table 2.7, as participants instructed to rate 'meaningfulness'.

Yesterday I the child a dog gave
The with feet aching man came yesterday home
Get me from the kitchen a big spoon
You can him not understand
Almost every Saturday she her house cleans
To me was interesting the movie

Table 2.7: Maclay and Sleator (1960)

The empirical problem with Equalization Strategy in its current form is that it predicts more variability among Satiation studies than is attested. In this light, one way to linguistically ground this hypothesis is to connect it to this apparent tendency for participants to extract meaning from anomalous form. If interpretability is the linguistic aspect that guides equalizing responses, then common effects found across the experiments reported may have an explanation. Augmenting the Equalization Strategy with a meaning-based criterion change – implemented consciously by participants in response to the apparent mismatch with expected rating distribution outcomes, makes inroads in accounting for the attested mismatch and variability as reasoned strategies are bound to vary between individuals, and between population samples.

Another aspect of these data that may provide some insight comes from the fact that some types of anomaly in the original seven *wh*-extraction violations are deviant in much more obvious ways than others. In essence, interpretability may in fact be a function of the edit distance between an anomalous and licit syntactic form. For example, Want for, That-trace and Whether Islands can be corrected by replacement or omission of one-word. In contrast, Adjunct Islands and Left Branch Condition sentences are not easily correctable in the same sense. Along these lines, Hiramatsu noted that some participants had crossed out the "that" in That-trace constructions on their response sheets. This indicates that it was clear enough to some what was meant by the anomalous phrasal type that they felt compelled to

modify the sentence to represent what it meant.

in a series of experiments directed as a response to sentence matching experiments conducted by Freedman and Forster (1985), Crain and Fodor (1987) argue that 'correctability' is a key component to the comprehension of anomalous sentences. They found that participants were slower to respond to correctable 'rule-based' (agreement errors, etc.) and 'constraint-based' (Subjacency violations, etc.) violation matches than their non-correctable counterparts. Of particular interest here is the contrast between constraint violations in (18) and (19).

(18) Correctable

a. Is this the coat which that want to buy?

b. How many dancers did they say that would be in the show?

c. What kind of sauce did Sam serve the pork chops and?

(19) Non-correctable

a. Who do they deny the claim that the burglar shot?

b. Who did you buy Picasso's portrait of?

c. Who is the fact that he lives with well known?

Their argument contends that some types of anomaly are 'spontaneously' corrected by participants in online sentence comprehension. Note that the correctable types contain cases of Want for and That-trace violations and non-correctable types contain a number of instances from Complex NP constructions. This apparent contrast between Complex NPs and Want For sentence types fits well with the findings in the current experiment (and to some degree That-trace constructions) and provides further evidence that there is a qualitative difference between the *wh*-violations tested in Satiation experiments that may provide a perspective on the reported effects hitherto discussed in the literature.

2.6 Summary

To conclude, we began this chapter with a review of the evidence for the Satiation Effect which mounted a case against the replicability of the effect. Against these claims, findings here suggest that the Satiation Effect *is* replicable – albeit somewhat inconsistent. The quite consistent findings that separate Whether Islands and Subject Islands on the one hand, from Adjunct Island and Left Branch Condition structures on the other argue in favor of distinct sources for the surface acceptability of these constructions. What does not corroborate the initial conclusions made by Snyder that rating change reveals a class of syntactically grammatical structures too difficult for working memory to process, is the fact that other sentence types, such as Want for constructions, also show effects that do not demonstrate the same hypothesized processing burdens.

The degree overlap in results between the various investigations argues against an account that is based on test-taking strategies alone. A consideration introduced here points out a potential qualitative difference between satiating and non-satiating sentence types. Satiating types are typically more 'interpretable'. Including interpretability as a factor rekindles the possibility that the explicit, reasoned component of judgment tasks provides participants the opportunity to change their response criterion. Criterion changes over the course of the experiment in this way may develop under particular pressures to equalize the distribution of rating responses to 50/50 in unbalanced designs in a non-spurious fashion.

The key question that remains is, what is the influence of non linguistic-based strategies versus the contribution of linguistic processing on increased rating scores? Chapter 3 aims to address these particular issues by taking a closer look at the parallels between Satiation and a well-documented sentence processing phenomenon, Syntactic Priming.

CHAPTER 3

Repetition effects in sentence processing

This chapter aims to begin to tease apart the potential sentence processing and task-based sources of the Satiation effect. As a first step, the parallels between Syntactic Satiation and the more well-documented phenomenon Syntactic Priming are explored. Evidence for automatic, implicit responses to exposure to syntactic form have been demonstrated in a large, and growing research program. Robust priming effects are found in language production, within and across languages. Yet there are fewer studies on priming in language comprehension and those studies suggest there are some complications in drawing a direct analog for effects in production to effects in language comprehension.

One key point to consider for the present investigation is the observation that there is a general lack of evidence for priming of anomalous syntactic structure which, in turn, raises the question to what extent priming is contingent on licit underlying forms. Another important facet is the degree to which 'mere-exposure' to syntactic form underlies the observed increase in rating scores or whether these differential effects shown in the Satiation literature are particular to decision-based exposure. Evidence from two experiments presented here argues for a non-judgment based measure to gauge the influence of sentence processing changes as a function of exposure in Satiation studies.

3.1 Syntactic Priming

People tend to repeat structure in spoken language, as in (1). Speakers may choose to voluntarily repeat some aspect of language for stylistic reasons (i.e. emphasis) or it may be unintentional.

(1) a. Individual speaker

i. Once you're *in it*, you can't get *out it*.

b. Between speakers

i. I *most certainly am* not

ii. You *most certainly am* are

3.1.1 Language production

Bock (1986) first attempted to explore this phenomenon in an experimental setting looking at syntactic structure in language production, in particular the Voice alternation between *active* (2a) and *passive* (2b).

(2) Voice Alternation

a. The ball hit the boy. (Active)

b. The boy was hit by the ball. (Passive)

In a picture-description task in which a participant listens to a sentence prompt, repeats the sentence and then describes an unrelated event in depicted in a picture, Bock found that participants repeating an *passive* sentence were more likely to describe the picture using an *passive* sentence as well.

Supporting evidence followed for other propositionally equivalent structures, most notably for Dative alternations (3).

(3) Dative Alternation

a. The girl handed an apple to the teacher. (Prepositional Object)

b. The girl handed the teacher an apple. (Double Object)

These and subsequent experiments point to a truly syntactic component to priming effects – demonstrating that priming does not rely on verbal subcategory properties (theta-roles) (Bock and Loebell, 1990) nor the repetition of function words (Bock, 1989). Furthermore, evidence also suggests that facilitory gains provided

by recent exposure can withstand up to ten intervening unrelated syntactic objects (Bock and Griffin, 2000).

Confirming both the observational and experimental evidence, a series of corpus analyses also show that in naturalistic contexts speakers tend to repeat structures that have used or comprehended in the recent discourse context (Bresnan, 2007; Gries, 2005). These studies corroborate another key aspect of the experimental evidence, namely that priming effects appear to be 'long-lasting', persisting despite intervening syntactic material. Furthermore, corpus analyses also provide an opportunity to gauge the extent to which the distributional frequencies of particular syntactic forms influences syntactic choice. Along these lines Jaeger and Snider (2007) report that in addition to recent syntactic exposure speakers are sensitive to the relative frequency of syntactic structure they are exposed to. Their data show that low frequency structures appear to prime more than high frequency structures (Inverse Frequency Effect).[1]

3.1.2 Language comprehension

Evidence for syntactic priming has also been found in sentence comprehension. Noppeney and Price (2004) looked at locally ambiguous early/late closure structures for which there are (temporarily) two distinct parses. In general speakers tend to parse the NP constituent 'the stage' as the object of the main clause verb (late closure (4)) and disprefer structure in which the relevant NP is taken as the subject of the preceding clause (early closure (5)).

(4) [Before the director had left the stage] the play began.

(5) [After the headmaster had left] the school deteriorated rapidly.

[1]Frequency effects in language processing refer to the observation that frequency of occurrence of a linguistic object (lexical items (Scarborough et al., 1977) and sentences (Juliano and Tanenhaus, 1994)) predicts ease of processing. The Inverse Frequency Effect refers to findings that low frequency linguistic objects show more sensitivity (i.e. relative change) to recent exposure than high frequency structures (Forster and Davis, 1984).

This study found that reading times were faster and neural activity was attenuated (fMRI) for dispreferred parses when prime sentences were structurally similar.

Arai et al. (2007) investigated the Dative alternation by recording eye-gaze times in a visual-world paradigm. Their findings show that participants tend to direct their gaze towards the picture of the recipient item after having DO constructions (6), and towards the theme in PO constructions (7).

(6) The assassin will send the dictator the parcel. (DO)

(7) The assassin will send the parcel to the dictator. (PO)

However, in a follow-up experiment Arai et al. (2007) found that these effects appear to be contingent on the repetition of the matrix verb –switching 'send' for 'give' in priming pairs eliminated the effect. A cross-experiment comparison suggests that verb repetition too plays a role in production, enhancing effects. Taken together, evidence points to differences in the robustness of detectable priming in production versus comprehension.

There has also been evidence that exposure to novel syntactic forms can condition priming. Kaschak and Glenberg (2004) found evidence that participants who read text excerpts containing the 'needs' construction (8) showed faster reading times, and higher accuracy on comprehension questions in a subsequent online experiment than participants that had not been previously exposed to the novel construction, rather the standard construction to convey the propositional meaning (9).

(8) Needs construction [2]

   a. The meal needs cooked.

   b. The grass needs cut.

(9) Standard construction

   a. The meal needs <u>to be</u> cooked.

   b. The grass needs <u>to be</u> cut.

These findings were taken as evidence that adult speakers of a language are capable of learning to comprehend new constructions in their native language.

More evidence for priming for novel structures comes from Luka and Barsalou (2005). Their study provides evidence that suggests that exposure to semi-grammatical sentences in a reading task conditions higher rating scores in a judgment task. Whereas participants gave higher rating scores for previously encountered syntactic structures, lexically identical sentences showed an even more robust effect. This evidences corroborates earlier evidence from comprehension of Dative structures that lexical repetition plays a key role in detecting effects in comprehension, but it also underlines the contribution of abstract syntactic structure on priming.

3.2 Implicit learning

The upshot from this review is that priming is observed in both experimental and naturalistic settings and appears to interface with syntactic structure. A leading characterization of syntactic priming is that it is an instance of *implicit learning*, knowledge acquired from exposure is long-lasting, not just carry-over from episodic memory and that learners are not explicitly aware the learning processes, nor what is learned. We have seen that there is evidence from both experimental and corpus studies that suggest that syntactic priming effects do in fact remain active over an extended amount of time and resist the influence of unrelated intervening syntactic forms.

Key evidence that participants do not have conscious awareness of the learning process comes from work with retrograde amnesiacs. Amnesiacs are by definition unable to encode information in explicit memory stores yet Ferreira et al. (2008) found that amnesiacs perform similar to control groups in standard syntactic priming tasks.

It has been suggested that Syntactic priming serves two main functions. First, coordination of speech patterns between interlocutors enables more fluid transmission of propositions and metalinguistic information. Choosing similar syntactic structure helps assure that speakers are making the same assumptions about structure-

meaning mappings. Alignment of speech for metalinguistic ends is known as Linguistic Accommodation (Bell, 1984) – a sociolinguistic function in which speakers align their speech patterns consciously and unconsciously for social aims. On the other hand, syntactic priming also has a decidedly implicit effect that is illustrated by 'Inverse Frequency' findings, namely that low frequency and generally less preferred structures show more priming than high frequency and more commonly preferred structures. In this way, syntactic priming is implicit learning of the frequency distributions of syntactic structures. It has been argued that this strategy works to promote underrepresented structures and avoid over-learning of a given structure (Ferreira and Bock, 2006).

3.3 Is Syntactic Satiation an instance of Syntactic Priming?

In light of this discussion the question is, is the Syntactic Satiation an instance of syntactic priming in comprehension?

The obvious similarity is that both effects include repeated exposure to syntactic structure as a key aspect of the phenomenon. A further link between Satiation studies and work in syntactic priming comes from a series of experiments by Luka and Barsalou (2005) that showed participants that had been exposed to 'semi-acceptable' sentences forms tended to rate sentences with a similar syntactic structure higher than participants that had not. These results were taken as evidence suggesting that implicit processing of syntactic structure can be indexed in a rating task. In a framework developed by Luka (2005) called MAPIL (Metacognitive Attribution in Implicit Learning), improved ratings are viewed as an indirect measure of the associated implicit retention of, and subsequent familiarity with a given syntactic structure. Processing of congruent syntactic form is facilitated (Structural Facilitation) producing a positive affect for a structure. This positive affect is then misinterpreted by participants as an index of grammaticality.

Another common feature is that Satiation and Syntactic priming effects are persistent. Priming in production was first observed for immediate effects of exposure.

Only after questions as to the nature of the priming effect began to be explored did evidence surface indicating that exposure biases could be detected for up to ten intervening unrelated syntactic structures. But for Satiation studies, the methodology integrates the assumption that effects remain despite the exposure to one to eighteen other deviant and/or grammatical forms. [3] On this point, there has been somewhat more difficulty detecting priming in sentence comprehension studies that do not integrate some level of lexical repetition (Arai et al., 2007; Kaschak and Glenberg, 2004; Luka and Barsalou, 2005). In those studies that have found priming for syntactic and lexical similarity, evidence suggests lexical overlap increases the effect.[4]

Furthermore, an apparent similarity between Satiation studies and Luka's work is that the sentence structures considered were anomalous to some degree. However, this similarity appears superficial on closer inspection. The"semi-grammatical" items employed in Luka and Barsalou (2005) do not appear to be anomalous for the same reasons as those sentence types investigated in the Satiation literature (on the whole). Examples of materials employed in this study appear in Table 3.1

I miss having any time to do anything.
A heart courageous never breathed scant.
You should absolutely come and visit me.
In China, wine is served in small cups, and in Turkey, coffee.
Sam recites poems as well as playing the piano.

Table 3.1: Examples of moderately grammatical sentence types in Luka and Barsalou (2005)

This asymmetry highlights the divergent goals of the bodies of work. On the one hand, Luka and Barsalou (2005) aimed at gauging the relative contribution

[3]The number of potential interveners varies based on the block size of the experiment. Another difference between the designs employed in Luka and Barsalou (2005) and in Satiation studies is that the typical measure in Satiation work is the rating change between the first and last two repetition exposures.

[4]Exposure intervals and lexical overlap in fact may contribute to the inconsistent reports of satiating structural types.

of exposure on familiarity – indirectly measured by change in rating scores. This approach did not aim to assess the possibility of obtaining increased rating scores for strictly ungrammatical stimuli. Rather the semi-acceptable status of the target items was necessary only in terms of the logistics of providing room for rating improvement. In fact, the investigators did not include ungrammatical distractor items in the exposure task in a deliberate attempt to avoid drawing attention to the grammatical status of the target items. The goals of Satiation work is quite different: the grammatical status of the test items *is* the object of inquiry.

Sprouse (2007b) suggests that this difference is of key importance in regard to claims about the source of the Satiation effect and its (non)linguistic basis. He argues that Luka and Barsalou's items are not ungrammatical in the technical sense, as is the case in Satiation studies, and that only grammatical structure can show syntactic priming effects. This position emphasizes the distinction between types of anomaly and asymmetrical effects in syntactic priming – assuming that syntactic priming, and facilitated processing can only stem from licit syntactic structure. The prediction, again, is that Satiation effects attested in the literature stem from unbalanced experimental designs and therefore is inherently a non-linguistic effect. Results from Experiment 1 challenge this stance on grounds that the effect should show more instability than is actually attested. However the possibility that uneven balance in rating tasks contribute to Satiation effects cannot be excluded without addressing this issue directly.

Another key prediction that stems from literature in Syntactic Priming concerns the nature of the exposure condition. One defining characteristic of studies on the Satiation effect is that they have exclusively used rating tasks as the exposure task and the dependent measure. Contrastively, priming literature has primarily employed auditory and/or reading tasks as exposure sessions. In fact, Luka and Barsalou purposely avoided exposure to clearly ungrammatical items in the exposure task in order to avoid systematic engagement of structural evaluation strategies on the part of participants. This raises the question, hitherto unexplored in the Satiation literature, – are repetition effects found in exposure contexts that do not

require linguistic intuition and structural evaluation on the part of the subject?

3.4 Summary and predictions

In summary, there is robust evidence that repeated exposure to syntactic form produces facilitory effects in production for a number of propositionally equivalent structures. Yet there is less evidence for an analogue in comprehension. Describing Satiation as an instance of Syntactic priming may also be challenged by the lack of evidence for repetition effects for decidedly anomalous syntactic structure. Those studies that do include anomalous structures have arguably considered forms that are less likely to be framed as syntactically anomalous. Considering this evidence, Sprouse (2007b) makes the prediction that licit syntactic structure is a prerequisite to priming effects. This proposal, in tandem with the Equalization strategy, underlies Satiation effects as task-based effects that arise an artifact of unbalanced experimental designs. Experiment 2 aims to test this prediction by exploring the effects of task bias. If licit syntactic structure is a prerequisite to priming effects and the anomalous *wh*-constructions employed in Satiation studies are in fact ungrammatical, differential rating scores are not predicted in a balanced-design context (standard practice in the Syntactic Priming literature).

Another key question is the extent to which the judgment process itself contributes to rating improvement. According to implicit learning evidence, rating-based exposure is not a pre-condition to facilitated processing. Experiment 3 looks at the potential influence of exposure type on rating improvement. If Satiation is implicit learning of syntactic form, then, reading-based exposure should constitute a comparable exposure environment to sustain detectable priming effects.

3.5 Experiment 2

3.5.1 Methods

Participants

Twenty-two undergraduate students from the University of Arizona enrolled in an introductory psychology course participated in the experiment. All were given course credit for their time. Participants completed a language background questionnaire aimed at assessing their exposure and proficiency in English and familiarity with other languages.[5] Based on the questionnaire data, 17 subjects were considered monolingual English speakers.[6] Participants ranged in age from 18 to 30 years of age (mean 19.5).

Materials and procedure

The experimental items included 85 interrogative sentences. Forty grammatical fillers and forty ungrammatical target items were constructed for the main experiment in order to present participants with a balanced set of theoretically acceptable and unacceptable items. Five additional items similar to those contained in the main session were presented as practice items. The forty ungrammatical items consisted of eight instances of each of five ungrammatical structures used in experiment 1. These types can be found in Table 3.2.

The base sentences items from Snyder (2000) (employed in Experiment 1 as well) were extended and modified for the current experiment to provide three extra repetition exposures and to make lexical composition of the items more homogenous within and across target sentence types. First, the original materials included five

[5]Experiments 2 and 3 originally consisted of two language background populations: monolingual English speakers and bilingual English/Spanish speakers. I only report findings for the monolingual population in this thesis.

[6]5 monolinguals claimed proficiency in another language: Italian, French (2) and ASL(3) and were excluded from analysis. The rubric for the monolingual participant groups can be found in the appendix A.2

Table 3.2: Target Items for experiment 2.

That-trace	Who does Mary think that likes John?
Subject Island	What does John know that a bottle of fell on the floor?
Complex NP	Who does Mary believe the claim that John likes?
Adjunct Island	Who did John talk with Mary after seeing?
Left Branch	How many did John buy books?

instances of each of seven violation types to correspond to five repetition frames. The materials here included eight instances, reflecting an increase in the number of exposures by three. Furthermore, the original materials included a number of potentially salient properties that were modified for the current set of items to follow standard practices in psycholinguistics; in particular the use of uncommon proper names and the uneven distribution of Wh-words employed.

Proper names in the previous item list were replaced with more common names for the participant age range. These names were harvested from the 50 most popular names registered with Social Security from 1985. Previous materials included an uneven scattering of What, Who, How many, Where and Why (23 What, 14 Who, 7 How many, 4 Where, 1 Why). Here That-trace, Subject Island, Complex Noun Phrases and Adjunct Islands were balanced between *What* and *Who* question words. Left Branch violations always employed *How many*. Grammatical controls were also balanced for question word in an equal proportion to target items (32 What, 32 Who, 16 How many). An effort was also made to distribute the main clause verbs across sentence types.

Items were grouped into eight ten-item blocks. Each block of ten items included an instance of each of the five violations together with five grammatical sentences. Eight presentational lists were created by distributing the repetition frame in which individual sentences appeared. Participants were assigned randomly these presentation lists. Repetition blocks were randomized by computer, with a different order for each subject in order to avoid within-block ordering effects.

Participants were asked to give acceptability ratings for the 85 interrogative sentences. Each participant was assigned to one of the eight presentation lists which

was presented on a PC computer. The experimental software DMDX (Forster and Forster, 2003) was used to coordinate the presentation of the items and to record participant responses. Each sentence was displayed individually, on one line, centered on the monitor and in new courier font. A default timeout was set at 30 seconds in order to give the respondent ample time to respond.[7]

The instructions for the experiment were contained within the experimental session. The main components of these instructions mirrored those used in the previous experiment.[8] However, the example sentences were excluded from the instructions themselves in lieu of practice items appended to the beginning of the main experimental session.[9] Participants were instructed to read each sentence and press the `Yes` button for acceptable sentences and the `No` button for unacceptable sentences. The session including consent, questionnaire and rating lasted from 15 to 25 minutes.

The questionnaire responses were entered by hand and response data by PHP script into a database for organization and preparation for data analysis. Both written and electronic data were connected by a participant code assigned at the beginning of the session and recorded on the questionnaire form and the experimental output file as an ID.

3.5.2 Results

Participants' overall ratings for each sentence type can be found in Figure 3.1.

The data were analyzed by both parametric (analysis of variance) and non-parametric (G^2 logistic regression analysis of deviance) statistical procedures. Relevant factors for both analyses were exposure Set(First,Last)[10] and sentence Type(That-trace, Subject Island, Complex NP, Adjunct Island, Left Branch Con-

[7] In fact, most participants never reached the timeout threshold and were otherwise unaware that there was a time limit imposed by the software

[8] These instructions appear in the appendix

[9] These practice items contained two violation types found in experiment 1 but not part of experiment 2 in order to avoid an extra repetition exposure for these particular sentence types

[10] First set included repetitions 1 and 2, and the last set repetitions 7 and 8.

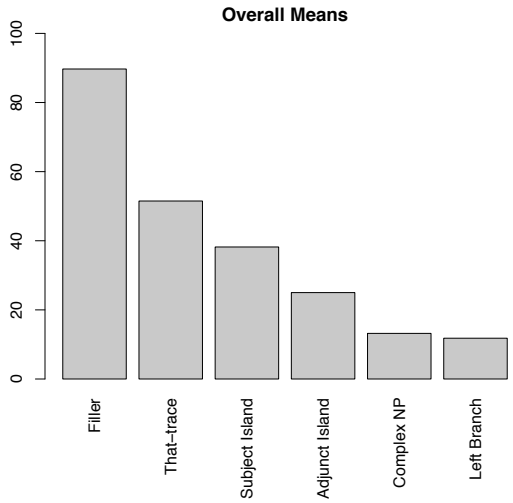


Figure 3.1: Overall mean acceptability for Experiment 2.

dition)[11]. The List(1-8) factor was used as a blocking factor in factorial tests and is not reported.

An overall 2 x 5 ANOVA was conducted revealing main effects for Set [$F(1, 32) = 4.8, p < 0.05$] and Type [$F(4, 80) = 117.9, p < 0.001$] with no interaction between the two. Planned comparison tests for Set for each sentence type revealed a difference between first and last set exposures for Subject Island [$F(1, 32) = 4.9, p < 0.05$] violations. Results from logistic regression measures confirm the parametric results (Subject Island violations ($G^2 = 4.0, p < 0.05$)) and no other effects for the other four violation types were found.

Rating stability was calculated by counting the number of participants that split yes/no ratings in both the first and last repetitions versus the number of participants that responded yes or no throughout the first and last set repetitions. These results are shown in Figure 3.3. The differences between these counts were submitted to Sign Test for each of the five levels of Type. Results found that Left Branch Condition ($p = 0.01$) and Complex NP ($p = 0.02$) sentence types were more stable

[11]Note only target ungrammatical types were considered in the statistical analysis

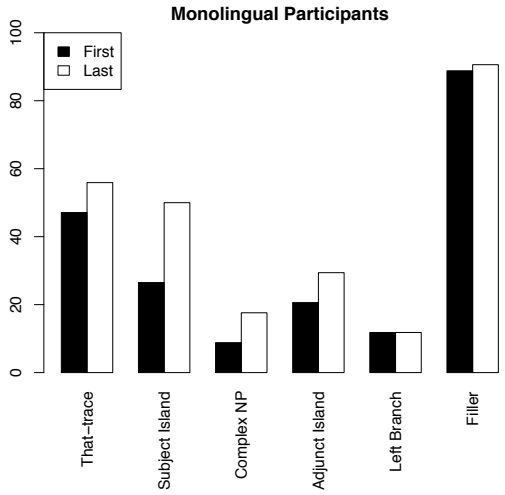


Figure 3.2: Acceptability ratings by exposure set in Experiment 2.

than Subject Islands ($p = 0.12$) and That-trace effect ($p = 0.50$) types. Marginal stability was found for Adjunct Islands ($p = 0.06$).[12]

3.6 Discussion

The results from Experiment 2 show Satiation effects for Subject Islands in a balanced experimental design. This finding pairs well with findings in Experiment 1 given that of the five types considered in the current experiment, only Subject Islands showed effects in both. The Equalization Strategy, in its basic form, does not predict effects in balanced designs. The fact that Subject Islands satiate in both studies conducted here suggests that there is some factor beyond a task-based response strategy underlying the observed effects.

Other evidence against the idea that a response bias is at the root of the Satiation effect comes from the overall mean ratings from first to last repetition set. The first thing to note is that ratings increase for both uneven and equal bias rating tasks. The relative differences between the experiments can be attributed to the

[12]The number of stable/variable pairs fall below best practice for binomial tests and should be interpreted with caution.

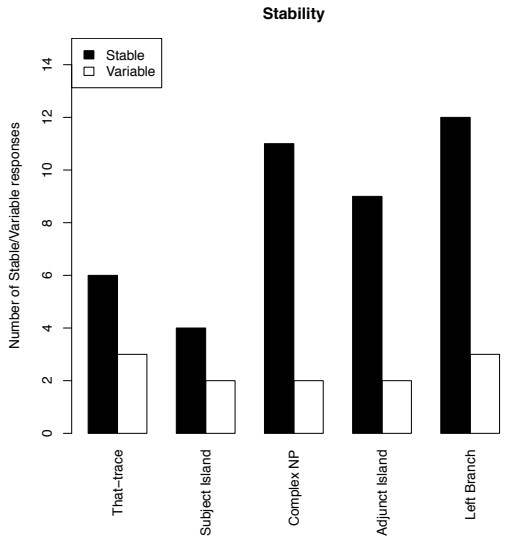


Figure 3.3: Rating stability: number of stable and variable responses in Experiment 2.

fact that only 30% of the items in Experiment 1 are clearly grammatical compared to 50% in Experiment 2. Under the Equalization strategy it seems that participants would be closer to 30% in Experiment 1 in the first repetition set. The fact that participants respond more favorably than is expected on the basis of the grammatical to ungrammatical ratio potentially speaks to some other test taking strategy that influences participants to rate sentences higher throughout the experimental session more generally.

Experiment	First Set (expected)	(observed)	Last Set (expected)	(observed)
1 (Bias)	30.0	49.8	30.0	52.5
2 (Balanced)	50.0	55.9	50.0	61.8

Table 3.3: Mean ratings for First and Last Set for Experiment 1 and 2.

Another comparison to be made between Experiment 1 and 2 is the distribution of relative acceptability. Of the types tested in both experiments, the order of acceptability is the same for That-trace, Subject Islands, Adjunct Islands and Left Branch Condition structures. One difference is the relative ordering for Complex

NPs. In the unbalanced context in Experiment 1, Complex NPs were found more acceptable than the other four types reported here but in Experiment 2 Complex NPs were judged less acceptable – on par with Left Branch Condition violations. Stability measures provide another source of detecting the effect of uneven design contexts. Results from Experiment 1 found Complex NPs the most unstable over the course of the experiment whereas the current experiment Complex NPs were much less variable. Thus, task balance may in fact play a role by contributing to a generalized instability in rating responses in the experimental session.

In short, the evidence from Experiment 2 suggests that Satiation effects are observable in balanced experimental designs, lending credence to the idea that repetition effects found in Satiation studies are related to sentence processing (priming or a memory bottleneck), and not task-bias effects. Contrasting evidence from unbalanced and balanced designs, however, does provide some support for the hypothesis that unbiased contexts do influence rating responses to some degree – just not a direct and clear influence on increased rating responses as put forward by the Equalization Strategy. The next step, explored in Experiment 3, is to address to what extent Satiation effects can be separated from the judgment process itself. If rating improvement is an index of syntactic priming then effects should also be obtained through reading-based exposure.

3.7 Experiment 3

3.7.1 Methods

Participants

Twenty-eight undergraduate students participated in the experiment. Undergraduates were students from the University of Arizona and were either enrolled in an introductory psychology course and received course credit for their participation. Participants completed a language background questionnaire aimed at assessing their proficiency in English and exposure to other languages. Based on the questionnaire data, 10 participants were excluded for proficiency in a language other

than English.[13] Participants ranged in age from 18 to 23 years of age (mean 19).

Materials and procedure

The materials were the same as used in experiment 2 (rating task).

Each participant was assigned to one of the eight presentation lists on a computer monitor. The experimental software DMDX (Forster and Forster, 2003) was used to present items and record participant responses. In contrast to experiment 2, the current task asked participants to provide acceptability ratings for the first (1st) and last (8th) repetition blocks.

In the second through seventh exposures participants were asked to perform a word-by-word reading task known as the Maze task (Freedman and Forster, 1985; Nicol et al., 1997; Forster et al., 2009)[14]. In essence the Maze task is a decision task in which participants are presented two word options such that one word continues a grammatical sentence and the other does not. A typical presentation for a sentence is represented in Figure 3.4.

Typically the presentation of a sentence is halted and feedback is displayed when the wrong lexical item is chosen. However, the current formulation aimed to exploit the underlying reading aspects of the Maze paradigm. Therefore, the options presented were reduced to word/non-words, every sentence continued to be displayed regardless of the response and no feedback was given for incorrect choices in order to ensure exposure to the entire sentence for all participants.

The questionnaire and response data were prepared and organized as in experiment 2. And as in the previous acceptability rating tasks, the first and last responses for each of the violation types in each list served as the basis for relative rating change over the experimental session. However, only the first and eighth item

[13] 10 monolinguals claimed proficiency in another language: ASL(3), French, Jgbo, German, Greek, Hebrew, Italian, Korean and Portuguese. The rubric for the monolingual participant groups can be found in the appendix

[14] The Maze task first used by Freedman and Forster (1985) and has been used to gauge sentence processing effects.

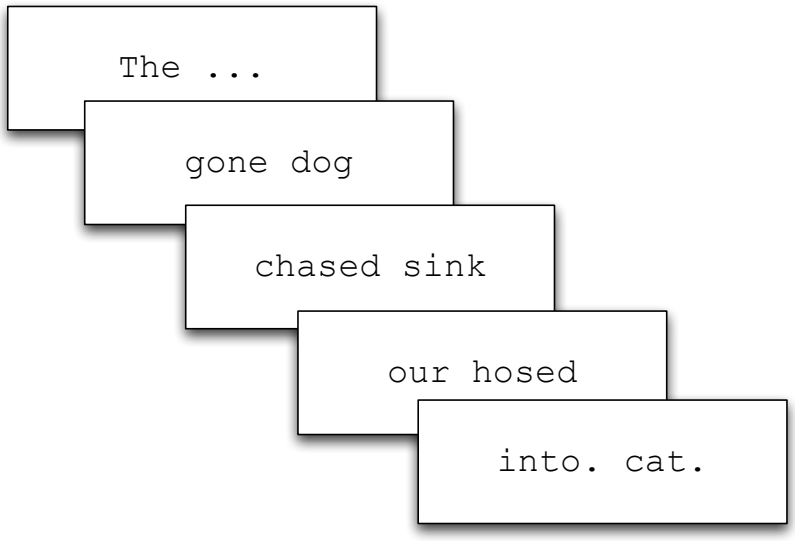


Figure 3.4: Illustration of the Maze task (presented one line at a time).

blocks were used in the final analysis.

3.7.2 Results

Response were evaluated by the same factorial groupings and statistical measures as in Experiment 2.[15] Means for the relative acceptability for each sentence type can be seen in Figure 3.5.

An overall ANOVA considering the factors Set and Type was conducted. Results reveal main effects for Type $[F(4, 170) = 43.7, p < 0.001]$ and not for Set. One-way ANOVAs revealed no differences in for the mean acceptability of the sentence types tested. Similarly, G^2 measures showed no differences as well. In order to assess the effect of exposure on rating responses, planned comparisons were run for Set by each sentence type. No significant effects were detected.

Given first and last set scores only contained one rating score rating stability was not calculated. The number of participants for increased, decreased and stable

[15]Note that the factor Set with the levels 'First' and 'Last' are synonymous with the first and eighth repetition exposures given that the second through seventh exposures took place in the reading task and no rating score was taken.

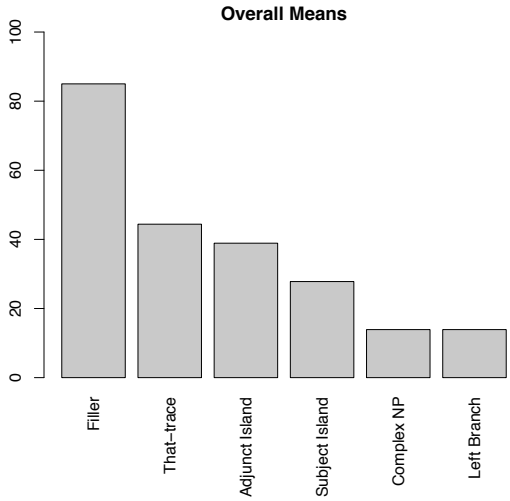


Figure 3.5: Overall mean acceptability for Experiment 3.

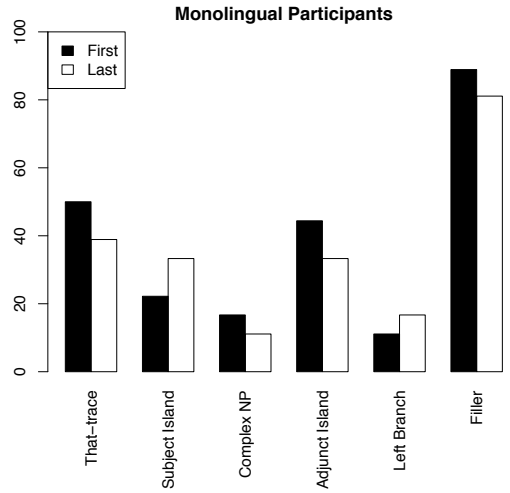


Figure 3.6: Acceptability ratings by exposure set in Experiment 3.

	That-t	Subject Island	Complex NP	Adjunct Island	Left Branch
Increase	6	4	1	3	1
Decrease	5	4	3	6	2
Stable	7	10	14	9	15

Table 3.4: Rating stability: number of increased, decreased and stable responses in Experiment 3.

judgment patterns appear in Table 3.4

3.8 Discussion

Results from Experiment 3 found no significant effects for rating change in a reading-based exposure task. This finding, on the surface, lends favor to the account that Satiation effects are particular to rating-based exposure and potentially not driven by mechanisms underlying Syntactic priming. On the other hand it may be the case that the Maze task is not an adequate exposure task to sustain syntactic priming effects. Although the task has been proven to detect processing differences (i.e. subject/object relative clause asymmetries Forster et al. (2009)), it is not clear that the relevant processing mechanisms are engaged that support syntactic priming when comprehension of sentences is word-by-word, as this type of comprehension is not 'natural' in any typical sense. It is also plausible that a decision task that incorporates non-word options further complicates typical comprehension from proceeding naturally. Moreover, it also appears to be the case that evidence from syntactic priming in comprehension has been somewhat more elusive to detect in general compared to evidence from production – varying from task to task, measure to measure and often relying on lexical repetition to show robust effects.

Other evidence that the Maze task is not a well-suited exposure task to detect syntactic priming effects comes from the distribution of overall mean ratings for Experiment 3, seen in Table 3.8. Rating means for Experiment 3 drop instead of increase, highlighting a potential unexpected interaction between the reading exposure

Experiment	First Set (expected)	(observed)	Last Set (expected)	(observed)
1	30.0	49.8	30.0	52.5
2	50.0	55.9	50.0	61.8
3	50.0	58.9	50.0	53.9

Table 3.5: Mean ratings for First and Last Set for Experiment 1, 2 and 3.

carried out via the Maze task and judgment ratings. Both previous experiments report increased acceptability from the first to last repetitions. Although it may have been expected that a reading exposure task would not provide a suitable context for increased ratings, a drop in overall ratings is not predicted on any account.

3.9 General discussion

Two experiments in this chapter aimed to gauge the non-linguistic, task-based strategies on the observed Satiation effects. First, experiment 2 looked at the hypothesis that unbalanced tasks induce Satiation effects observed in the literature. Findings show that despite balanced experimental design, Subject Islands show Satiation effects. These results show consistencies with Experiment 1 in terms of the types of constructions that undergo Satiation effects, the overall rating bias from first to last repetition set and relative acceptability ordering. Neither the consistency in overall mean rating change, nor specific sentence types is predicted by the Equalization Strategy. However, null results were found in Experiment 3 which aimed to address the extent to which Satiation effects are contingent on rating-based exposure. It was hoped that results from reading-based exposure would be more telling as to the sentence processing strategies associated with the Satiation effect.

A line of inquiry opened in Chapter 2 designed to ground the task-based explanations for Satiation effects was to consider the influence of interpretability and spontaneous correction on rating change. One observation is that participants naturally try to interpret sentences in rating tasks even against formal instruction to focus on structural acceptability. This hypothesis received support from the general

qualitative differences between more readily interpretable satiating types and less interpretable non-satiating types. Again here in Experiment 2 there is corroborating evidence for a distinction between satiating types based on interpretability. However the current evidence for Satiation comes from a balanced experimental design. If meaning is a relevant factor, it appears to be applied under repetition in general regardless of the biased ration between grammatical and ungrammatical test items.

3.10 Summary

The goal of this chapter was to begin to isolate the potential sources of the Satiation effect. Evidence from two experiments directed at addressing the potential task-based and processing-based influences on rating change were conducted, but revealed mixed results. On the one hand a significant Satiation effect was found in a balanced-design rating task. This finding suggests that task bias alone cannot explain improved rating as a function of exposure – however results do suggest that bias may contribute to a generalized instability across sentence types. A second experiment aimed at assessing the extent to which rating change is dependent on the repeated rating exposure and the judgment process in general found no significant change in acceptability for any of the five sentence types tested. Considering the data from overall acceptability for the experimental session, the adequacy of the word-by-word task to serve as an exposure task for Satiation effects is challenged. This result, in tandem with little evidence supporting robust priming effects in sentence comprehension in general suggests that exposure type plays a critical role in producing detectable effects.

In sum, the evidence from this chapter argues for a corroborating source of evidence to distinguish the relative contributions of the decision-making process and language processing on the Satiation effect. Chapter 4 aims to fill this gap by coordinating Satiation effects in an offline rating task with data from online reading time measures.

CHAPTER 4

Coordinating rating and reading time measures

The current chapter explores the hypothesis that rating improvement for satiating sentence types is due to facilitated processing by providing evidence coordinating rating and reading time measures. Given the variable and composite nature of the intuition process, data linking the Satiation effect to processing in a more standard manner is key in assessing the relative contribution of the judgment process and sentence processing on acceptability. To address this distinction a two-part experiment was conducted in which participants were exposed through a rating task to a set of syntactically anomalous sentences in grammatical and ungrammatical bias sets and subsequently participated in a self-paced reading session.

Results from the rating task again show Satiation effects regardless of grammatical bias contexts, as in Experiment 2, and some indication that the interpretability of anomalous types co-varies with Satiation. Reading time data suggest participants' exposed to satiating syntactic anomaly in the judgment task are faster to read these types according to self-paced reading measures. The findings here make an important methodological contribution to the connection between the syntactic processing and linguistic intuition for unacceptable sentence types, and underline the complexity of the linguistic introspection and the difficulty in isolating any one performance factor as the sole source of Satiation effects.

4.1 Mitigating variability in rating tasks

There are a number of influences ultimately bear on an acceptability judgment rating. These influences are both linguistic and non-linguistic in nature. It is clear that at some level, every acceptability judgment reflects the grammatical status of syntactic structure. There is little doubt that the clear-cut difference between the

surface (un)acceptability in (1) derives from an assessment of grammatical well-formedness.

(1) a. Colorless green ideas sleep furiously.

b. * Furiously sleep ideas green colorless.

What is less clear is when the grammatical status of a linguistic structure is obscured by some other factor or factors other than syntactic well-formedness.

Experimental syntax proponents suggest that some of these factors can be mitigated with controlled data collection techniques. Typical confounding factors purported to be diminished include task, and informant variability. For example, the presentation order of sentences to be evaluated can have an influence on acceptability ratings. Greenbaum (1976) found that when the sentence in (4) was judged in coordination with (2) it was consistently found more grammatical than with (3).

(2) We didn't dare answer him back.

(3) We dared not answer him back.

(4) We didn't dare to answer him back.

This indicates that structural comparison between test items can influence the evaluation process. Experimental approaches attempt to reduce the effects of sentence ordering by implementing counter-balancing and randomization techniques typical in group testing practices. This approach aims to distribute potential effects that arise due to direct comparisons across all participants.

Another type of variability found to affect acceptability ratings deals with the mental state of informants. Carroll et al. (1981) found that participants that provided judgments under 'objectively aware' (viewing oneself as others may) conditions tended to use meaning and communicativeness as a criterion to assess grammaticality, whereas participants that were 'subjectively aware' (viewing oneself as an unreflective participant) assessed linguistic anomaly in response to their structural prop-

erties despite receiving the same formal instructions.[1] Informal elicitation settings can vary widely and potentially bias participants in terms of objective/subjective mental states. On the other hand, experimental methods typically employ standardized procedures to diminish differences in the session from one participant to the next, addressing this type of intraspeaker variation. Any session-based effect on participants that may potentially have an influence on ratings, including mental state, is held constant for all informants.

Interspeaker variation perhaps poses the largest challenge to informal data collection. People differ in many ways, some linguistic and others non-linguistic. More often than not, informal methodologies include relatively small participant samples. Dialect and metalinguistic differences between participants, if not balanced by larger sample sizes, increase the possibility that any given participant contributes something unexpected to a rating score. In general, experimental approaches are better equipped to deal with this type of variability through the the use of larger population samples. Moreover, experimental methods often elicit questionnaire data from informants. This data can be used, post hoc, to distinguish the potential source of differences between participants (dialect, age, education, linguistic training, etc.). Informal methods often downplay the importance of these variables, in part due to the fact that informal methods are ill-equipped to account for variance associated with small samples sizes.

4.2 Variability introduced by experimental methods

Task and informant variability has been the area where advances of experimental approaches to acceptability judgments have been most substantial; reducing extraneous variability and revealing contrasts previously undetected. Evidence from Satiation, in terms accounted for by Snyder (2000), suggests that rating change throughout the experimental session provides an additional perspective on variabil-

[1] An objective mental state was induced by positioning a mirror facing participants in the booth where ratings were collected.

ity associated with language processing. However, two alternative approaches raised in the light of investigations on Satiation effects propose that experimental methodologies may inadvertently introduce, instead of reduce, confounding factors and that these factors have been misinterpreted as evidence for detecting processing effects in judgment tasks. One potential factor, noted in the Equalization strategy hypothesis, concerns the decision-making process and the influence of experimental design in testing conditions. Sprouse (2009) notes that the majority of Satiation studies have implemented designs in which the experimental materials were predominately anomalous. Based on a series of replication studies in which no Satiation effects are found in balanced designs, he proposes that Satiation effect reflects a task-based strategy in which participant attempt to equalize the number of 'yes' to 'no' responses across the entire experimental session.

This hypothesis is of key importance to assessing Satiation effects for the current investigation, but also makes a more general claim about experimental approaches to syntax, namely that the particular configuration of an experimental design itself can systematically alter rating responses.

A second potential confounding factor concerns the informants typically recruited to participate in experiments. Arguments have been made that in order to avoid experimenter bias, acceptability judgments should not performed by linguists themselves. As such, experimental syntax has been typically based on data from naive participants. This of course has its advantages, beyond the mitigation of experimenter bias influences (Hawthorne Effect), by reducing the aforementioned individual differences inherent to human subjects. However, naive participants – untrained in syntax, by definition – appear to be susceptible to confusing structural well-formedness for meaningfulness, or typicality (Maclay and Sleator, 1960). As noted in Chapter 2, on the surface those sentence structures that have shown Satiation effects in the literature are more readily interpretable than non-satiating structures. The Interpretability approach raised here suggests that participants actively or passively adopt a meaning-based criterion despite formal instruction to evaluate the structural properties of sentences as a function of repeated exposure.

Consequently, the benefits from employing naive subjects in experimental settings may come at a cost.

Together the Equalization strategy and Interpretability approach raise the possibility that participants search out task and sentence material properties, and use this information to predicate acceptability ratings. However, a key prediction is that both these strategies are particular to the explicit nature of rating tasks. As these factors may contribute to more variability in Satiation studies reviewed and conducted here, neither strategy is predicted to show effects outside judgment-based exposure, nor rating responses.

4.3 Assessing processing effects

The main question of theoretical interest still remains; do Satiation effects provide the linguist a new tool to distinguish processing effects from syntactic effects? It is difficult to say given the evidence supporting processing-based Satiation may itself be obscured by the variable nature of linguistic intuition and acceptability tasks. What is needed is a corroborating measure for Satiation effects from a task more typically employed to measure online processing effects.

Evidence of this type is limited. The only direct investigation of the extent to which rating improvement in a judgment task can be indexed by an independent processing measure was conducted by Braze (2002, 2003). Braze investigated the possibility of finding a non-judgment based Satiation effect in an eye-tracking experiment. Two of the seven original Island violations where selected from Snyder's 2000 paper; Whether-Islands and Adjunct Islands. These were selected due to the fact that Whether Islands showed the most robust effects, and Adjunct Islands the least. Grammatical control versions of each violation were constructed, as seen in (5a) and (6a).

(5) Whether Items

a. When did Valerie wonder whether Don had liked the story?

b. What did Valerie wonder whether Don liked?

(6) Adjunct Items

a. When did Fraser clean the bathroom while Judy mopped the kitchen?

b. What did Fraser clean the bathroom while Judy mopped on Saturday?

Participants read a context sentence and then pressed a button to get either the violation or control sentence (counter-balanced across subjects). Half of the sentence sets presented contained a comprehension probe to assure that participants were reading for comprehension. Results showed that Whether Island showed differential reading times compared to controls as a function of exposure, whereas Adjuncts experienced no change. These findings suggest that Satiation effects do have a processing correlate.

Braze frames these results in terms of memory limitations, and underlines the notion that differential responses to repeated exposure indexes processing asymmetries – along the lines proposed by Snyder (2000). However, considering the mixed findings from Satiation studies reviewed in this thesis it is difficult to determine if improved ratings reported for all satiating sentence types stem from processing burdens. Want for and That-trace effects have no clear processing account yet still show effects in various studies. Furthermore, if the influence of exposure is processing based, then the question arises, why don't we find consistent Satiation effects for Complex NPs and to some extent Subject Islands in experiments that test them?

Although Braze's experiment makes inroads to connecting judgment-based Satiation to differential effects from measures more typically employed to gauge processing effects, this study does not coordinate the same participants in both studies. As it stands if Satiation effects are found in a rating task, and independently found in a reading task it does not follow that the same factors stem both effects. It is potentially the case that some effects that influence ratings in a judgment task and/or a reading task are task-specific. Coordinating the same participants would provide an important perspective on effects in a rating task and their subsequent effects in a reading task.

4.4 Summary and predictions

Considering the multi-faceted nature of acceptability ratings a clear understanding of the source of improved ratings as a function of exposure is difficult to attain from the current methodologies (ratings as exposure and measure). Given the claims regarding the source of the Satiation effect it makes intuitive sense to assess both the influence of task bias on ratings in coordination with measures gauging the influence of bias versus exposure through measures more typically applied to gauge processing effects. Experiment 4 and 5 aim to fill this gap by conducting a two-part experiment in which the same set of participants provide data for rating and self-paced reading tasks.

If task bias is an operative characteristic of the Satiation effect, ungrammatically biased contexts should produce more Satiation effects than grammaticality biased contexts. Evidence for a strong influence of experimental design was mitigated in Experiment 2, but given the detected effect on stability there is reason to suspect that bias can attribute more variability overall in an experimental session – and as a consequence produce more Satiation effects than non-bias and reverse-bias contexts. If the ability for participants to interpret certain sentences drives Satiation effects, then, all else being equal, exposure, not task balance should produce improved rating scores.

Experiment 4 aims at assessing the relative contributions of these processes on rating scores but also serves as the exposure task for a subsequent self-paced reading task (Experiment 5). Claims that Satiation reveals processing asymmetries that stem from syntactic priming predict detectable reading time asymmetries for satiating versus non-satiating sentence types as a function of exposure, regardless of bias context. The Memory Bottleneck account makes an even more specific prediction in that only those sentence types hypothesized to be anomalous due to working memory demands should show differential processing effects (Subjacency violations).

4.5 Experiment 4

4.5.1 Methods

Participants

One hundred and thirty-eight undergraduate students participated in the experiment. Undergraduates were enrolled in an introductory psychology course at the University of Arizona and received course credit for their participation. Participants completed a language background questionnaire aimed at assessing their proficiency in English and exposure to other languages, primarily Spanish. Based on the questionnaire data, 69 subjects were considered monolingual English speakers with no other relevant language background experience. Participants ranged in age from 18 to 24 years of age (mean 18.8).

Materials and procedure

Each participant rated a total of 55 interrogative sentences; 50 main session sentences and 5 practice items. To address the potential influence of the ratio of grammatical to ungrammatical sentences on the stability acceptability ratings two lists were created in which one (Grammatical bias) contained 3 ungrammatical sentences for every 10 and another (Ungrammatical bias) containing 7 ungrammatical for every 10 items. These 10 item groupings formed 5 repetition frames.

Given that this task also served as an exposure task for a subsequent self-paced reading task two other lists were created for each bias list. One list (A) contained That-trace, Complex NP and Adjunct Island violation types, the other (B) contained the Want for, Whether Island and Coordinate Constraint structures. For the two Grammatical bias lists these were the only anomalous sentence types to appear. Four additional violation types were added to the Ungrammatical bias lists Needs Construction, Infinitive Subject, Subject Island and Left Branch. Those sentence types that did not appear in previous experiments appear in Table 4.1.

The structure of the four treatment lists is seen in Tables 4.2 and 4.3.

Coordinate Constraint	Who did Kate meet new member and for dinner?
Needs Construction	What does David say Amber believes needs cut?
Infinitive Subject	Who was to tell the results at the exam cruel?

Table 4.1: Additional violation types in Experiment 4

Set A	Set B
That-trace	Want for
Complex NP	Whether Island
Adjunct Island	Coordinate Constraint
Grammatical filler (x 7)	...

Table 4.2: Grammatical bias

Each treatment list was balanced for repetition frame which created 20 presentation lists. Sentences within each frame were randomized automatically on presentation to avoid potential ordering effects.

Items that appeared in previous experiments were reused here with minor modification. The length of the sentences was modified for each sentence type in order to assure that the reading windows explored the subsequent self-pace reading task were of equal number. An effort was made to distribute the tense of the auxiliaries used (i.e. *do*, *did*, *will*). New items also conformed to these modifications and other characteristics take into account in previous materials creation.

Participants were asked to give acceptability ratings for the 55 interrogative sentences. Each participant was assigned to one of the 20 presentation lists. The experimental software DMDX (Forster and Forster, 2003) was used to present items and record participant responses. Each sentence was displayed individually, on one line, centered on the monitor and in new courier font. A default timeout was set at 30 seconds in order to give the respondent ample time to respond.[2]

The instructions for the experiment were contained within the experimental session. The main components of these instructions mirrored those used in the previous

[2]In fact, most participants never reached the timeout threshold and were otherwise unaware that there was a time limit imposed by the software.

Set A	Set B
That-trace	Want for
Complex NP	Whether Island
Adjunct Island	Coordinate Constraint
Needs Construction	...
Infinitive Subject	...
Subject Island	...
Left Branch	...
Grammatical filler (x 3)	...

Table 4.3: Ungrammatical bias

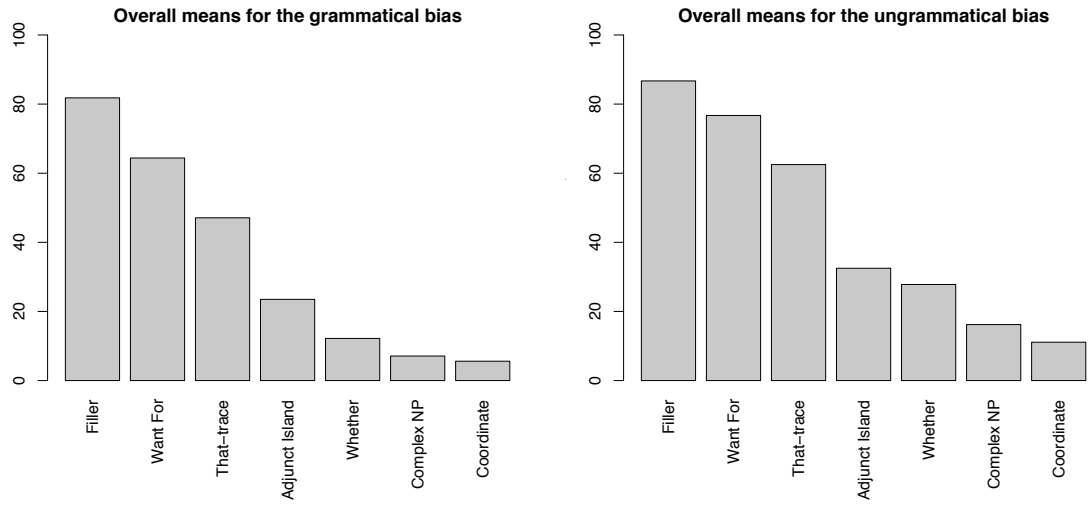
experiment.[3] However, the example sentences were excluded from the instructions themselves in lieu of formal practice items appended to the beginning of the main experimental session. Participants were instructed to read each sentence and press the `Yes` button rating for acceptable sentences and the `No` button for unacceptable sentences. The session including consent, questionnaire and rating lasted from 15 to 20 minutes. Furthermore, participants were made aware that there participation included two independent sessions on the computer.

The questionnaire responses were entered by hand and response data by PHP script into a database for organization and preparation for data analysis. Both written and electronic data were connected by a participant code assigned at the beginning of the session and recorded on the questionnaire form and the experimental output file as an ID. Those participants whose error rates on comprehension questions in the subsequent reading task exceeded 20% were excluded from the data analysis in both Experiment 4 and 5.

4.5.2 Results

Factors Type(That-trace, Complex NP, Adjunct Island, Want For, Whether, Coordinate), Bias(Grammatical, Ungrammatical), Set(First, Last) and the blocking group List(1-5) were used in both parametric and non-parametric tests as in previ-

[3]These instructions appear in the appendix A.2.



(a) Grammatical

(b) Ungrammatical

Figure 4.1: Overall acceptability for bias conditions in Experiment 4.

ous experiments.

A 2 (Bias) x 7 (Type) (x 2 (List)) ANOVA revealed main effects for Bias $[F(1, 138) = 23.6, p < 0.001]$ and Type $[F(10, 759) = 220.6, p < 0.001]$ with no interaction between them. Follow up comparisons by Bias for each sentence type was conducted. Results showed that That-trace $[F(1, 138) = 4.0, p < 0.05]$, Whether Islands $[F(1, 138) = 6.9, p < 0.01]$ and grammatical fillers $[F(1, 138) = 6.2, p < 0.05]$ differed significantly in grammatically and ungrammatically biased contexts with both Want For $[F(1, 138) = 3.8, p = 0.05]$ and Complex NP $[F(1, 138) = 3.5, p = 0.06]$ violations showing marginal effect. Deviance scores from a G^2 test also found a a contrast for That-trace ($G^2 = 223.4, p < 0.05$), Whether Islands ($G^2 = 173.2, p < 0.01$) and grammatical fillers ($G^2 = 1563.0, p < 0.05$) as well as the marginal effects found for Want For ($G^2 = 214.9, p = 0.07$) and Complex NP ($G^2 = 114.3, p = 0.06$) sentence types. No other contrast was found by either measure.

Turning to the influence of repeated exposure, a 2 (Set) x 7 (Type) ANOVA was conducted collapsing over bias lists and then grammatical and ungrammatical bias conditions independently. The results for all bias contexts combined show main

effects for Type $[F(10, 748) = 181.0, p < 0.001]$ and Set $[F(1, 136) = 8.2, p < 0.01]$ and no two-way interaction. Follow up contrasts between Set and each level of Type revealed effects for Want For $[F(1, 136) = 4.6p < 0.05]$ and Whether Island $[F(1, 136) = 7.8p < 0.01]$. G^2 scores also found effects for Want For $G^2 = 172.0, p < 0.05$ and Whether Islands $G^2 = 137.2, p < 0.01$.

For grammatical bias conditions the Set x Type ANOVA also revealed main effects for Set $[F(1, 68) = 8.3, p < 0.01]$ and Type $[F(6, 238) = 134.4, p < 0.001]$ and no interactions. One-way ANOVA comparisons for sentence types by Set revealed a significant difference for Whether Islands $[F(1, 68) = 5.7, p < 0.05]$ – reflected in the G^2 test as well ($G^2 = 44.6, p < 0.05$). The ungrammatical bias condition showed a main effect of Set $[F(1, 66) = 87.7, p < 0.001]$ but not Type nor their interactions. But mixed results were obtained for simple effects in the one-way ANOVA and the G^2. Coordinate constructions were significant in the parametric $[F(1, 66) = 2.4, p < 0.05]$ but not in the non-parametric test ($G^2 = 47.9, p = 0.13$). Of note were the marginal effects for Want For, Complex NP and Coordinate structures and the overall distribution of trends in repetition effects, seen in Table 4.5 and 4.6.

Bias	First Set (expected)	(observed)	Last Set (expected)	(observed)
All	50	54.00	50	58.10
Grammatical	70	62.70	70	68.40
Ungrammatical	30	45.00	30	47.50

Table 4.4: Mean rating change across types by bias condition

Given the relatively small sample size a series of one-tailed t-test were conducted to follow up the results for the one-way ANOVAs for each level of type for Set. Results for the grammatical bias condition confirm the results of earlier test, but also found the marginal effect for Complex NP to be significant ($t(42.3) = -1.72, p = 0.05$). For the ungrammatical bias condition, Whether Islands ($t(67.1) = 1.83, p < 0.05$) reached significance and both Want for ($t(67.1) = -1.39, p = 0.09$) and Coordinate ($t(85.1) = -1.5, p = 0.07$) sentence types nudged

Type	F score	p level	Dev score	*Chi*	t score	p level
Want For	2.7	0.11	89.4	0.13	-1.5	.07
That-trace	0.66	0.42	93.2	0.46	-.7	.24
Adjunct Island	0.79	0.38	75.8	0.40	-.8	.20
Whether Island	5.7	0.02	44.6	0.02	-2.3	.01
Complex NP	3.2	0.08	37.4	0.08	-1.7	.05
Coordinate	0.34	0.56	24.6	0.55	.58	.71

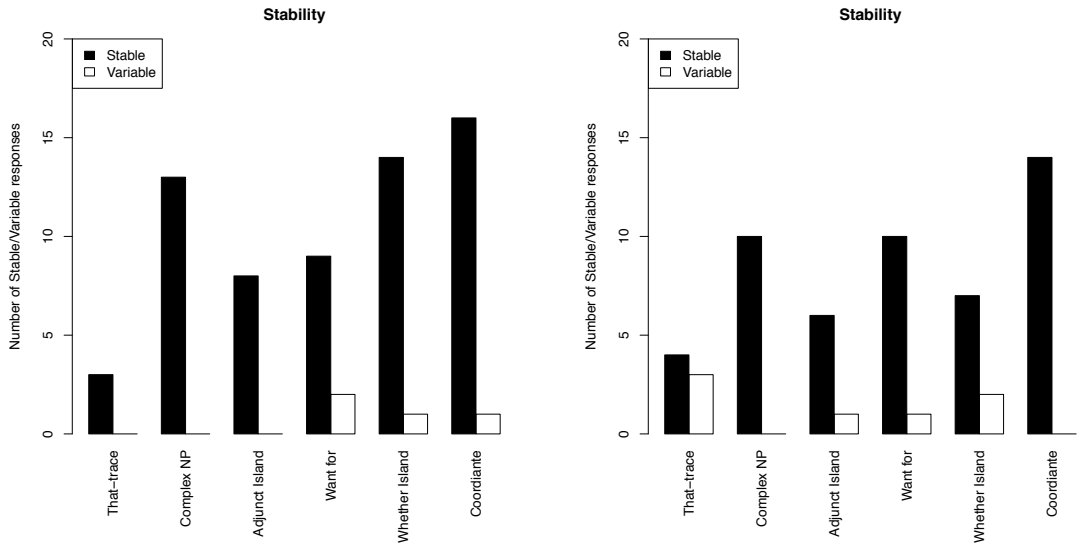
Table 4.5: Complete list of variance and deviance scores for the grammatical bias condition in Experiment 4. (**Bold** - strong effect, Underline - marginal effect)

Type	F score	p level	Dev score	*Chi*	t score	p level
Want For	2.6	0.11	76.6	0.16	-1.4	.09
That-trace	0.27	0.61	84.7	0.61	-.5	.31
Adjunct Island	0.06	0.80	83.5	0.79	.25	.60
Whether Island	3.5	0.07	83.6	0.07	-1.8	.04
Complex NP	0.14	0.71	51.9	0.72	-.35	.64
Coordinate	2.9	0.03	47.9	0.13	-1.5	.07

Table 4.6: Complete list of variance and deviance scores for the ungrammatical bias condition in Experiment 4. (**Bold** - strong effect, Underline - marginal effect)

closer to significance than in the two-tailed analysis.

Rating stability was calculated for both the grammatical and ungrammatical conditions as in previous rating tasks. Results for grammatical bias for each of the six levels of type found that Complex NP ($p = 0.000$), Adjunct Islands ($p = 0.01$), Whether Islands ($p = 0.000$ and Coordinate Constructions ($p = 0.000$) sentence types were more stable than That-trace ($p = 0.25$), types. Marginal stability was found for Want for types ($p = 0.07$). Ungrammatical bias showed instability for That-trace ($p = 1$), Adjunct Islands ($p = 0.13$), Whether Islands ($p = 0.18$) and more stability for Complex NP ($p = 0.001$), Want for ($p = 0.01$), Coordinates ($p = 0.000$).



(a) Grammatical bias

(b) Ungrammatical bias

Figure 4.2: Stability measures for bias conditions in Experiment 4.

4.6 Discussion

Evidence for Satiation effects was found in both grammatical *and* ungrammatical bias conditions. This evidence suggests that even under conditions that are hypothesized to influence participants to reduce acceptability as a function of exposure (not increase), Satiation effects are observed. Notably, strong effects are found for Whether Islands in both ungrammatical and grammatical bias conditions. Whereas Want for constructions show marginal effects in both conditions, they are also highly grammatical overall suggesting Want for sentences experience a ceiling effect in which ratings are capped by the overall acceptability, rather than by weak Satiation effects. Taken together, Satiation effects do not appear to be directly contingent on task bias.

The evidence for an interpretability account for Satiation effects is also limited. First, although Whether Islands, Want for and Complex NP Islands are qualitatively interpretable, Coordinate constraint violations are not. Furthermore, it is not clear why effects are not found for That-trace constructions, which appear to be more

meaningful than Coordinate constructions. Secondly, an interpretability approach as framed, predicts exposure to be the key factor in producing Satiation effects. However, of the four sentence types that show Satiation patterns only two (Whether Islands and Want for violations) occur in both bias contexts, suggesting task bias does play a role in rating improvement to some degree.

A point of consistency, both bias conditions show the same relative ordering in overall scores. Considering the findings in Experiment 1 and 2 in which relative acceptability appeared to be affected by task bias, participants in Experiment 5 responded quite consistently despite the bias mismatch (Figure 4.1).

A clear difference between bias conditions, participants tended to rate anomalous sentences higher generally in ungrammatical contexts. Despite this overall asymmetry the increased rating is distributed across the entire experimental session and does not arise exclusively in the second portion of the task (Table 4.5.2). Furthermore, neither bias condition showed a difference in the number of sentence types that were stable/unstable, in contrast to findings from Experiment 1 and 2. However, task bias does appear to affect rating stability to some degree.

The most striking results pertain to the unexpected Satiation effects for Complex NP and Coordinate Constraint forms. On the one hand, Complex NP violations have shown effects in previous studies – however not in the current work. But unexpectedly, effects are attested in grammatical contexts, and not in ungrammatical contexts. This finding is odd from the standpoint of consistency between experimental sessions. If a sentence type satiates in one condition and not in another chances are the effect stems from more variable participant related factors and not from exposure alone. Furthermore, this finding does not bode well for the case for bias-induced satiation effects in that Equalization predicts effects only in ungrammatical condition sets – against current findings.

On the other hand, Satiation effects for Coordinate constraint constructions is not predicted on any account. Coordinate forms were selected in this manipulation to pair against another decidedly syntactically anomalous structure (Adjunct Islands). Overall acceptability for Coordinates, and for Complex NPs for that mat-

ter, are extremely low attesting to the robust unacceptability for these structures. Increased acceptability for these highly unacceptable structures suggests that participants are not evaluating grammatical structure alone, in particular for Coordinates.

Processing accounts do not find much support either. Although the Memory Bottleneck hypothesis correctly predicts findings for Whether Islands, and the marginal effect for Complex NPs, findings for Want for and Coordinate structures are not expected. An Syntactic Priming approach may make inroads in accounting for Whether Islands, Complex NPs and Want for sentence types as grammatical, and therefore 'prime-able', but again, Coordinates to do fit clearly into this picture. In addition, processing accounts must also account for the variability found between bias contexts. All else being equal both processing approaches predict effects regardless of bias.

These results further highlight the fact that rating tasks as exposure and measures do not provide a straight-forward way to tease apart the underlying contributions to Satiation effects. A novel perspective on Satiation effects from reading time measures is explored in Experiment 5.

4.7 Experiment 5

4.7.1 Methods

Participants

All participants in Experiment 4 also completed the current experiment as the second part of a two part session.

Materials and procedure

Participants read 60 target sentences. Half of these were ungrammatical sentences from the previous rating task session. Any given participant had seen half of these. The other half were novel grammatical fillers, none of which participants saw in the previous task. The ungrammatical sentences were of 6 distinct types, corresponding

to each of 3 violation constructions from set A and set B in the previous session, as listed in Table 4.7.

Type	Example
That-trace	What did Lisa believe that frightened Kyle's wife?
Complex NP	What will Robert hate the fact that Sarah is buying?
Adjunct Island	Who will Heather talk about Ed after seeing?
Want for	Who will Ursula want for Brandon to kiss?
Whether Island	What will Mason wonder whether Rachel discovered?
Coordinate Constraint	Who does Cathy see the pediatrician and every month?

Table 4.7: Ungrammatical Sentence Types in Experiment 5

One of each of the 6 violation types appeared together with 6 grammatical fillers, matched for salient surface distributional characteristics as in previous manipulations. These items were grouped into 5 blocks. For each block a presentation list was created in order to distribute the relative contribution of any given sentence across all exposure positions. Furthermore, sentences were randomized within each 12 item block to diminish ordering effects.

A series of 30 comprehension questions of two types were also created; true/false (7) and word verification (8).

(7) True/false

   a. What did Richard ask whether Eleanor preferred? (Test sentence)

   b. Was Richard concerned about Eleanor's preferences? (Comprehension question)

(8) Word verification

   a. Who did Jacob claim that voted for Andrea? (Test sentence)

   b. Did the word 'voted' appear in the sentence? (Comprehension question)

The appearance of a comprehension question and the type of question was distributed across all grammatical and ungrammatical sentence types equally. These questions were randomized in parallel with the target sentences within each repetition frame.

After completing the rating portion of the experimental session, participants were asked to read 60 interrogative sentences on a computer screen. Participants were randomly assigned to one of the 5 presentation lists which was displayed on a computer monitor and coordinated by the experimental software DMDX (Forster and Forster, 2003). The words in each sentence initially appeared with each character replaced by an asterisk (*). Pressing the `RIGHT` button on the button box revealed the sentence two words at a time. Yes/No comprehension questions also appeared pseudorandomly, to which the participant was asked to press `YES` or `NO` (also corresponding to `LEFT` and `RIGHT` buttons). Target sentences and comprehension questions appeared individually, on one line, centered and in new courier font. This portion of the experimental session lasted from 10 to 15 minutes.

Responses from the questionnaire completed before the rating task, the rating task data and the reading task data were entered into a database for organization and preparation for data analysis. Both written and electronic data were connected by a participant code assigned at the beginning of the session and recorded on the questionnaire form and the experimental output file as an ID.

4.7.2 Results

Response times (RTs) collected were prepared for analysis in the following manner. First, error rates were calculated for each participant. Participants whose error rates for comprehension questions exceeded 20% were excluded from the data analysis.[4] Second, outliers were adjusted by replacing all RTs that were +/-2 standard deviations from each participant's own mean reading time for critical regions with the value of either 2 standard deviations above or below. Thus, the high/low cut-off points were dynamic and varied according to individual reading patterns.

These scores were then transformed into residual reading scores. Following stan-

[4] As noted in section 4.7.1, this is the same error rate used to exclude participants in Experiment 4, the exposure task for the current experiment. Rating data and RTs were collected and analyzed for the same participants. Thus, only data from participants that participated in both sessions at a minimum level of engagement were used for analysis.

dard procedure for neutralizing the effect of the varying number of characters in any given reading window, the expected reading time for a given window length of each window, for each type, for each participant (Ferreira and Clifton, 1986; Trueswell and Tanenhaus, 1991; Phillips, 2004) was calculated. The difference between the observed and expected reading times formed the residual score.

Finally, two regions were identified for each sentence type as the hypothesized point at which processing difficulty would be most pronounced. Data for regions before and after these critical regions were excluded from analysis. Mean residual response latencies for these critical regions from lowest to highest are shown in Table 4.8.

Types	RTs	Critical Regions
Complex NP	0.66	What will Robert hate [the fact - that Sarah] is buying?
Want For	27.0	Who will Ursula want [for Brandon - to kiss?]
That-trace	36.7	What did Lisa believe [that frightened Kyle's wife?]
Whether	38.9	What will Mason wonder [whether Rachel - discovered?]
Adjunct Island	47.7	Who will Heather talk [about Ed - after seeing?]
Coordinate	53.9	Who does Cathy see the pediatrician [and every - month?]

Table 4.8: Residual Response Times for critical regions in Experiment 5.

These data were submitted to a 2 x 2 x 7 (x 5) repeated-measures ANOVA with Exposure(Novel, Familiar) a as within-subject and within-item factor, Type(That-trace, Complex NP, Adjunct Island, Want For, Whether Island, Coordinate) as a within-subject and a between-item factor and Bias(Grammatical,Ungrammatical) as between-subject and within-item grouping factor. List (1-5) was included as a blocking factor and not reported.

The overall analysis revealed no main effects for Type, Exposure or Bias in the subjects analysis, but main effects were observed for Bias ($[F_1(1,136) = 2.1, p = 0.15]$; $[F_2(1,136) = 61.31 = p < 0.001]$) and for Exposure ($[F_1(1,136) = 0.0, p = 0.99]$; $[F_2(1,136)4.9 = p < 0.05]$) in the by items test. Significant interactions between Type * Bias occurred in both analyses ($[F_1(5,408) = 2.72, p < 0.5]$; $[F_2(5,408) = p < 0.05]$) and Type * Exposure in the item analysis ($[F_1(1,68) =$

$0.9, p = 0.34]$; $[F_2(1, 68) = 13.1 = p < 0.001]$). No two-way interaction was found for Bias * Exposure, nor a three-way interaction between the main manipulations.

Planned comparisons between familiar and novel exposure types revealed significant contrasts for Want For ($[F_1(1, 68) = 5.7, p < 0.05]$; $F_2(1, 68) = 35.4, p < 0.001]$), Whether Island ($[F_1(1, 68) = 3.8, p = 0.05]$; $[F_2(1, 68) = 25.1, p < 0.001]$) and Coordinate ($[F_1(1, 68) = 5.3, p < 0.05]$; $F_2(1, 68) = 13.9, p < 0.05]$) constructions in both analyses. In addition, the by items analysis also detected a reliable contrast for Complex NP ($[F_1(1, 68) = 0.6, p = 0.43]$; $F_2(1, 68) = 12.3, p < 0.05]$) structures.

Given the interaction for Bias and Type in the overall ANOVA a follow-up series of pairwise contrasts for each sentence type was performed. Differences in reading times for participants that saw Complex NP $[F_1(1, 31) = 4.8, p < 0.05]$; $[F_2(1, 31) = 16.1, p < 0.05]$ and Adjunct Island $[F_1(1, 31) = 4.8, p < 0.05]$; $[F_2(1, 31) = 30.7, p < 0.001]$ violations in distinct grammatical bias contexts in the rating task were found to be significant as were Whether Island ($[F_1(1, 35) = 1.2, p < 0.27]$; $[F_2(1, 35) = 24.0, p < 0.001]$) and That-trace ($[F_1(1, 35) = 2.0, p < 0.16]$; $[F_2(1, 35) = 17.6, p < 0.05]$) constructions in the item analysis only.

4.8 Discussion

Results show differential reading times as a function of exposure for Whether Islands, Want for and Coordinate Constraint violations, and a marginal effect for Complex NP violations. This suggests that there is some degree of correlation between improved rating scores and reading times. The same four structures showed significant effects, or trends toward Satiation effects in the previous rating task. Both the Equalization and Interpretability accounts predict task-specific Satiation effects and do not predict this level of carry over between tasks.

The outcome for processing accounts is less obvious. On the surface, effects for Whether Islands and Complex NPs provides positive evidence in favor of the Memory Bottleneck hypothesis. These two structures are the only two violation

types subsumed by the Subjacency condition. However, differential reading times found for Want for and Coordinates downplay the potential that Satiation effects tap into, and mitigate working memory burdens as the only source of Satiation effects.

The wide variety of sentence types that show effects makes it difficult to explain Satiation in terms of Syntactic Priming as well. The case could be made that carry-over effects for Whether Islands, Complex NPs and Want for violations stem from implicit learning effects, but findings for Coordinates complicate this story, if licit structure is a pre-requisite to syntactic priming effects.[5]

Task bias appears to play a part in reading times as well. Complex NPs and Adjunct Islands show significant effects in both subject and item tests, and Whether Islands and That-trace constructions showed significant effects only in the items analysis. No interaction was found between Exposure and Bias, which suggests that differential reading times occur independent of exposure. In other words, Complex NPs and Adjunct Islands were read faster if the participant had participated in a rating task that was biased toward ungrammatically even if the participant had not seen a Complex NP or an Adjunct Island violation. The only structures not to show differential reading times for Bias were Want for and Coordinate violations. Of interest here is the fact that these structures both showed the fastest (Want for) and slowest (Coordinate) reading times overall. This suggests that the grammatical bias of a rating task can have a generalized effect on reading times. Supporting evidence for this notion comes from reading times for grammatical distractors. Grammatical distractors, novel constructions for all informants, were read faster after having participated in a ungrammatical bias condition.

[5]It is not clear from priming literature to what degree licit structure refers to complete sentence objects, sub-trees or phrases.

4.9 General discussion

4.9.1 Task-specific effects

Reading time data suggest that Satiation effects, to some degree, *do* carry over from rating to reading tasks. What is more, a variety of anomaly types show facilitation effects – complicating a straight-forward adoption of any single hypothesis hitherto explored. On the one hand, task-specific explanations for Satiation cannot account for corroborating effects between independent tasks. The Equalization strategy predicts that decision-making influences provide participants with a rationale to change their response criterion, critically towards the latter part of the experimental session. Effects found in both ungrammatical *and* ungrammatical bias conditions argue against this approach. Furthermore, Satiation effects appear to have an analog in reading time measures, strengthening the argument against Equalization as the primary source of rating change in Satiation studies.

Another potential task-specific influence on Satiation effects that was considered was the role of interpretability. The observation that satiating sentence types tend to be more readily interpretable than non-satiating types suggested that repeated exposure to anomalous constructions could contribute to improvements in rating scores as naive participants confuse interpretability with grammaticality. The influence of interpretability is predicted to be limited to rating tasks, corroborating effects from reading time measures is counter-evidence to this proposal. Furthermore, this formulation of the influence of interpretability on acceptability ratings makes the prediction that repeated exposure, regardless of task bias, should produce Satiation effects. A mismatch between the types of structures that show effects in the grammatical and ungrammatical bias conditions is additional evidence against Interpretability.

4.9.2 Processing-based effects

Although predictions made by the task-specific accounts explored here are not confirmed, processing accounts of Satiation also find less than convincing support. In

terms of corroborating effects between rating and reading tasks both processing accounts fair about the same on the surface. The Memory Bottleneck account correctly predicts effects in both rating and reading tasks for Subjacency violations (Whether Islands and Complex NP structures) but not the effects found for Want for and Coordinates.

A Syntactic Priming interpretation of Satiation effects can account for Whether Islands, Complex NPs *and* Want for structures if we assume that these structures are grammatically well-formed and [6] are marked in terms of familiarity, influencing initial acceptability ratings and recent exposure increases familiarity of the structure. An explanation for Coordinates cannot proceed in the same way. Coordinates are robustly unacceptable, and cannot plausibly be viewed as syntactically well-formed under any syntactic theory.

4.9.3 Residual issues

In light of the discussion of the current findings, there are two main questions that remain: 1) What drives the apparent mismatch in satiating types between bias conditions in rating tasks? 2) What underlies Satiation effects in rating and reading tasks for Coordinate constructions?

Rating task

The first step in addressing these issues is to address the question of why a number of participants find Coordinate constructions acceptable on repetition. The position advocated here: Coordinates constructions are interpretable. On the surface this claim appears unfounded: Coordinates are *not* interpretable. There is a clear difference when considering the extent to which Coordinates (9), and other satiating types, such as Whether Islands (10) are readily interpreted as meaningful.

[6] The assumption made here is that if Whether Islands and Complex NPs show Satiation effects in both rating and reading tasks independent of bias condition – one of the two processing accounts for Satiation of both which claim that effects are predicated on licit syntactic structure, must be operative.

(9) Coordinate Construction

a. Mark will play the guitar and the cello during the concert.

b. *What$_i$ will Mark play the guitar and t$_i$ during the concert?

(10) Whether Island

a. Eric asked whether Ted had visited the Eiffel Tower.

b. *What$_i$ did Eric ask whether Ted had visited t$_i$?

A key observation that makes Want for, That-trace and Whether Islands relatively more readily interpretable is the fact that these structures are also easily 'correctable'. Through one-word deletion or replacement each of these structures is well-formed.

(11) What did Mark want (for) Vanessa to buy? (Want for)

(12) Who did Jacob claim (that) voted for Andrea? (That-trace)

(13) Who does John wonder (whether/if) Alice blamed? (Whether Island)

Unexpectedly Coordinate constructions are also correctable in a similar fashion. A review of Crain and Fodor's (1987) study on correctability in sentence matching reveals that Coordinates were employed as a correctable structural constraint. In there formulation the replacement of 'and' for 'with' renders the ungrammatical structure grammatical, seen in (14).

(14) a. *What$_i$ will Mark play the guitar **and** t$_i$ during the concert?

b. What$_i$ will Mark play the guitar **with** t$_i$ during the concert?

This correction does not seem as readily recognizable as Want for, That-trace and Whether Island corrections. But once applied, the structure (and subsequent tokens of the same structure) is readily interpretable, albeit with a very distinct propositional meaning.

(15) Coordinates

a. What did Angela drink a shot with at the party?

b. Who does Cathy see the pediatrician with every month?

c. What did Mark steal the bread with last week?

The difficulty in recognizing the grammatical alternative to Coordinates is similar to detecting the alternative readings for some grammatical structures. Note that (16) can readily be perceived as grammatical. (17) can be initially found ungrammatical, if 'smoked' is taken as a participle, but as adjectival modifier for 'pot' the ungrammatical structure is clearly acceptable - although marked.

(16) Tom has been smoking pot.

(17) *Tom is having smoked pot.

McCawley (1982) noted that:

> ... the extent to which comical virtuosity is required for a person to notice the existence of such an interpretation, as contrasted with the ease with which one recognizes its acceptability once it has been pointed out, shows that the strings of words on which grammatically judgments are allegedly made exist only as typographical or acoustic objects, not as perceptual or cognitive objects, just as the Necker cube exists only as a graphic object and not as a perceptual object, ... (p. 79)

So what encourages participants to search out an alternative, correctable version of Coordinates and not apply the same strategy for more clearly correctable That-trace structures? The pressure is potentially two-fold. First, Satiation effects for Coordinates only appear in the ungrammatical bias condition. Participants, under pressure from a response bias interpret the sentence as a typographical mistake, and choose to respond to the sentence based on the corrected form instead of the presented form. In the grammatical bias condition, the pressure to search out and apply an alternative criterion is reduced and in turn, there is no rating improvement.

All else being equal, grammatical bias should also provide sufficient context for That-trace constructions to be recognized as correctable, but the evidence here suggests that, if this does occur, it does not translate into improved rating scores.

A particular characteristic of the experimental design that is different for Coordinates and That-trace structures concerns the other sentence violation types contained in the presentation list. Coordinates were one of three target structures in Set B that appear to be deviant by one word. Conversely, That-trace structures are the only structure in Set A that are one-word correctable. In essence a correctability strategy is 'bootstrapped' by contextual influence given a number of other anomalous sentence types are correctable in similar ways. On this view, a coordination of task bias, and composition of the other sentence materials in the exposure set induces pressure to find correctable alternatives of a certain type.

Reading task

If the interpretable-by-correction approach is on the right track, we have an explanation for why Satiation is found for Coordinates, only in ungrammatical bias contexts and not found for That-trace effects regardless of bias context. However, an interpretability account is hypothesized to be an explicit process in which a combination of task bias and reasoning-based strategies encourage participants to respond more favorably to correctable sentence types over the course of the experimental session. Consequently no effects are predicted to carry over to a reading task. Nevertheless, all structures proposed to satiate due to correctability show reading time effects as well. But how does the apparent connection between rating and reading time measures fit into the larger picture of repetition effects and mitigated memory burdens?

A first step to address the conflict is to look at the sentence types that undergo satiation in grammatical bias conditions. Key to the interpretable-by-correction approach is external pressure to search out alternative structures – ungrammatical bias provides the rationale but the grammatical bias condition does not. Whether Islands and Complex NPs satiate in the grammatical bias condition, suggesting that a correction bias is not the source of these findings. Viewed in this way, rating

change for Subjacency violations suggest that memory limitations can be detected in judgment tasks as originally reported by Snyder (2000).

More difficult to establish are rating and reading time effects for Want for and Coordinate constructions given neither seem plausibly memory-based in the straightforward way that Subjacency violations are. On the one hand, effects for Want for sentences may be construed as an instance of Syntactic Priming. Motivation for this view comes from overall acceptability ratings. Want for structures are not judged to be anomalous to the same degree as other sentence types. Furthermore Satiation effects for Want for constructions are found independent of Bias condition, and therefore appear to be conditioned directly by exposure.

On the other hand, Coordinate sentence types are found very unacceptable and a syntactic priming account, in which judgment and reading Satiation stem require syntactic structure, unlikely. Another approach is to explore the degree to which some aspect(s) of the interpretable-by-correction strategy may carry over to a reading task. One approach to Coordinates is to assume that participants are actively carrying out full parsing routines, and instead differential reading times stem from implicit learning. Where syntactic priming is operationalized as contingent on licit syntactic structure, implicit learning represents the internalization of more abstract patterns –not necessarily grammatical structure. Attentional focus on correctable patterns during the rating task potentially provides a structure-to-meaning association that potentially enables faster reading times. It is important to note that implicit learning may be operational in the same fashion for other correctable types as well. The net effect: stronger effect sizes in terms of judgment and reading Satiation as a function of the number and strength of the strategies in operation.

Another approach to explain exposure effects for Coordinates in the reading task is to suggest that participants recognize instances of Coordinates, and explicitly deploy a non-natural parsing routine through the course of reading. In other words, participants that develop a interpretable-by-correction strategy in the rating task consciously identify ungrammatical-but-correctable types and abort a full parse attempt. Note that this operation may be applied to all 'correctable' types (Want

for, Whether Islands) more generally. This approach to reading times would predict lower reading latencies as a function of previous exposure. [7]

4.10 Summary

Experiments 4 and 5 aimed to assess the influence of task-based processes and sentence processing-based processes on Satiation effects. This approach was designed to complement earlier research on a potential processing analog to judgment Satiation. The initial evidence from Braze (2002) supported the Memory Bottleneck account to Satiation effects, yet only two constructions were tested (Whether and Adjunct Islands). The current investigation expanded the investigation to include a more diverse set of Island violations that have shown Satiation effects in the literature.

Results suggest that framing Satiation effects in terms of any one source explored in this chapter is misguided. On the one hand, evidence suggests that the Equalization strategy, although not directly supported, may contribute to criterion changes based on particular properties which are salient in the other sentence materials. On the surface, these findings would suggest that Satiation effects arise due to inadvertent characteristics of judgment tasks including task bias and the particular lexical and structural similarities (or dissimilarities) of the sentences employed. But Satiation effects for Whether Islands and Complex NPs in grammatical bias conditions highlight the potential that memory limitations in terms originally formulated by Snyder (1994, 2000) may still underly rating change in experimental designs – potentially obscured by the aforementioned task confounds.

Nevertheless Satiation effects for Want for and Coordinate violations cannot follow from a strict implementation of memory limitations, and are more difficult to explain in a unified way. Findings here do underline that participants both actively

[7] Another point of interest, Coordinates do not show an interaction effects for Bias and Exposure conditions. In other words, previous exposure regardless of the grammatical bias condition predicts reading time differences. This suggests that despite the lack of Satiation effects in a grammatically biased rating task, some participants may recognize the correctability of Coordinates, yet do not change their judgment criterion to reflect it.

and passively implement a variety linguistic and non-linguistic strategies in rating tasks as it is available to them under the particulars of the experimental design. Furthermore, results point to 1) areas of further investigation in to the influences of task, informant and materials on acceptability scores and 2) the importance of corroborating evidence to better refine our understanding of the variables that interact in the introspective process.

CHAPTER 5

Concluding Remarks

The aims of this dissertation have been threefold: 1) to explore evidence for the Syntactic Satiation effect and its source(s), 2) to assess the claim that Satiation effects highlight gains of the Experimental Syntax program and 3) to further contribute to an understanding of the nature of linguistic introspection and the relationship between grammaticality and acceptability.

Results from experiments presented here find robust Satiation effects for judgment tasks and varied effects in reading tasks. A distinction emerges between Whether and Subject Islands on the one hand, that consistently show rating improvement in multiple experiments and Adjunct Islands, Left Branch and That-trace violations on the other which do not show effects in any of the tasks in this thesis. This striking contrast between 'weak' and 'strong' violations, on the surface, provides further evidence in support of Snyder (2000) that low initial acceptability for satiating sentence types stems from working memory demands and not syntactic constraint.

Despite these robust findings, there is evidence that a straight forward account of Satiation effects in terms of working memory demands is misguided. Effects for Want for constructions in Experiment 1, 4 and 5 are not predicted on a memory account. It is important to note, however, that the anomaly demonstrated for these sentence types is quite different from other anomalous constructions explored in this thesis and more characteristic of the 'semi-grammatical' sentences employed in Luka and Barsalou (2005). Robust Satiation effects in conjunction with high overall acceptability suggest that rating improvement for Want for sentences stems from facilitated syntactic parsing, akin to Syntactic Priming. Approaching findings for Want for constructions from this perspective dovetails nice with the Luka (2005) proposal that rating tasks can index priming in comprehension. Nevertheless, these

findings highlight the influence of at least two distinct processing effects underlying rating improvement in judgment tasks.

These robust findings suggest judgment tasks, performed experimentally, can detect differences in sources of anomaly. However, evidence from variable findings for some sentence constructions across judgment tasks and null results for reading-based exposure complicate the hypothesis that rating improvement in judgment tasks can be used as diagnostic test to separate syntactic constraint from processing constraint. Two potential accounts for the varied replicability found for Wh-violations in Satiation studies suggest that increased rating scores arise as artifacts of particular features of some rating task designs. Sprouse (2007d) proposes that Satiation effects arise as a by-product of ungrammatically-biased experimental designs that bias participants to accept sentences more readily in the later-half of the experiment in an attempt to 'equalize' the total number of 'yes' to 'no' responses across the entire session. A test-based approach such as this one makes inroads in explaining mismatches in findings across Satiation studies and does appear to covary with rating (in)stability (in particular for Complex NP violations), yet as an inherently non-linguistic strategy more variability is predicted than is found.

A novel observation documented in this thesis is that consistent satiating *and* non-satiation types show asymmetries in overall interpretability. To some extent satiating types appear to be more readily 'correctable' and/or 'interpretable' than non-satiating types. Initially introduced as a way to restrict the predicted variability of the Equalization strategy, the Interpretable-by-correction approach also incorporates another source of variability associated with judgment tasks: despite explicit instruction participants tend to confuse 'communicative' with 'grammatical'.

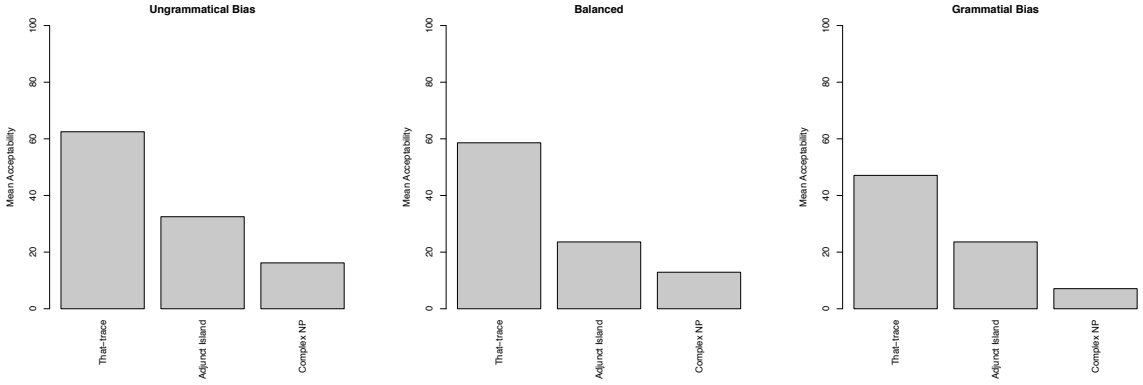
Satiation effects found in Experiment 4 and 5 for Coordinate Constraint structures, robust syntactic violations, suggest that participants responses can be influenced by item-based biases; strategies developed according to the particular characteristics of the sentence types that appear in the testing session. This type of confound may work independent and/or in conjunction with test-based strategies such as the Equalization strategy to produce unexpected rating changes over the

course of an experimental session.

Another prediction of processing-based accounts of Satiation that is not borne out is that effects should be contingent on exposure, not the judgment task itself. Evidence from reading-based exposure in Experiment 3 in which participants were exposed to Wh-constructions in a word-by-word reading task suggest that mere exposure is not the underlying source of Satiation effects. However, a strict interpretation of the null results found for reading-based exposure as evidence against Satiation as either one of the processing effects discussed should be tabled for further investigation. Although the Maze task has been shown to detect reliable effects for well-documented processing effects, such as subject/object extraction asymmetries and garden path effects (Forster et al., 2009), it is not clear that word-by-word segmentation of a sentence is adequate to tap into the mechanisms underlying priming and/or working memory operations to the degree necessary to detect Satiation effects. Furthermore, the extent to which the Maze task can support priming in sentence comprehension more generally need be established before data from investigations into exposure to anomalous sentence types can be accommodated.

In sum, findings from this thesis suggest that judgment tasks can detect reliable differences across experiments, yet there are aspects of the testing environment that also can produce variability. So, what does this mean for Experimental Syntax? One the one hand consistent findings from overall acceptability scores between experiments suggest that group testing procedures are effective in providing the tools necessary to mitigate the more variable sources of non-grammatical noise in rating tasks. Standardized sessions, participant sampling and counter-balancing techniques do aid in capturing graded differences, and replicating them. Illustrated in the consistency in relative differences and mean acceptability found across experiments in this thesis, seen in Figure 5.1.

On the other hand, this thesis adds to a literature that suggests procedures adopted in experimental approaches to judgment tasks are not perfect, and highlight areas for further work. Moreover, if not addressed certain design characteristics of experimental methodologies themselves can inadvertently introduce confounds,

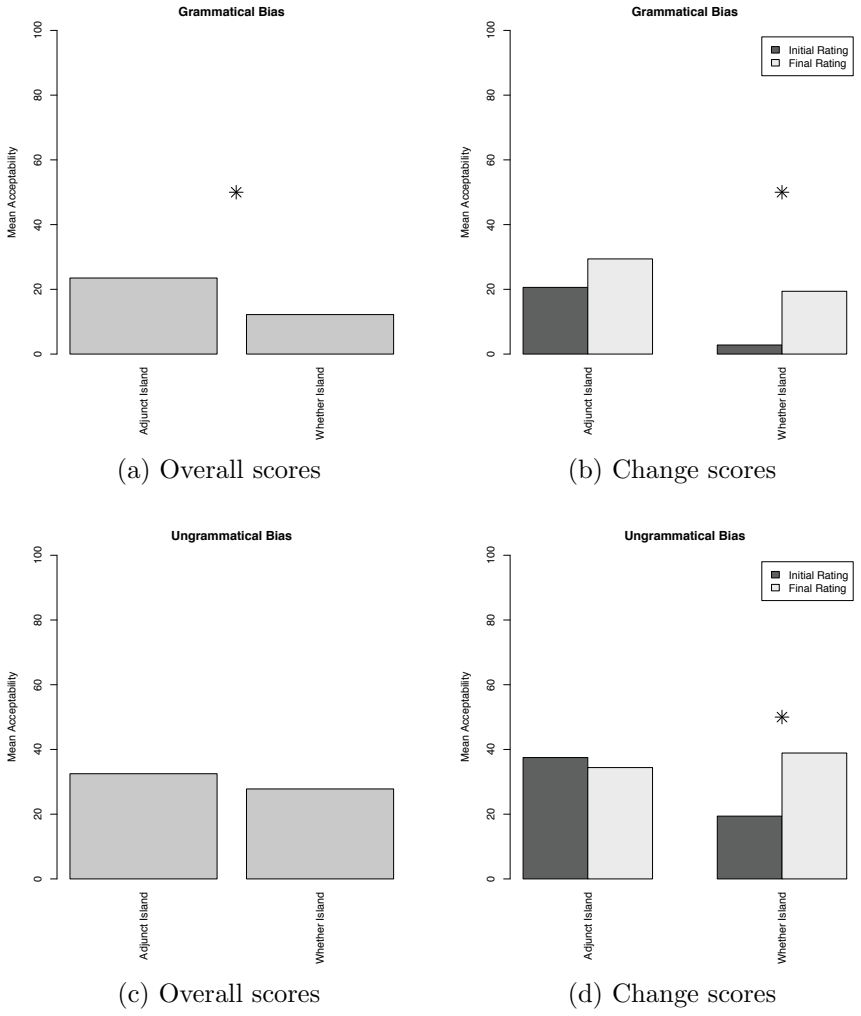


(a) Experiment 4b (b) Experiment 2 (c) Experiment 4a

Figure 5.1: Relative contrasts across experimental tasks.

wrongly interpreted as syntactically relevant. Whereas typical experimental approaches to syntactic inquiry explore the overall mean differences in acceptability between syntactic structures, Satiation studies provide a unique perspective on the relative differences between initial and final ratings for participants. As seen here, there is a potential that overall differences in relative acceptability are colored by particular interactions between task, material and participant-based non-grammatical biases that may develop over the course of the experiment. Focusing solely on the overall acceptability scores, in (5.2a) or (5.2c) in Figure 5.2 for example, misses a potentially revealing difference between the stability of the rating scores as a function of exposure, seen in (5.2b) and (5.2d) respectively. What is more, finding a significant difference, or not, between these two sentence types may be contingent on task design (explored here) and not exclusively on the underlying grammatical status of these constructions.

Findings from experiments in this dissertation point to test and item-related biases as active extra-grammatical confounds in judgment tasks. Furthermore, this evidence suggests that acceptability data are not reliable without a better understanding of extra-grammatical influences on rating scores and how these strategies interact in the testing session. This assertion should not be viewed, in the author's opinion, as a point against the Experimental Syntax program. Rather, the identi-



(a) Overall scores (b) Change scores

(c) Overall scores (d) Change scores

Figure 5.2: Overall and change scores for acceptability judgments.

fication of areas in which experimental methodologies need development reflects an important step forward in refining group testing approaches to linguistic intuition. If our understanding of the introspective process and the relationship between competence and performance systems is to proceed, it surely must include the rigorous development of methods and reporting practices typical in experimental approaches to doing syntax.

Finally, this thesis highlights the importance of taking into account multiple approaches to syntactic inquiry. Linguistics is fundamentally a mentalist enterprise, and as such any single data collection method is inherently limited in scope and alone cannot provide an unfettered view into the total contents of linguistic knowledge. Adopting techniques and methods developed in psychology is at the heart of the Experimental Syntax program. But exploring other techniques, methods and practice offers the syntactician opportunities to revisit familiar problems from new perspectives. In sum, this thesis advocates for a more 'experimental' Experimental Syntax in which linguistic intuition is but one of a host of tools employed to garner data on tactic linguistic knowledge.

APPENDIX A

Experiment materials

A.1 Experimental items

A.1.1 Experiment 1

Want For

What does Barbara want for Jim to invest in?
What does Bob want for Vanessa to buy?
What does George want for Mary to give Biff?
What does Sally want for Andrea to have?
What does Ursula want for Sam to do?

Whether

What did Mildred ask whether Ted had visited?
What did Richard ask whether Eleanor would prefer?
What does Henry wonder whether George discovered?
What does John wonder whether Mary fixed?
Who does Mary wonder whether John met?

That-trace

Who does Bob believe that killed the millionaire?
Who does Fred believe that frightened Bob?
Who does Mary claim that voted for Reagan?
Who does Mary think that likes John?

Who does Susan think that admires Alfred?

Subject

What does Edwin think that a stack of was discarded?
What does John believe that a claim about will be discussed?
What does Marvin think that a picture of will appeal to Molly?
What does Pierre admit that a bag of is missing?
What does Sally know that a bottle of fell on the floor?

Complex NP

What does Margo accept the idea that Felix will do?
What will Sven claim that Ingrid took?
Where did Alex believe the claim that Pablo had put the gold?
Who did Raj whisper that Margaret knew?
Who does Bob like the book that Ted gave?

Adjunct

What did Alice talk to Ed after seeing?
What did John solve the crossword puzzle while ordering?
What did Karl eat dessert before drinking?
Who did Norman talk to Kate while watching?
Who did Paula write two novels before meeting?

Left Branch

How many did Jill hear Antonio receive phone calls?

How many does Alberto suspect that Fred stole magazines?

How many does Bill know that Alice smoked cigarettes?

How many does Edwin think Margaret read books?

How many does Jonathan admit that Marsha stole cars?

Grammatical controls

How does Daniel prefer to take his coffee?

How many books does it seem that Margaret read?

What claim does Ellen believe?

What did Justin see Vanessa throw over the fence?

What did Mary put on the table yesterday?

What did Mary put the last package in?

When did Quentin celebrate his fifteenth birthday?

Where did Max put his favorite sunglass?

Where does Cathy believe that Justin kept the documents?

Where does Mario buy his beautiful neckties?

Who does Edward believe greatly admires Sally?

Who does my cousin Susie think is a genius?

Who does Satoshi greatly admire for their honesty?

Who wonders whether Beatrice saw Benedict?

Why did Patrick begin to bake a cake?

A.1.2 Experiment 2 and 3

That-trace

What did Lisa believe that frightened Kyles wife?

What did Sean profess that destroyed the city?

What does Charles say that stains clothing?

What does Paul believe that killed the millionaire?
Who did Jacob claim that voted for Andrea?
Who did Kathryn confess that hid the watch?
Who does Shannon think that likes Kellys painting?
Who does Susan think that admires Alfred?

Subject Island

What did Nick admit that a jar containing was lost?
What does Angela know that a bottle of fell on the floor?
What does Eric predict a company of will perform tonight?
What does Richard think that a stack of was discarded?
Who did Jessica remember a partner of leaving the firm?
Who did Mark say a fight with started a national scandal?
Who did Rebecca want a discussion with to take place?
Who does Aaron think a picture of will appeal to Katie?

Complex Noun Phrase

What did Alex believe the claim that Amy buried?
What did Laura fight the story that Mark stole?
What does Amber accept the idea that David will study?
What does Robert hate the fact that Sarah is buying?
Who did Sarah hear the rumor that Mary kissed?
Who did Tiffany learn the news that Jeremy pushed?
Who does Bob like the book that Tiffany gave?
Who does Peter love the car that Jim traded to?

Adjunct Island

What did Amanda solve the puzzle while ordering?
What did Christopher fall asleep after doing?
What did Karl eat dessert before drinking?
What did Kyle read a magazine while cooking?
Who did Heather talk about Ed after seeing?
Who did Norman talk to Kate while watching?
Who did Paula write a novel before meeting?
Who will Simon watch a movie before calling?

Left Branch

How many did Adam accept that Beatrice hid cookies?
How many did Jill hear Eric receive phone calls?
How many did Kimberly discover that Bill worked jobs?
How many does Bill know that Alice smoked cigarettes?
How many does James think Rachel read books?
How many does Jim suspect that Kevin stole magazines?
How many does Jonathan admit that Ashley stole cars?
How many will Ryan predict that Megan will drink beers?

Grammatical controls

How many books does it seem that Jamie read?
How many boxes does Shannon fill while working?
How many eggs will Steve eat while at the dinner table?
How many pencils will Timothy decide to donate?
How many people does the train carry during long trips?
How many shoes did Brandon say Erica purchased?
How many shrimp did Jenna put in the freezer before lunch?
How many sugar cubes does Daniel take in his coffee?

What did Alicia think Brian put the last package in?
What did Beth hear worries Karen about the future?
What did Dustin see Vanessa throw over the fence?
What did Max put his favorite sunglasses on after work?
What did Patrick decide that Rebecca wants for dinner?
What did Scot suspect led the police to the Kristins house?
What did the president claim made the economy suffer?
What does Benjamin find in his expensive neckties?
What does Cathy believe that Charles stores the files in?
What does Courtney hope that Tyler buys before Sunday?
What does Greg hate about the way Emily acts?
What does Margo think John discovered at the academy?
What rumor does Ellen believe that Allison started?
What will Adam talk with Ed about at the next meeting?
What will April think about Travis when he leaves school?
What will Steven suggest is the root of the problem?

Who did Jenny do the project with last year?
Who did John say celebrated his birthday before Mary?
Who did Kristen doubt felt sorry for working late?
Who did Laura say saw William at the movies?
Who did Michelle begin to bake a cake for while at home?
Who did Ryan hate before really getting to know him?
Who did Zac give the car keys to before leaving?
Who does Anthony greatly admire for their honesty?
Who does Crystal think stole her watch?
Who does Joshua think has spread the rumor?
Who does Lisa try to console during hard times?
Who does my cousin Susie think is a math genius?

Who will Edward believe admires Sally after the speach?
Who will Jessica look for during the concert on Friday?
Who will Justin write after the conference is over?
Who will Nathan scold for having parked on the street?

A.1.3 Experiment 4 and 5

Set A

That-trace

What will Sean profess that destroyed the city?
What did Lisa believe that frightened Kyles wife?
What does Paul think that killed the millionaire?
Who does Shannon think that likes Kellys painting?
Who did Jacob claim that voted for Andrea?

Complex Noun Phrase

What does Amber accept the idea that David will study?
What will Robert hate the fact that Sarah is buying?
Who does Peter love the car that Jim traded to?
Who does Bob like the book that Tiffany gave?
What did Alex believe the claim that Amy buried?

Adjunct Island

What does Christopher smoke a cigarette after doing?
What did Amanda solve the puzzle while ordering?
Who will Heather talk about Ed after seeing?
What did Karl eat a croissant before drinking?

Who did Norman talk to Kate while watching?

Set B

Want for

What does Robert want for Peter to invest in?
What did Mark want for Vanessa to buy?
Who does Cathy want for Andrea to meet?
Who will Ursula want for Brandon to kiss?
What does Sarah want for Amber to donate?

Whether Island

What did Eric ask whether Ted had visited?
What will Mason wonder whether Rachel discovered?
Who does John wonder whether Alice blamed?
Who did Lisa wonder whether Zac met?
What did Richard ask whether Eleanor preferred?

Coordinate

What will Mark play the guitar and during the concert?
What did Angela drink a shot and at the party?
Who does Cathy see the pediatrician and every month?
What did Mark steal the bread and last week?
Who did Kate meet new member and for dinner?

Ungrammatical controls

Needs construction

What does David say Amber believes needs cut?
What does Sarah think wants walked after dinner?
Who did Mike say requires treated for the injury?
Who will Evelyn say wants praised for the article?
What did Ben recommend needs cooked before he arrives?

Infinitive subject

What will to admit in public be easier someday?
What is to reveal before the end of the book unfair?
Who was to tell the results at the exam cruel?
Who will to surprise on their birthday be fun?
What is to sell in this market not recommendable?

Subject Island

What did Nick admit that a jar containing was lost?
What does Angela know that a bottle of fell on the floor?
Who did Mark say a fight with started a national scandal?
Who did Rebecca want a discussion with to take place?
What does Richard think that a stack of was discarded?

Left branch

How many does James think Rachel read books?

How many does Jim suspect that Kevin stole magazines?

How many does Bill know that Alice smoked cigarettes?

How many did Kimberly discover that Bill worked jobs?

How many did Jill hear Eric receive phone calls?

Grammatical controls

How many books does it seem that Jamie read?

How many eggs will Steve eat while at the dinner table?

How many people does the train carry during long trips?

How many shoes did Brandon say Erica purchased?

How many shrimp did Jenna put in the freezer before lunch?

How many sugar cubes does Daniel take in his coffee?

How many pencils will Timothy decide to donate?

What rumor does Ellen believe that Allison started?

What did Dustin see Vanessa throw over the fence?

What did Alicia think Brian put the last package in?

What will April think about Travis when he leaves school?

What did Max put his favorite sunglasses on after work?

What does Cathy believe that Charles stores the files in?

What does Benjamin find in his expensive neckties?

What does Margo think John discovered at the academy?

What will Adam talk with Ed about at the next meeting?

What did Beth hear worries Karen about the future?

What did the president claim made the economy suffer?

What will Steven suggest is the root of the problem?

What does Courtney hope that Tyler buys before Sunday?

What did Patrick decide that Rebecca wants for dinner?

Who will Jessica look for during the concert on Friday?

Who does Joshua think has spread the rumor?

Who did Ryan hate before really getting to know him?
Who does Crystal think stole her watch?
Who did Jenny do the project with last year?
Who will Justin write after the conference is over?
Who did John say celebrated his birthday before Mary?
Who does my cousin Susie think is a math genius?
Who does Anthony greatly admire for their honesty?
Who did Laura say saw William at the movies?
Who did Michelle begin to bake a cake for while at home?
Who will Nathan scold for having parked on the street?
Who did Kristen doubt felt sorry for working late?
Who does Lisa try to console during hard times?

A.2 Experimental instructions

Experiment 1

Experiment 2, 3, 4

Press the foot pedal to proceed through the instructions. In this task, you are asked to provide a judgment of grammatical acceptability for a number of sentences. Every test sentence is in the form of a question. We would like you to judge whether each sentence is a 'grammatically possible' sentence of English. By 'grammatically possible' we mean possible for you, as a native speaker of English. We are not concerned about whether the sentence would be acceptable to a writing teacher - points of style and clarity are not the issue and we are not concerned about 'who' versus 'whom' or ending a sentence in a preposition. Instead, we are interested in whether the sentence could be accepted as 'English,' in your opinion.

This is not a memory test. You are not required to remember the judgments

that you gave for previous items. Many of the items will be quite similar to one another --to the best of your ability, you should simply ignore this fact and provide an 'independent' judgment on each individual item. Press the button on your RIGHT for 'Yes'. Or the button on the LEFT for 'No'. Here are some for practice.

+ *What did Susan have Ed put on the dining room table?
- *Who does John want whether for who to meet Bob?
+ *What did Bethany think deeply about?
+ *Who claims that Ellen professes to believe Mary?
- *What will Mike want for Andrew to cook tonight?

End of practice. Please continue.

Experiment 3

In this task, you will be given two letter combinations -- one word and one non-word, side by side. Press the Right or Left key to indicate which is the word.

The following items are for practice.

Experiment 5

Press the foot pedal to proceed through the instructions.
You will see sentences presented on the screen, two words at a time. The characters for the remaining words, will appear as asterisks (*). Press the YES/RIGHT button, to read through the sentence. After you have read the sentence,, you will be presented a Yes/No question to answer. Answer the question by pressing,, the YES/RIGHT or NO/LEFT button Please proceed

as quickly and accurately as possible.

Here are sentences for practice.

Press the foot pedal to begin.

REFERENCES

Arai, M., van Gompel, R., and Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, 54(3):218–250.

Bard, E., Robertson, D., and Sorace, A. (1996). Magnitude Estimation of Linguistic Acceptability. *Language*, 72(1):32–68.

Bell, A. (1984). Language Style as Audience Design. *Language in Society*, 13(2):145–204.

Berwick, R. and Weinberg, A. (1984). *The grammatical basis of linguistic performance: Language use and acquisition*. MIT Press.

Bever, T. G. (1970). The cognitive basis for linguistic structures. *Cognition and the development of language*, pages 279–362.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

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

Bock, K. and Griffin, Z. (2000). The persistence of structural priming: Transient activation or implicit learning. *Journal of Experimental Psychology: General*, 129(2):177–192.

Bock, K. and Loebell, H. (1990). Framing sentences. *Cognition*, pages 1–39.

Braze, D. (2003). The Role of Experience in Sentence Acceptability and Sentence Processing. Presentation Slides.

Braze, F. (2002). Grammaticality, acceptability, and sentence processing: A psycholinguistic study. *Unpublished Ph. D. Dissertation, University of Connecticut*.

Bresnan, J. (2007). A few lessons from typology. *Linguistic Typology*, 11(1):297–306.

Carroll, J., Bever, T., and Pollack, C. (1981). The non-uniqueness of linguistic intuitions. *Language*, pages 368–383.

Chomsky, N. (1955). *The Logical Structure of Linguistic Theory*. Plenum Press.

Chomsky, N. (1957). *Syntactic Structures*. Mouton & Co. -'S-Gravenhage.

Chomsky, N. (1959). A review of b.f. skinner's verbal behavior. *Language*, 1:26–58.

Chomsky, N. (1964). Degrees of grammaticalness. *The Structure of Language: Readings in the Philosophy of Language*, pages 384–389.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, N. and Miller, G. (1963). Finitary models of language users. *Handbook of mathematical psychology*, 2:419–491.

Cowart, W. (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgements*. Sage Publications.

Crain, S. and Fodor, J. D. (1987). Sentence matching and overgeneration. *Cognition*, 26(2):123–169.

Featherston, S. (2005). Universals and grammaticality: wh-constraints in German and English. *Linguistics*, 43(4):667–711.

Ferreira, F. and Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, 25(3):348–368.

Ferreira, V. and Bock, K. (2006). The functions of structural priming. *Language and cognitive processes*, 21(7-8):1011.

Ferreira, V., Bock, K., Wilson, M., and Cohen, N. (2008). Memory for syntax despite amnesia. *Psychological science: a journal of the American Psychological Society/APS*, 19(9):940.

Fodor, J. (1978). Parsing strategies and constraints on transformations. *Linguistic Inquiry*, pages 427–473.

Forster, K. and Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(4):680–698.

Forster, K. I. and Forster, J. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35:116–124.

Forster, K. I., Guerrera, C., and Elliot, L. (2009). The Maze Task Forced Incremental Parsing. *Behavior Research Methods*, 41(1):163–171.

Freedman, S. and Forster, K. I. (1985). The psychological status of overgenerated sentences. *Cognition*, 19(2):101–131.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68:1–76.

Gibson, E. and Warren, T. (2004). Reading-time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax*, 7(1):55–78.

Goodall, G. (2004). On the syntax and processing of wh-questions in spanish. In Schmeiser, B., Chand, V., Kelleher, A., and Rodriguez, A., editors, *WCCFL 23 Proceedings*, pages 101–114. Cascadilla Press.

Goodall, G. (2005). The Limits of Syntax in Inversion. In *Proceedings of CLS*, volume 41.

Greenbaum, S. (1976). Contextual influence on acceptability judgments. *Linguistics*, 187:5–11.

Gries, S. (2005). Syntactic Priming: A Corpus-based Approach. *Journal of Psycholinguistic Research*, 34(4).

Haegeman, L. (1994). *Introduction to Government and Binding Theory*. Blackwell Publishers.

Hiramatsu, K. (2000). *Accessing Linguistic Competence: Evidence from Children's and Adults' Acceptability Judgments*. PhD thesis, University of Connecticut.

Jaeger, T. F. and Snider, N. (2007). Implicit Learning and Syntactic Persistence: Surprisal and Cumulativity. *Universif of Rochester Working Papers in the Language Sciences*, 3(1).

Johnson, K. (2008). *Quantitative methods in linguistics*. Blackwell Pub.

Juliano, C. and Tanenhaus, M. (1994). A constraint-based lexicalist account of the subject/object attachment preference. *Journal of Psycholinguistic Research*, 23(6):459–471.

Kaschak, M. and Glenberg, A. (2004). This Construction Needs Learned. *Journal of Experimental Psychology General*, 133(3):18.

Keller, F. (1996). How do humans deal with ungrammatical input? experimental evidence and computational modelling. *Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielefeld, October*, 2734.

King, J. and Just, M. (1991). Individual differences in syntactic processing: the role of working memory. *Journal of memory and language*, 30(5):580–602.

Kluender, R. (1998). On the distinction between strong and weak islands: a processing perspective. *Syntax and semantics*, pages 241–280.

Kluender, R. (2005). Are subject islands subject to a processing account. *Proceedings of the 23rd West Coast Conference on Formal Linguistics*, pages 475–499.

Kluender, R. and Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and Cognitive Processes*, 8(4):573–633.

Lasnik, H. (1984). On the nature of proper government. *Linguistic Inquiry*, pages 235–289.

Luka, B. J. (2005). *Polymorphous linguistics: Jim McCawley's legacy*, chapter A Cognitively Plausible Model of Linguistic Intuition, pages 479–502. MIT Press.

Luka, B. J. and Barsalou, L. W. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language*, 52:436–459.

Maclay, H. and Sleator, M. (1960). Responses to Language: Judgments of Grammaticalness. *International Journal of American Linguistics*, 26(4):275.

McCawley, J. (1982). How far can you trust a linguist. *Language, Mind and Brain*.

Montrul, S., Foote, R., and Perpiñán, S. (2008). Knowledge of Wh-movement in Spanish L2 Learners and Heritage Speakers. In *Selected Proceedings of the 2006 Hispanic Linguistics Symposium. Somerville, MA: Cascadilla Proceedings Project*.

Nagata, H. (2003). Judgments of Grammaticality of Japanese Sentences Violating the Principle of Full Interpretation. *Journal of Psycholinguistic Research*, 32(6):693–709.

Nicol, J., Forster, K., and Veres, C. (1997). Subject–Verb Agreement Processes in Comprehension. *Journal of Memory and Language*, 36(4):569–587.

Noppeney, U. and Price, C. (2004). An fMRI Study of Syntactic Adaptation. *Journal of Cognitive Neuroscience*, 16(4):702–713.

Olarrea, A. (1996). *Pre and postverbal subject positions in Spanish: a minimalist account*. University of Washington.

Ordóñez, F. and Olarrea, A. (2001). Weak subject pronouns in Caribbean Spanish and XP pied-piping. *Features and Interfaces in Romance. Essays in honor of Heles Contreras*, pages 223–239.

Ordóñez, F. and Olarrea, A. (2006). Microvariation in caribbean/non carebbean spanish interrogatives. *Probus*, 18:59–96.

Phillips, C. (2004). The real-time status of island constraints. *Ms., University of Maryland.[Submitted.]*.

Ross, J. (1967). *Constraints on Variables in Syntax, unpublished Ph. D.* PhD thesis, dissertation, MIT.

Sag, I., Arnon, I., Estigarribia, B., Hofmeister, P., Jaeger, T., Pettibone, J., and Snider, N. (2008). Processing Accounts for Superiority Effects.

Sag, I., Hofmeister, P., and Snider, N. (2007). Processing Complexity in Subjacency Violations: The Complex Noun Phrase Constraint.

Scarborough, D., Cortese, C., and Scarborough, H. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, 3(1):1–17.

Schütze, C. (1996). *The Empirical Base of Linguistics: grammaticality judgments and linguistic methodology*. University of Chicago Press.

Severance, E. and Washburn, M. (1907). The loss of associative power in words after long fixation. *American Journal of Psychology*, 18:182–186.

Snyder, W. (1994). A psycholinguistic investigation of weak crossover, islands, and syntactic satiation effects. *Poster presented at the CUNY Sentence Processing Conference, CUNY Graduate Center, New York*.

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry*.

Sobin, N. (1987). The variable status of Comp-trace phenomena. *Natural Language & Linguistic Theory*, 5(1):33–60.

Sobin, N. (2002). The Comp-trace effect, the adverb effect and minimal CP. *Journal of Linguistics*, 38(03):527–560.

Sorace, A. and Keller, F. (2005). Gradience in linguistic data. *Lingua*, 115(11):1497–1524.

Spencer, N. (1973). Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research*, 2(2):83–98.

Sprouse, J. (2007a). *A Program for Experimental Syntax: Finding the relationship between acceptability and grammatical knowledge*. PhD thesis, University of Maryland.

Sprouse, J. (2007b). Continuous Acceptability, Categorical Grammaticality, and Experimental Syntax. *Biolinguistics*, 1:118–129.

Sprouse, J. (2007c). Experimental syntax: What does it get you? Manuscript.

Sprouse, J. (2007d). Syntactic satiation: Toward an etiology of linguist's disease. Manuscript.

Sprouse, J. (2009). Revisiting Satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*, 40(1).

Torrego, E. (1984). On inversion in spanish and some of its effects. *Linguistic Inquiry*, 15(1):103–129.

Tremblay, A. (2005). Theoretical and methodological perspetives on the use of grammaticality judgment tasks in linguistic theory. *Second Language Studies*, 24(1):129–167.

Trueswell, J. and Tanenhaus, M. (1991). Tense, temporal context and syntactic ambiguity resolution. *Language and Cognitive Processes*, 6(4):303–338.

Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):10.

Wundt, W. (1900). *Die Sprache*. W. Engelmann, Leipzig.